Hasso Plattner Institute for Digital Engineering, University of Potsdam
Data Analytics and Computational Statistics Research Group
Prof. Dr. Bernhard Renard

Integrative Biomarker Detection Using Prior Knowledge on Gene Expression Data Sets

Dissertation

submitted in partial fulfillment
of the requirements for the academic degree of

Doctor of Engineering (Dr.-Ing.)

in the scientific discipline practical computer science

to the Digital Engineering Faculty
at the University of Potsdam
by

Cindy Perscheid

March 4, 2023

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-58241
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-582418

Abstract

Gene expression data is analyzed to identify biomarkers, e.g. relevant genes, which serve for diagnostic, predictive, or prognostic use. Traditional approaches for biomarker detection select distinctive features from the data based exclusively on the signals therein, facing multiple shortcomings in regards to overfitting, biomarker robustness, and actual biological relevance. *Prior knowledge approaches* are expected to address these issues by incorporating prior biological knowledge, e.g. on gene-disease associations, into the actual analysis. However, prior knowledge approaches are currently not widely applied in practice because they are often use-case specific and seldom applicable in a different scope. This leads to a lack of comparability of prior knowledge approaches, which in turn makes it currently impossible to assess their effectiveness in a broader context.

Our work addresses the aforementioned issues with three contributions. Our first contribution provides formal definitions for both prior knowledge and the flexible integration thereof into the feature selection process. Central to these concepts is the automatic retrieval of prior knowledge from online knowledge bases, which allows for streamlining the retrieval process and agreeing on a uniform definition for prior knowledge. We subsequently describe novel and generalized prior knowledge approaches that are flexible regarding the used prior knowledge and applicable to varying use case domains. Our second contribution is the benchmarking platform *Comprior*. Comprior applies the aforementioned concepts in practice and allows for flexibly setting up comprehensive benchmarking studies for examining the performance of existing and novel prior knowledge approaches. It streamlines the retrieval of prior knowledge and allows for combining it with prior knowledge approaches. Comprior demonstrates the practical applicability of our concepts and further fosters the overall development and comparability of prior knowledge approaches. Our third contribution is a comprehensive case study on the effectiveness of prior knowledge approaches. For that, we used Comprior and tested a broad range of both traditional and prior knowledge approaches in combination with multiple knowledge bases on data sets from multiple disease domains. Ultimately, our case study constitutes a thorough assessment of a) the suitability of selected knowledge bases for integration, b) the impact of prior knowledge being applied at different inte-

gration levels, and c) the improvements in terms of classification performance, biological relevance, and overall robustness.

In summary, our contributions demonstrate that generalized concepts for prior knowledge and a streamlined retrieval process improve the applicability of prior knowledge approaches. Results from our case study show that the integration of prior knowledge positively affects biomarker results, particularly regarding their robustness. Our findings provide the first in-depth insights on the effectiveness of prior knowledge approaches and build a valuable foundation for future research.

Zusammenfassung

Biomarker sind charakteristische biologische Merkmale mit diagnostischer oder prognostischer Aussagekraft. Auf der molekularen Ebene sind dies Gene mit einem krankheitsspezifischen Expressionsmuster, welche mittels der Analyse von Genexpressionsdaten identifiziert werden. Traditionelle Ansätze für diese Art von *Biomarker Detection* wählen Gene als Biomarker ausschließlich anhand der vorhandenen Signale im Datensatz aus. Diese Vorgehensweise zeigt jedoch Schwächen insbesondere in Bezug auf die Robustheit und tatsächliche biologische Relevanz der identifizierten Biomarker. Verschiedene Forschungsarbeiten legen nahe, dass die Berücksichtigung des biologischen Kontexts während des Selektionsprozesses diese Schwächen ausgleichen kann. Sogenannte *wissensbasierte Ansätze* für Biomarker Detection beziehen vorhandenes biologisches Wissen, beispielsweise über Zusammenhänge zwischen bestimmten Genen und Krankheiten, direkt in die Analyse mit ein. Die Anwendung solcher Verfahren ist in der Praxis jedoch derzeit nicht weit verbreitet, da existierende Methoden oft spezifisch für einen bestimmten Anwendungsfall entwickelt wurden und sich nur mit großem Aufwand auf andere Anwendungsgebiete übertragen lassen. Dadurch sind Vergleiche untereinander kaum möglich, was es wiederum nicht erlaubt die Effektivität von wissensbasierten Methoden in einem breiteren Kontext zu untersuchen.

Die vorliegende Arbeit befasst sich mit den vorgenannten Herausforderungen für wissensbasierte Ansätze. In einem ersten Schritt legen wir formale und einheitliche Definitionen für vorhandenes biologisches Wissen sowie ihre flexible Integration in den Biomarker-Auswahlprozess fest. Der Kerngedanke unseres Ansatzes ist die automatisierte Beschaffung von biologischem Wissen aus im Internet frei verfügbaren Wissens-Datenbanken. Dies erlaubt eine Vereinfachung der Kuratierung sowie die Festlegung einer einheitlichen Definition für biologisches Wissen. Darauf aufbauend beschreiben wir generalisierte wissensbasierte Verfahren, welche flexibel auf verschiedene Anwendungsfalle anwendbar sind. In einem zweiten Schritt haben wir die Benchmarking-Plattform *Comprior* entwickelt, welche unsere theoretischen Konzepte in einer praktischen Anwendung realisiert. Comprior ermöglicht die schnelle Umsetzung von umfangreichen Experimenten für den Vergleich von wissensbasierten Ansätzen. Comprior übernimmt die

Beschaffung von biologischem Wissen und ermöglicht dessen beliebige Kombination mit wissensbasierten Ansätzen. Comprior demonstriert damit die praktische Umsetzbarkeit unserer theoretischen Konzepte und unterstützt zudem die technische Realisierung und Vergleichbarkeit wissensbasierter Ansätze. In einem dritten Schritt untersuchen wir die Effektivität wissensbasierter Ansätze im Rahmen einer umfangreichen Fallstudie. Mithilfe von Comprior vergleichen wir die Ergebnisse traditioneller und wissensbasierter Ansätze im Kontext verschiedener Krankheiten, wobei wir für wissensbasierte Ansätze auch verschiedene Wissens-Datenbanken verwenden. Unsere Fallstudie untersucht damit a) die Eignung von ausgewählten Wissens-Datenbanken für deren Einsatz bei wissensbasierten Ansätzen, b) den Einfluss verschiedener Integrationskonzepte für biologisches Wissen auf den Biomarker-Auswahlprozess, und c) den Grad der Verbesserung in Bezug auf die Klassifikationsleistung, biologische Relevanz und allgemeine Robustheit der selektierten Biomarker.

Zusammenfassend demonstriert unsere Arbeit, dass generalisierte Konzepte für biologisches Wissen und dessen vereinfachte Kuration die praktische Anwendbarkeit von wissensbasierten Ansätzen erleichtern. Die Ergebnisse unserer Fallstudie zeigen, dass die Integration von vorhandenem biologischen Wissen einen positiven Einfluss auf die selektierten Biomarker hat, insbesondere in Bezug auf ihre biologische Relevanz. Diese erstmals umfassenderen Erkenntnisse zur Effektivität von wissensbasierten Ansätzen bilden eine wertvolle Grundlage für zukünftige Forschungsarbeiten.

While this dissertation is a "child" of mine of rather intellectual and technical nature, it still took a village full of colleagues, friends, and family to bring this up. I am dearly thankful for each of them, as without these people this work would not have achieved the quality it has today. First and foremost, I thank Professor Hasso Plattner, for allowing me to pursue my research at his chair at a great and inspiring workplace. I am most grateful that Professor Bernhard Renard agreed to take over supervision at the final stages, and for him providing guidance and always just the right feedback. I thank my colleagues from the Enterprise Platform and Integration Concepts chair for memorable off-site events and a great work environment. Thanks also go to my colleagues from the Data Analytics and Computational Statistics chair, for the enjoyable get-togethers and particularly for warmly welcoming me into their group after I changed supervisors. Special gratitude goes to Milena Kraus, with whom I shared not only an office but also the worries that occur along the PhD journey, especially when trying to reconcile with family life. Our supportive and inspiring discussions on both private and work-related topics often gave just the right impulses to change my perspective and see issues from a different angle. I further thank my colleagues Ralf Teusner, Johannes Hügle, and Katharina Baum for revising large parts of this dissertation and providing honest feedback. I thank my student Bastien Grasnick, who conducted the first experiments during his Master's thesis that laid the foundation for my research. I thank David Weese for his continuous mentoring throughout the years. Thanks also go to our chair assistance Marilena Davis, for her patience with my incorrect paper forms and always having cookies around.

I am deeply grateful for the everlasting support of my husband Michael Perscheid who always lifted me up when I thought I could not go any further, for pointing my view to my achievements when self-doubts tried to prevail, for living a truly equal partnership and being my love, best friend, and partner in crime in one person. I also thank our wonderful daughter Theresa, who constantly reminded me of what really counts in life. Much appreciation goes to my mother- and sister-in-law, Petra and Heike Perscheid, for taking over childcare when the kindergarten once more had to close. I am sincerely grateful for my longtime friends Laura Börner, Andrina Mascher, Claudia Lehmann, Cornelia Rehbein, and Julia Schewe; for their supportive community whose empowering conversations allowed me to carry on my work. Finally, thanks go to my parents Sylvia and Andreas Fähnrich, who brought me up to be the person with the discipline, perseverance, and resilience that it takes to finish a doctorate. I dedicate this work to my father Andreas, who played a decisive role in me starting and following down my path at HPI. It is with a heavy heart that I go these last meters without him. I dearly hope that he can see me reach the finish line — from wherever that may be.

Contents

1 Introduction 1
- 1.1 The Importance of Biomarkers for Precision Medicine 1
- 1.2 Integration of Biological Context into Biomarker Detection 2
- 1.3 Contributions 4
  - 1.3.1 Definition of Prior Knowledge and Classification of Prior Knowledge Approaches 4
  - 1.3.2 Comprior — Implementation and Benchmarking of Prior Knowledge Approaches 5
  - 1.3.3 Assessment of the Impact of Prior Knowledge 6
  - 1.3.4 Conclusion 6
- 1.4 Outline 7

2 Background 9
- 2.1 From DNA to Proteins 10
  - 2.1.1 Gene Expression — Building Functional Products 10
  - 2.1.2 How Altered DNA Affects Gene Expression 11
- 2.2 Gene Expression Profiling — Measuring Gene Activity 12
  - 2.2.1 Transcriptomics Data from Microarrays 12
  - 2.2.2 Transcriptomics Data from RNA Sequencing 13
  - 2.2.3 Transcriptomics Data Postprocessing 15
- 2.3 Biomarker Detection — Analyzing Gene Expression Data 17
  - 2.3.1 Differential Expression Analysis 18
  - 2.3.2 Traditional Feature Selection Approaches 18
  - 2.3.3 Shortcomings of Traditional Approaches 19
- 2.4 Annotating Gene Sets with Biological Knowledge 19
  - 2.4.1 Annotation Knowledge Bases 20
  - 2.4.2 Interaction Knowledge Bases 22
  - 2.4.3 Meta Knowledge Bases 24
  - 2.4.4 Programmatic Access to Knowledge Bases 26
- 2.5 Summary 28

3 Related Work . . . 29
3.1 Biomarker Detection Using Prior Knowledge . . . 29
3.1.1 Prior Knowledge Approaches — State of the Art . . . 29
3.1.2 Insights . . . 31
3.2 Benchmarking in Bioinformatics . . . 32
3.2.1 Benchmarking Tools — State of the Art . . . 32
3.2.2 Insights . . . 33
3.2.3 Implications for the Contributions of this Thesis . . . 35
3.3 Summary . . . 35

4 Making Prior Knowledge Approaches Flexible: Defining Prior Knowledge and Integration Strategies . . . 37
4.1 A Conceptual Definition of Prior Biological Knowledge for Gene Expression Data . . . 37
4.2 Transforming Prior Knowledge Levels . . . 39
4.2.1 Transforming Higher-Level Prior Knowledge into Lower-Level Prior Knowledge . . . 40
4.2.2 Transforming Lower-Level Prior Knowledge into Higher-Level Prior Knowledge . . . 41
4.3 Strategies for Integrating Prior Knowledge into Feature Selection . . . 42
4.3.1 Modifying Approaches . . . 44
4.3.2 Combining Approaches . . . 45
4.3.3 Network Approaches . . . 47
4.4 Generalized Approaches to Flexibly Integrate Prior Knowledge into Feature Selection . . . 48
4.4.1 Modifying Prior Knowledge Approaches . . . 48
4.4.2 Prefiltering Approach . . . 49
4.4.3 Postfiltering Approach . . . 49
4.4.4 Extension Approach . . . 49
4.4.5 Combining Approach . . . 50
4.4.6 Network Approach . . . 50
4.5 Summary . . . 52

5 Comprior: A Software Tool to Effortlessly Implement and Benchmark Prior Knowledge Approaches . . . 53
5.1 General Description of Comprior . . . 53
5.1.1 Supported Processing Functionality . . . 53
5.1.2 Tool FAIRness . . . 55
5.2 Architecture Design . . . 56
5.3 Ensuring Extensibility by Custom Functionality . . . 58
5.3.1 Standardized Interfaces between Components . . . 58
5.3.2 Including External Code . . . 58

5.4 Reducing the Implementation Effort for Comprehensive Benchmark Experiments . . . 59
5.4.1 Experiment Configuration . . . 60
5.4.2 Enabling a Flexible Combination of Feature Selection Approaches 61
5.4.3 Enabling a Comprehensive Result Assessment . . . 62
5.4.4 Handling Multiple Identifier Formats . . . 64
5.5 Uniform Access to Prior Knowledge . . . 66
5.5.1 Knowledge Base Concept . . . 67
5.5.2 Processing Pathways . . . 67
5.5.3 Mapping Prior Knowledge Levels . . . 68
5.6 Summary . . . 69

6 Assessment of the Impact of Prior Knowledge Integration During Biomarker Detection . . . 71
6.1 Data Sets . . . 71
6.2 Experiment Setup . . . 74
6.2.1 System Specifications . . . 74
6.2.2 Feature Selection Approaches . . . 75
6.2.3 Identifier Mapping . . . 76
6.2.4 Prior Knowledge Retrieval . . . 76
6.2.5 Classification . . . 77
6.2.6 Enrichment Analysis . . . 77
6.2.7 Feature Set and Enrichment Robustness . . . 78
6.3 Results . . . 78
6.3.1 Coverage of Diseases in Knowledge Bases . . . 79
6.3.2 Runtime Performance . . . 81
6.3.3 Performances on Data Sets from Different Disease Domains . . . 85
6.3.4 Comparing Traditional Approaches with Adaptations Therof Using Prior Knowledge . . . 97
6.3.5 Comparing Results of Different Complexity Levels of Integration . 101
6.3.6 Comparing Results of Different Knowledge Bases . . . 104
6.3.7 Threats to Validity of the Findings from the Case Study . . . 109
6.4 Summary . . . 110

7 Discussion . . . 113
7.1 Improving the Applicability of Prior Knowledge Approaches . . . 113
7.1.1 Generalized Approaches and Unified Definitions for Prior Knowledge . . . 114
7.1.2 A Technical Infrastructure for Development and Evaluation . . . 114
7.1.3 Transferability of our Concepts to Other Omics Domains . . . 115
7.1.4 Does More Flexibility Result in Improved Application in Practice? 117
7.2 The Impact of Integrating Prior Knowledge into Feature Selection . . . 119

7.2.1 The Choice of Knowledge Base Affects Performance Results 119
7.2.2 Prior Knowledge Approaches are Feasible, but not Real-Time 120
7.2.3 Marginal Improvement in Classification Performance, but more Enrichments and Higher Robustness 121
7.2.4 Different Integration Levels Affect Biomarker Results 122
7.2.5 Modifying or Network Approaches are the Methods of Choice 122
7.2.6 Do Prior Knowledge Approaches Keep Their Promises? 123
7.3 Directions for Future Work 125

8 Conclusion 127

9 Publications 137
9.1 Journal Articles 137
9.2 Conference Articles 137
9.3 Workshop Articles 138
9.4 Technical Reports 138

10 Appendix 141

Bibliography 181

1

Introduction

In May 2017, the U.S. Food and Drug Administration (FDA) approved the first cancer treatment that is not based on the type of tumor but rather specific patient characteristics[1]. The approval of this treatment was an important step towards precision cancer treatment: for the first time, a cancer patient could be treated based on their individual molecular profile, rather than just on the tumor's tissue type. Subsequently, in October 2018, the second such drug was approved[2] which further paved the way for precision medicine.

The term *precision medicine*, which is often used interchangeably with *personalized medicine*, "refers to the tailoring of medical treatment to the individual characteristics of each patient" [141]. It allows for a fine-grained classification of patients according to their individual characteristics in regards to disease susceptibility, biology, prognosis, or treatment response. Consequently, precision medicine promises to establish more effective treatments with less side effects for patients.

The advances in molecular technologies, e.g. next-generation sequencing (NGS), provided an impetus to applying precision medicine broadly. Large-scale computational analyses of whole cohorts on the molecular level provide insights into the interplay of particular characteristics and clinical outcomes, e.g. treatment response or chance of survival. In recent years, research has increasingly focused on identifying characteristics that are truly correlated to clinical outcomes, as these are the key enabler for precision medicine. Such characteristics are called *biomarkers*.

1.1 The Importance of Biomarkers for Precision Medicine

The *Biomarkers, EndpointS, and other Tools* (BEST) Resource, jointly established by the FDA and the National Institutes of Health (NIH), defines a biomarker as follows:

[1] https://www.fda.gov/news-events/press-announcements/fda-approves-first-cancer-treatment-any-solid-tumor-specific-genetic-feature

[2] https://www.fda.gov/news-events/press-announcements/fda-approves-oncology-drug-targets-key-genetic-driver-cancer-rather-specific-type-tumor

> "A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers. A biomarker is not an assessment of how an individual feels, functions, or survives." [72]

Biomarkers are therefore anything that can be objectively measured in a patient: heart rates, blood pressure, hormone concentration, radiometric imaging, genes, or even whole molecular complexes. There are different types of biomarkers, which allow for — amongst others — a precise disease diagnosis (*diagnostic*), selection of appropriate treatment (*predictive*), and assessment of disease progression (*prognostic*) [72]. For example, levels of the prostate-specific antigene (PSA) are used to diagnose prostate cancer [251]. Expression rates of the HER2 gene are examined to decide on breast cancer treatment with trastuzumab, a medication that supresses HER2 expression rates [238]. MammaPrint©, a set of 70 marker genes detected by Van't Veer et al., is used to assess breast cancer recurrence after treatment [240].

The detection of these biomarkers was made possible by the advances in molecular technology. In the last decade, the amount of data generated on the molecular level which captures genetic information, protein activity, and other molecular information, has been growing rapidly. High-dimensional data sets containing measurements from tens of thousands of molecular examination points, e.g. genes, are generated within a very short time. This has left research with the challenge of finding ways to analyze these high-dimensional data sets and identify biomarkers. This is merely a task of feature reduction or extraction: the high-dimensional feature space must be reduced to those genes that achieve best performance, e.g. in classification, clustering, or prediction tasks. However, it soon became apparent that biomarkers retrieved from data-driven methods, i.e. methods that assess a feature based on its statistical characteristics, are not robust and of questionable biological relevance [48, 52, 53, 82, 88, 138, 163, 262]. These observations were also due to the high error-proneness and data layout of high-throughput data sets, which are generated at an ever increasing speed, but at the cost of data quality. While they are of high dimensionality — containing multiple tens of thousands of features — they only have a small proportion of samples. Consequently, more complex machine learning approaches tend to overfit, causing a random signal to be interpreted as relevant. This issue concerns gene expression data in particular, as data sets generated via DNA sequencing tend to exhibit strong random bias, i.e. that randomly selected features show a robust performance in classification or prediction tasks [200].

1.2 Integration of Biological Context into Biomarker Detection

What is therefore missing during analysis is the biological context, i.e. that features are also assessed on the basis of their interactions and involvement in biological processes,

e.g. cancer hallmarks. In response to this issue, recent research now focuses on *integrative* analyses that aim to view and assess the data on a holistic level based on biological factors. Such integrative analyses have shown to improve analysis results and lead to more robust and biologically meaningful biomarkers [19, 151]. Currently, research puts a major focus on the analysis of multi-omics data sets: integrating multiple artifacts of the same object, e.g. gene expression, mutation, or regulatory data, for a holistic view of cell processes. However, this kind of integrative analysis is not feasible if only one type of data is available. In such cases, biological context can still be incorporated into the analysis, via prior biological knowledge from publicly available knowledge bases.

Nowadays, a growing number of knowledge bases provide the most recent and highly-curated insights from research, e.g. on gene-gene or gene-disease interactions, gene functions and co-expressions, or signaling pathways [10, 43, 60, 101, 128, 143, 228]. Even meta knowledge bases that integrate information from various well recognized resources are emerging, e.g. DisGeNET and Open Targets [108, 165]. However, these resources are not applied during the actual analysis. Instead, they are used afterwards to validate the biological significance of the identified biomarkers. This observation leads us to the problem statement of this thesis:

Problem Statement: *Despite the abundance of biological knowledge which is currently available and being generated, it is not applied to the analysis when assessing the biological relevance of a biomarker.*

Strategies that already incorporate biological knowledge from external resources during biomarker detection are referred to as *prior knowledge approaches*. With the increasing volume and availability of knowledge bases, prior knowledge approaches could turn out to be a powerful alternative for an integrative analysis when no multi-omics data is at hand.

However, prior knowledge approaches are not widely used in practice. There are three major, mutually dependent reasons for this: lack of applicability, missing comparability, and insufficient research on the effectivity of prior knowledge integration. Most approaches are custom, standalone solutions that cannot be flexibly modified for other use cases, e.g. to use another knowledge base. Instead, most approaches focus on only a few specific knowledge bases, despite the availablility of many more, and even meta knowledge bases. What is more, most approaches are rarely made available to the public, e.g. by providing the source code or sample applications. Consequently, comparisons between prior knowledge approaches are rare. The usefulness of a new approach is most often only evaluated in context of non-integrative approaches. Because of the aforementioned conditions, knowledge about the actual effectivity of prior knowledge integration is scarce. The capabilities of the different integration levels of prior knowledge are unknown. The impact of a chosen knowledge base on the results is unclear. The current state of research makes it impossible to thoroughly assess the usefulness and the ef-

fect of prior knowledge integration. Based on the aforementioned observations, we have formulated the following research questions that will be addressed in this thesis:

RQ1: *How can we improve the practical applicability of prior knowledge approaches in practice and subsequently enable better comparability?*

In comparison to integrative approaches, traditional approaches are widely used. The main reason for this is their easy applicability: most traditional approaches are use-case-independent and available as packages in common programming languages, e.g. R or Python. Traditional approaches can thus be seamlessly integrated into any analysis. It should be the goal for integrative approaches – and in the scope of this thesis, for prior knowledge approaches – to achieve applicability at a similar level. An increased practical applicability of prior knowledge approaches facilitates comparisons with other prior knowledge approaches regarding quantitative performance and biological relevance. This calls for a corresponding evaluation infrastructure that allows evaluation strategies to be specified effortlessly and to select uniform measures to assess biomarker robustness, accuracy, and biological relevance. This thesis aims to provide such an evaluation infrastructure that allows prior knowledge approaches, knowledge bases, and traditional approaches to be compared with respect to their effective performance and usability.

RQ2: *What is the impact of integrating prior biological knowledge on different analysis levels of biomarker detection regarding the*

a) delivery of interpretable and biologically meaningful results,

b) robustness across approaches and data sets, and

c) computational complexity and transparency?

For traditional approaches there are many quantitative and qualitative evaluations and subsequent usage recommendations [22, 82, 94, 112]. This is not the case for prior knowledge approaches. Due to their limited applicability and comparability, it is unclear what impact knowledge bases and integration strategies have on the outcome. Additionally, existing approaches are seldom evaluated in regards to biomarker robustness across data sets. The objective of this thesis is to provide a comprehensive, first-time study on prior knowledge approaches and their performance in a broader context.

1.3 Contributions

We address the aforementioned research questions through multiple contributions. In doing so, we focus on the integrative analysis of gene expression data sets.

1.3.1 Definition of Prior Knowledge and Classification of Prior Knowledge Approaches

The general concepts applied in traditional approaches are well described with a subsequent classification of existing approaches into distinct categories [86, 130, 181]. Al-

though prior knowledge approaches have been around in research for over a decade now, a clear definition of the general concepts, e.g. of prior knowledge, and a characteristic-based categorization of prior knowledge approaches are still missing. With this work, we provide a definition of the formal concepts for integrating prior knowledge into biomarker detection. As such, we identify and describe what kind of prior knowledge is suitable for integration and under what assumptions transformations between different types of prior knowledge are possible. We then describe strategies for integrating prior knowledge into biomarker detection and derive a subsequent classification of existing prior knowledge approaches.

Parts of the concepts described in this work have been published in the following publications:

> **C. Perscheid**. "Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches". *Briefings in Bioinformatics* 22.3 (2020), bbaa151
>
> B. Grasnick, **C. Perscheid**, and M. Uflacker. "A Framework for the Automatic Combination and Evaluation of Gene Selection Methods". In: *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Ed. by F. Fdez-Riverola, M. S. Mohamad, M. Rocha, J. F. De Paz, and P. González. Springer. Cham: Springer International Publishing, 2019, pp. 166–174

1.3.2 Comprior — Implementation and Benchmarking of Prior Knowledge Approaches

We developed and implemented Comprior to improve the practical applicability, extensibility, and comparability of prior knowledge approaches. Comprior provides the technical infrastructure to rapidly implement prior knowledge approaches and to evaluate them against both traditional and prior knowledge approaches in regards to robustness, quantitative performance, and biological relevance. Comprior provides easy access to multiple knowledge bases and flexible combination options for traditional approaches to biomarker detection. Comprior was designed with a modular and extensible architecture, providing well-defined interfaces for adding new functionality as needed, e.g. with regards to preprocessing, biomarker detection approaches, knowledge bases, or evaluation. Comprior also provides an evaluation infrastructure that enables automated comparisons across approaches. It provides standardized measures regarding quantitative performance and biological relevance. It also allows biomarker robustness to be assessed by providing cross-validation strategies within and across data sets. Approaches for biomarker detection provided by and integrated in Comprior can be seamlessly embedded in custom analysis workflows for use in practice.

Comprior and underlying concepts have been described in prior publications:

C. Perscheid. "Comprior: facilitating the implementation and automated benchmarking of prior knowledge-based feature selection approaches on gene expression data sets". *BMC Bioinformatics* 22.1 (2021), pp. 1–15

C. Perscheid, B. Grasnick, and M. Uflacker. "Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches". *Journal of Integrative Bioinformatics* 16.1 (2019), p. 20180064

1.3.3 Assessment of the Impact of Prior Knowledge

Currently, comparisons of prior knowledge approaches are limited: they seldom compare themselves with other prior knowledge approaches, and only with few traditional approaches. Most approaches do not examine the robustness across data sets, although this is one of the key advantages assumed for integrative approaches. Comparisons are typically limited to showing the improvement of an approach that slightly adapts an already existing approach — e.g., by extending a Lasso strategy and introducing biological relevance via a penalty term [261]. While this allows relative improvement to be demonstrated, it does not allow the actual usefulness of the approach in a broader context to be described, e.g. by comparisons to other strategies. Many questions remain unanswered. Is it already sufficient to use a low-complexity prior knowledge approach to achieve the same performance as a high-complex embedded traditional approach? What is gained by applying a dense integration of prior knowledge, compared to simple filtering strategies? What knowledge bases are most suitable? How strongly does the choice of a knowledge base affect result sets? Do we already have a sufficient coverage of knowledge bases to achieve a good performance?

To answer these questions, we have carried out a case study with multiple data sets. We have used Comprior to evaluate prior knowledge approaches from all different integration levels and multiple selected traditional approaches which are representative for their different types. We have applied these approaches on multiple cancer and Alzheimer's disease data sets. We have assessed the effectivity of an approach based on its quantitative performance, e.g. classification accuracy, and biological relevance. We have put special emphasis on examining the robustness of biomarkers across data sets.

Findings from these case studies will be published in a separate publication:

C. Perscheid. "The impact of integrating prior knowledge during biomarker detection: A case study on high-dimensional gene expression data" (2022). in preparation

1.3.4 Conclusion

Our contributions increase the practical applicability of integrative approaches by enabling a flexible integration of biological context, particularly prior biological knowledge, into the analysis of gene expression data. For the first time, Comprior enables researchers

to compare their own approach to others and to assess the effectiveness of the applied integration concepts. Ultimately, our work further promotes a widespread use of prior knowledge approaches in the future, taking research another step towards identifying robust biomarkers for precision medicine.

1.4 Outline

The remainder of this work is structured as follows:

Chapter 2 introduces the general background, ranging from biological fundamentals related to gene expression, generation and status quo analysis of gene expression data, and existing knowledge bases that are suitable for prior knowledge approaches.

Chapter 3 presents prior work related to our research. It provides an overview of benchmarking systems related to Comprior. Furthermore, it describes the current status quo of prior knowledge approaches, unsolved issues, and resulting challenges.

Chapter 4 describes formal concepts for integrating prior knowledge into biomarker detection. It defines the different types prior knowledge suitable for integration and transformations thereof, and subsequently categorizes and describtes prior knowledge approaches based on the degree of prior knowledge integration. Finally, for each category of prior knowledge approaches, it provides generalized concepts that allow to flexibly integrate prior knowledge.

Chapter 5 depicts the technical design and implementation of Comprior. It describes key features, provides details on the architecture design, and gives insights into the technical realization of selected features.

Chapter 6 describes the setup and outcomes of our case study to assess the influence of prior knowledge integration on detected biomarkers. In the case study, we have applied multiple data sets from Alzheimer's Disease, breast cancer, and glioma to evaluate selected traditional and prior knowledge approaches. We combine the applied prior knowledge approaches with multiple knowledge bases to examine their impact on result sets. We further assess the effectiveness of the tested approaches in regards to quantitative performance and biological relevance, with special focus on biomarker robustness.

Chapter 7 discusses major findings from our case study in a broader context, but also limitations of the presented approach, highlighting open challenges and providing new impulses for the research community.

Chapter 8 concludes our work by summarizing our major findings and promising aspects to address in future work.

2

Background

This chapter provides background information that is required to understand the domain of biomarker detection on gene expression data sets. It explains the biological details of DNA, gene expression, and how altered DNA can affect this biological process. It further describes the background of the two main strategies that are used for measuring gene expression activity and elaborates on necessary data processing steps. Finally, this chapter explains the overall objectives and shortcomings of such analyses, and how available biological knowledge contributes to a better assessment of the results.

Figure 2.1 depicts how the process of analyzing gene expression information spans across four areas. First are the biological processes taking place in the cells. These biological processes can be measured with the help of molecular technology to transform the information into a machine-readable format. The machine-readable gene expression information is then analyzed using computational methods. Finally, resulting biomarker candidates are assessed for their biological and clinical relevance by annotating them with biological information.

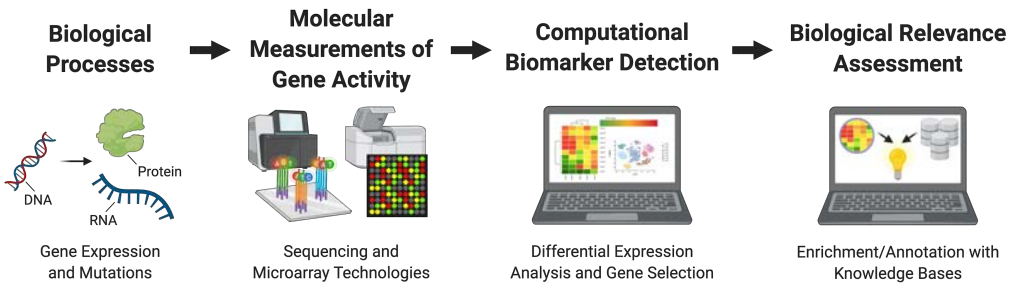


Fig. 2.1: The domain of biomarker detections spans across four areas: the actual biological processes taking place in the cells, molecular measurement of gene activity, the computational analysis of the resulting data sets, and annotation of these results with biological information for assessing their actual biological relevance. Figure created with BioRender.com

2.1 From DNA to Proteins

The human *DNA* is a double-helix-shaped molecule containing all hereditary information. It is made up of chemical building blocks called *nucleotides*. Nucleotides consist of three parts: a phosphate group, a sugar group, and one of four types of nitrogen bases: Adenin (A), Cytosin (C) Guanin (G), and Thymin (T). Each base is complemented by another base (from the other side of the DNA) to build a *base pair*. The human DNA consists of 3 billion base pairs.

The DNA is divided into functional units called *genes*. Genes are subsequences of bases within the DNA that provide instructions for synthesizing functional products. *Functional products* can be proteins or Ribonucleic Acids (*RNA*s). Proteins play an essential role in the human body: they are involved in the biological processes taking place in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins consist of multiple smaller units called *amino acids*, which are concatenated to a long chain. There are 20 different types of amino acids that occur in human DNA. The way these amino acids are concatenated determines the structure and specific function of a protein. RNAs, as the second type of functional products, are single-stranded molecules that are involved in the synthesis of proteins, e.g. by regulatory actions. Instead of carrying the base Thymin, RNAs contain Uracil (U). Depending on their functions, there are multiple types of RNA, e.g. messenger or transfer RNA.

A gene is made up of two types of regions: *coding regions* and *non-coding* regions. Coding regions provide the actual building instructions for functional products. Non-coding regions of a gene are not actively 'read' like coding regions. Instead, they are involved in the building process of the gene's functional product, e.g. by providing binding sites for enzymes that start protein synthesis. Coding regions make up only about 1% of the human DNA, whereas the remaining parts are covered by non-coding regions [242].

2.1.1 Gene Expression — Building Functional Products

Gene expression describes the process of reading and synthesizing the genetic information encoded in a gene's DNA to build a functional product. This process is also referred to as the *central dogma of molecular biology*. The quantity — how often gene expression is carried out, i.e. the production rate for the functional product — is called *expression level*. Gene expression is separated into the steps of *transcription* and *translation* as depicted by Figure 2.2.

During **Transcription**, a single DNA strand acts as a template to create an RNA *transcript*. For that, the DNA is first unwound and split into two separate strands. Along one of these strands, an enzyme called RNA polymerase subsequently adds one ribonucleotide after the other, which is complementary to the nucleotide at the DNA strand. The result is an RNA transcript that is complementary to the template DNA strand, with the exception that Thymin (T) is replaced with Uracil (U). If the expressed gene

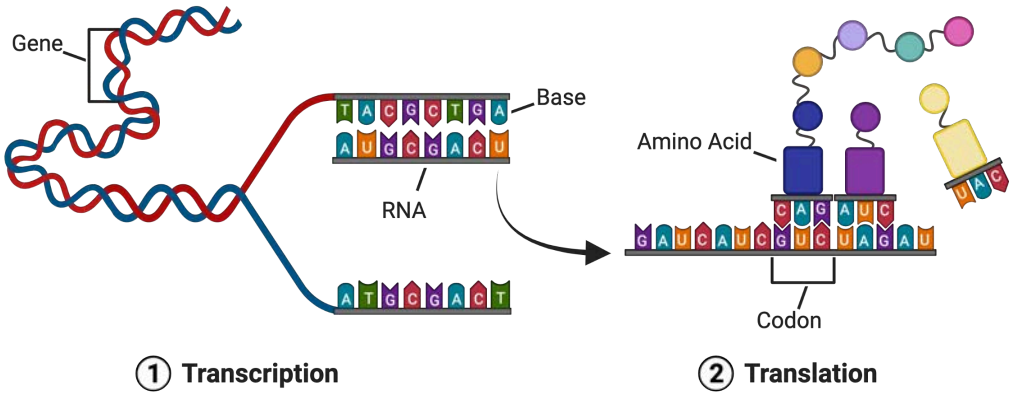


Fig. 2.2: Process of gene expression in a cell. 1 Transcription: the DNA is unwound and complimentary RNA is synthesized along one strand. 2 Translation: The complimentary RNA is translated into a protein by chaining together amino acids, which are encode by codons. Figure created with BioRender.com

encodes a protein as functional product, the created RNA transcript is *messenger RNA* (mRNA) that is further used during translation. If this is not the case, the RNA transcript is the final gene product and gene expression ends. The created RNA transcript is then used during other cell processes, e.g. as transfer or ribosomal RNA (t/rRNA) to help translating mRNA into a protein.

During **Translation**, mRNA is decoded to specify the amino acid sequence of a protein. This is carried out by a molecule called ribosome. The ribosome couples with the mRNA to synthesize a protein. For that, it reads triplets of bases, which are also called *codons*. Each codon defines a particular amino acid. To each codon, the ribosome binds an anticodon of complementary bases that has an amino acid attached. Anticodons of the same type always have the identical type of amino acid. Once an anticodon binds to a codon, its aminoacid is chained to the formerly bound anticodons to build the protein.

2.1.2 How Altered DNA Affects Gene Expression

Structural changes in the DNA can substantially affect the gene expression process. Structural changes, which are generally referred to as *genetic variants*, range from changes of a single nucleotide (*Single Nucleotide Polymorphism - SNP*), insertions or deletions of base sequences (*InDels*), to more complex structural changes like inversions or duplications. Depending on its location on a coding or non-coding region, a genetic variant can alter the quantity or actual functional product that is produced during gene expression.

If a coding region is affected by genetic variants, gene expression can produce an instable or even different functional product. When building RNA as a functional product, the created RNA is likely to loose its function as it can no longer bind to other molecules. This can have a negative regulatory effect on the expression of other genes. When building a protein as a functional product, a base change in a codon can lead to a different

amino acid being used. This alters the protein structure, resulting in an instable or completely different protein that is likely to be unusable in the original biological process. For example, mutations on the breast cancer (BRCA) 1 and 2 genes can increase the risk not only of breast cancer, but also of multiple other cancer types [61, 111]. BRCA1 and BRCA2 are tumor suppressor genes, i.e. they help repair DNA breaks that can lead to cancer and uncontrolled growth of tumors [260]. However, a mutation on the coding region of these genes can lead to a premature cessation of protein synthesis, having only produced the first part of the protein that cannot fulfill its original function.

Genetic variants located on non-coding regions can have a negative regulatory effect on gene expression. As non-coding regions provide binding sites for regulatory molecules, a sequence change can break the original binding site or create new binding sites for other molecules. For example, mutations on the promoter region of the telomerase reverse transcriptase (TERT) gene are correlated to multiple cancers [90, 245]. A promoter region of a gene is a short DNA sequence to which the RNA polymerase enzyme binds and subsequently initiates DNA transcription. Mutations on the promoter region of the TERT gene can generate new binding sequences for regulatory elements, which can upregulate TERT expression.

2.2 Gene Expression Profiling — Measuring Gene Activity

Identifying changes in gene activity, i.e. gene expression levels, leads to a better understanding of biological processes and their alteration in diseases. Modern technology employed in molecular biology enables *gene expression profiling*, i.e. measuring the activity of multiple thousand genes at once. Data from gene expression profiling is also referred to as *transcriptomics* data and generated either via Microarrays or RNA sequencing technology. Before transcriptomics data sets can be analyzed, they must undergo particular postprocessing steps. In the following sections, we describe the technologies and methods that are required to generate analysis-ready transcriptomics data sets.

2.2.1 Transcriptomics Data from Microarrays

Microarrays, also called DNA chips or gene chips, allow the expression level of a predefined set of genes to be measured simultaneously, up to multiple thousands. Microarrays used to be the standard method for gene expression profiling because they are cost-effective and have well-established protocols. However, the design of microarrays does not enable the detection of novel transcripts. In addition, microarrays are not very sensitive, making lowly expressed genes hard to detect.

Figure 2.3 depicts the principle behind microarray technology: First, paired samples of mRNA are collected from a reference cell and a cell that has the experimental condition. Second, each sample is converted into complementary DNA (cDNA) and labeled with

a fluorescent of a different color. Typically, reference samples receive green, and experimental samples receive a red color. Third, the created cDNAs of both samples are then mixed together and put on the microarray slide. The microarray slide contains many tiny spots, each of which has on it a particular known DNA sequence, typically from a gene. These are also called *probes* or *oligos*. The subsequent process which occurs on the microarray is referred to as *hybridization*: on each spot of a microarray, the samples' cDNA molecules try to bind to the probes. A binding only takes place if cDNA and probe are complementary, as only complementary base pairs create hydrogen bonds between each other. The more base pairs bind, i.e. are matched correctly, the stronger the bond between cDNA and probe. Afterwards, the microarray is washed so that only the cDNAs that have a strong binding to a probe remain. Fourth, the microarray is scanned

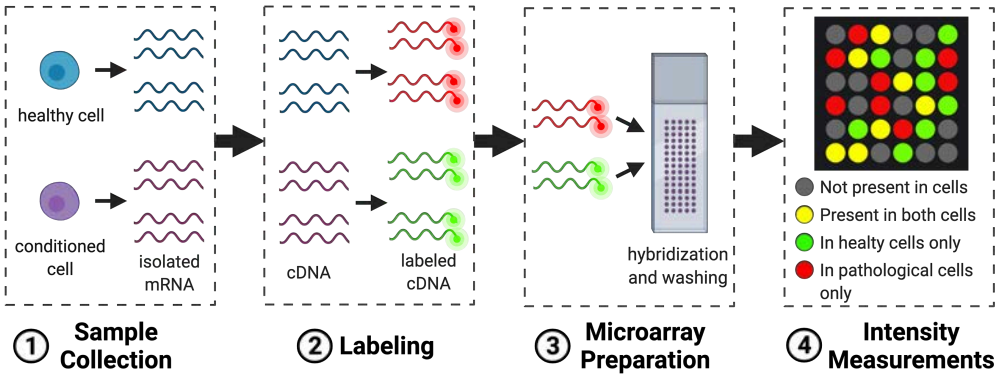


Fig. 2.3: Microarray processing steps: 1) paired mRNA samples are collected, 2) cDNA is created from mRNA samples and tagged with a fluorescent, 3) tagged samples are mixed on a Microarray, where hybridization and subsequent washing is carried out, 4) light intensities are measured. Figure created with BioRender.com.

to measure light intensities. The more cDNAs bind to a probe, the more intense the color, and the higher the expression level of the respective gene from which the original mRNA was taken. In our example, green would indicate that the reference sample has a higher expression level; yellow signifies no difference in gene expression between reference and experimental sample; red signifies a higher expression level of the experimental sample. The light intensities are then transformed into a computer-readable format and follow the postprocessing schema described in Section 2.2.3.

2.2.2 Transcriptomics Data from RNA Sequencing

RNA sequencing (RNAseq) uses sequencing technology to measure RNA quantity in a sample. In contrast to microarray experiments, which are limited to the measurement of a predefined set of RNAs, RNAseq can measure the full *transcriptome* of a cell, i.e. the complete set of RNA transcripts including coding and non-coding. Consequently, RNAseq generates huge data sets spanning many tens of thousands of genes. Compared

to microarrays, RNAseq is more robust and sensitive, as it allows the detection of lowly expressed genes. However, RNAseq is more cost-intensive and requires more complex computational analyses for subsequent processing.

Figure 2.4 depicts a schematic workflow of a sequencing run to create a transcriptomics data set, consisting of library preparation and amplification, sequencing, and alignment. For the sake of completeness, we focus here specifically on next-generation sequencing (NGS) technology as the currently most prominent sequencing technology used.

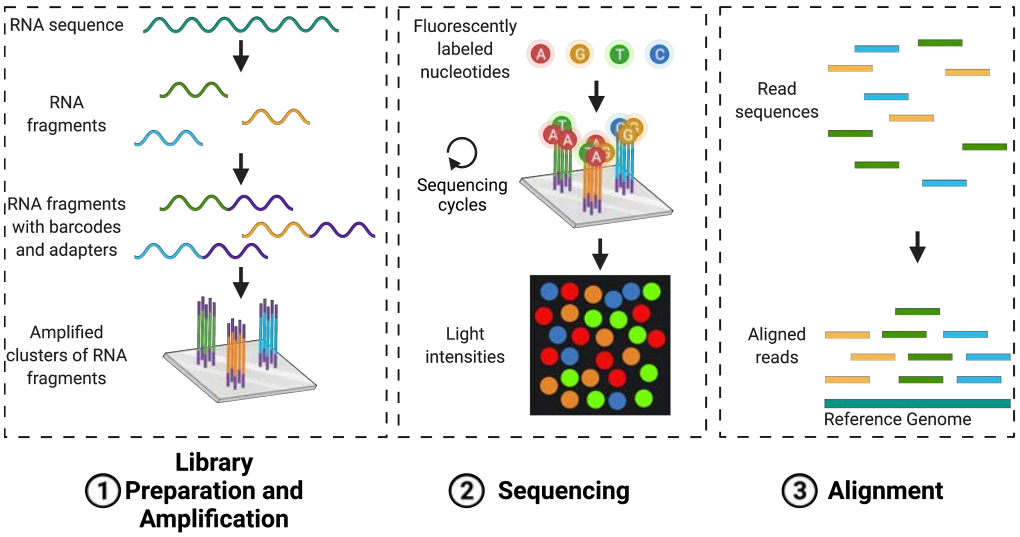


Fig. 2.4: Next-generation sequencing process on Illumina machines: 1) an RNA sequence is split into fragments that are labeled and amplified, 2) reverse strands are synthesized with colored nucleotides (one nucleotide per sequencing cycle) and light intensities are measured, 3) resulting read sequences are aligned to a human reference genome. Figure created with BioRender.com.

Library Preparation and Amplification

The RNA taken from a sample is first split up into thousands of fragments. From each RNA fragment, a cDNA sequence is created, along which a complementary strand will later be sequenced. Each cDNA sequence is then equipped with a custom barcode sequence for sample identification and an adapter sequence to bind to particular locations on the sequencing plate. Following this, the cDNAs are copied many thousand of times during amplification to yield a stronger signal.

Sequencing

cDNAs are sequenced with specialized sequencing machines. While the actual procedure and base detection differs between vendors, we shortly introduce the sequencing principles applied by the market leader for sequencing, Illumina. The sequencing machine

synthesizes the reverse strand one base at a time: first, nucleotides of all types are released to the cDNA strands. At Illumina sequencing, nucleotides are equipped with a fluorescent, each nucleotide type having a specific color. Second, the released nucleotides bind to their complementary nucleotides on the cDNA strand. Only a correct binding can create a strong bond. Third, the cDNA strands are washed to remove weak bindings, leaving behind only the correctly matched nucleotides. Fourth, the fluorescents attached to the matched nucleotides are then captured on camera: the strongest light signal determines the type of nucleotide. Third-generation sequencing technology joins the library preparation step with the actual sequencing and recognizes base types by sending them through a membrane that detects the relevant ions. This process is repeated until the complete strand has been sequenced. The output of the sequencing step is a file containing millions of RNA transcripts in a computer-readable format.

Alignment

Although the RNA transcripts have been detected during sequencing, it is not known to which gene they belong. Consequently, RNA transcripts are aligned to a reference genome, i.e. the blueprint of a human's complete DNA sequence. The expression level of a gene then corresponds to the number of RNA transcripts that have been aligned to the gene's region.

2.2.3 Transcriptomics Data Postprocessing

The results from RNAseq or microarray experiments are computer-readable data files containing the raw expression levels for multiple thousand genes. Table 2.1 shows an excerpt from a labeledexample gene expression file, which is typically encoded in CSV format. Gene expression levels are described in a large matrix of size mxn. The number of rows is determined by $m = |S|$, with S being the set of samples s_i. The number of columns is determined by $m = |G|$, with G being the set of genes g_i whose expression level was measured. The cells contain the actual expression levels e_{ij} for a sample s_i and gene g_j. The higher the value of e_{ij}, the more RNA transcripts from gene g_j were measured. For microarray data, e_{ij} corresponds to a light intensity as a floating point value that was measured by a camera and converted into numbers. For RNAseq data sets, e_{ij} is a discrete count value, i.e. it depicts the number of transcripts that were allocated to a region that is assigned to gene g_j.

The precise sizes of G and S vary depending on the data generation method and cohort size. For microarray data sets, G is a predefined set of genes for which the expression level is measured and depends on the applied chip set. For RNAseq data sets, G contains all genes to which the measured transcripts could be mapped, i.e. $|G|$ can currently be up to 65,000 genes.

The samples in S are separated into groups to test a hypothesis, i.e. there are control samples and samples with a particular condition. Samples can originate from different

Sample	TSPAN6	ICA1	WDR54	MARK4	POMT2	TRIO	PIK3CB
TCGA-3C-AAAU	4.5646	2.1123	3.6403	6.1926	4.4596	2.2539	4.4907
TCGA-3C-AALK	2.6000	3.8675	2.4137	7.1817	5.4446	0.2901	6.0221
TCGA-4H-AAAK	3.8736	7.8626	1.4673	6.4273	5.4832	6.4352	5.9835
TCGA-5T-A9QA	7.6018	6.1081	5.2293	7.5730	11.6279	4.4050	8.0847
TCGA-A1-A0SD	2.2634	0.9705	9.4098	8.3250	6.4352	8.1577	6.4521
TCGA-A1-A0SF	5.6545	6.4262	7.8105	5.5573	6.5429	2.5227	4.6756
TCGA-B6-A0I1	5.9587	5.8287	5.7448	4.9222	2.3237	1.7505	0.8385
TCGA-A1-A0SN	6.9936	2.6853	1.7796	5.9559	3.6377	8.1368	5.0545

Table 2.1: Excerpt from an example gene expression data set. The first row contains sample identifiers; subsequent columns contain the actual expression values for a particular gene, e.g. *TSPAN6*.

individuals, e.g. from patients with and without cancer, or from the same individual but from different tissue, e.g. from normal and tumor tissue. The assignment of samples to a particular group is delivered with a separate file containing sample meta data, e.g. diagnosis, gender, or survival time. The format of such meta data files, however, depends on the individual study and is not standardized. Once the gene expression data set has been created via microarray or RNAseq, it must be processed further to remove erroneous samples and genes and account for technical biases.

Filtering

Some genes and samples may have had errors in their processing, resulting in almost zero values or abnormally high expression levels. Such samples or genes must be removed from the analysis, as they would distort the subsequent analysis. Typically, genes and samples with a particular percentage (threshold) of zero values are removed. Sometimes, also samples with abnormally high expression levels get removed as well beforehand. Setting these thresholds is typically a manual process that is adjusted for each data set individually.

Accounting for Technical Biases

Gene expression data, whether generated via microarray or RNAseq, always represents a mixture of true biological variation and artifical variation introduced by technical bias. The primary objective during postprocessing of gene expression data is thus to remove the non-biological variation from the data by accounting for technical biases. According to Abbas et al., there are three types of technical bias that must be accounted for depending on the kind of subsequent analyses [1]: gene length, library size, and technical artifacts across samples.

While gene length is not relevant for microarray data, it must be accounted for in RNAseq data sets when subsequent analysis aims to compare expression levels between

different genes, i.e. applies intra-sample comparisons. The rationale behind is that the longer the sequence of a gene is, the more reads can be aligned to it. Popular approaches that account for gene length are Reads or Fragments per Kilobase per Million (R/FPKM) and Transcripts per Million (TPM) [136].

Accounting for library size is crucial when inter-sample comparisons are applied in subsequent analysis. RNAseq runs of samples differ in library sizes, i.e. sequencing depth. The higher the sequencing depth, the more reads are generated, which results in higher total counts for the entire sample. Consequently, read counts in each sample must be scaled by a sample-specific factor that reflects the library sizes. Methods like upper quartile (UQ), trimmed means of m-values (TMM), or relative log expression (RLE) are currently the peferred choices to account for library size [6, 31, 179]. According to Dillies et al., TMM and RLE are the most stable normalization methods [51].

Technical artifacts across samples, also referred to as batch effects, are caused by differing environmental circumstances when preparing batches for further processing, e.g. when batches are prepared in different labs, by different personnel, or different lab instruments. If such confounding factors are known in advance, they can be included in a model when accounting for library size. For small sample sizes, ComBat for microarray data and its extension ComBat-seq for RNAseq data can be applied separately [96, 263]. If confounding factors are not known in advance, they can be detected prior by applying remove unwanted variation (RUV), surrogate variable analysis (SVA), or principal component analysis (PCA) [115, 168, 176]. While these methods can also be used to remove the effects of the identified confounding factors, Abbas et al. recommend to not separately normalize the data set but rather incorporate these factors into the design matrix [1].

2.3 Biomarker Detection — Analyzing Gene Expression Data

Once the data set has been preprocessed, it is ready for the actual biomarker detection analysis. Biomarker detection on gene expression data can be modeled as a dimensionality reduction problem: the aim is to reduce noise and redundancy by identifying the most discriminative features, e.g. genes, that achieve best performance for a given task, e.g. survival predictions or distinguishing healthy samples from cancerous ones [112, 181]. Traditionally, the assessment of relevance of a gene is based on its expression levels in the data sets. Common approaches for biomarker detection are thus either differential expression analysis or standard feature selection — in this context, also known as *gene selection* — approaches. Throughout the remainder of this work, we will use the term *feature selection*, however interchangeably refer to features or genes that are selected by feature selection approaches.

2.3.1 Differential Expression Analysis

Differential expression analysis is often conducted for two-group comparisons and is beneficial for small sample sizes, when machine learning approaches tend to overfit. It can also be used as a preprocessing step before feature selection [80]. Differential expression analysis aims to identify differentially expressed genes in the data set. A gene is differentially expressed if it shows a statistically significant difference in its expression levels between conditions. For example, the BRCA1 gene shows a much higher expression level in cancerous samples when compared to healthy samples [260].

Differential expression analysis is thus a statistical analysis of gene expression data to discover quantitative changes in expression levels between experimental groups. Simply stated, the expression levels of a gene in one group are compared to the expression levels of that same gene in another group of samples. The analysis result contains information on the *fold change* of a gene, i.e. how much its expression level changes with respect to the other group. The fold change is defined by the ratio between the expression levels of the two groups. In addition to the fold change, each gene group comparison receives a *p-value* that indicates how likely the null hypothesis — i.e. that the expression of that gene does not change for the examined groups — is true. Genes having a low p-value are likely to be differentially expressed, with the fold change indicating the quantity of their up- or downregulation. For microarray data, Limma is the current state-of-the-art tool used for differential expression analysis [177]. For RNAseq data, DeSeq2 and EdgeR are widely used for differential expression analysis [124, 178].

2.3.2 Traditional Feature Selection Approaches

Traditional feature selection identifies discriminative features in gene expression data based on their statistical characteristics, also by examining feature dependencies. Literature classifies feature selection approaches into five categories: filter, wrapper, embedded, ensemble, and hybrid approaches [7, 82, 86, 94, 130].

Filter approaches rank genes according to a statistical measure, e.g. based on variance or Information Gain [45]. Filter approaches are widely used for their feasibility and usability [181]. They are low-complex measures that have an acceptable accuracy at a scalable performance. However, most filter approaches are univariate, i.e. they evaluate each feature separately without considering dependencies between them. As biological processes consist of gene interactions, univariate filter approaches cannot adequately reflect and identify the underlying biological processes in the data. In response, multivariate filter approaches like ReliefF also assess the relevance of a feature based on inter-feature dependencies [106].

Wrapper and *embedded* approaches provide more accurate results than filter approaches, but have a higher computational complexity. As more complex machine learning strategies are applied, wrapper and embedded approaches tend to be viewed as a computational black box by its users. Wrapper approaches, e.g. SVM-RFE or genetic algorithms,

interact with a classifier by iteratively creating multiple feature subsets, running the classification, and evaluating the results [78, 145]. Embedded approaches for feature selection are directly integrated into the learning algorithm that is used for subsequent classification, e.g. Random Forest and regularization approaches [50, 230].

Ensemble and *hybrid* approaches aim to combine the best characteristics of multiple feature selection approaches. Ensemble approaches run different feature selection approaches independently and combine their results into a final set of features [122, 258]. In turn, hybrid approaches combine feature selection approaches, e.g. a filter with a wrapper approach [116, 137]. In this way, ensemble and hybrid approaches exploit the advantages of both filter and wrapper approaches, leading to a higher accuracy than filter approaches and a computational feasibility better than for wrapper approaches.

2.3.3 Shortcomings of Traditional Approaches

The traditional approaches described base their decisions exclusively on data set characteristics. A gene is considered 'relevant' if its expression behavior shows a statistical significance in the data [7, 86]. However, statistical significance does not imply biological relevance. For example, oncogenes are not selected because they do not necessarily show a differential expression behavior; instead, they influence other genes along a signaling pathway [36, 262]. In turn, multiple genes from the same pathway can be selected because they all exhibit similar, statistically significant expression patterns [54, 117, 249]. This can, however, lead to a particular biological process being overrepresented in a set of selected genes, which introduces undesired redundancy.

A major obstacle for traditional feature selection approaches is the manifestation of random noise due to the high-dimensional data layout. Genes showing accidental correlation and not participating in a relevant biological process make machine learning approaches, in particular, likely to overfit. What is more, RNAseq data sets in particular suffer from what is referred to as random bias. A randomly selected gene signature exhibits true and robust, predictive power surpassing the expected value of its statistical significance [200]. As a consequence, the major shortcoming of existing feature selection approaches is their low robustness. This has been validated in multiple studies [48, 52, 53, 82, 88, 138, 163, 262]: two approaches applied on the same data set yield different gene signatures. In turn, applying an approach on a data set and then using the gene signature for analyzing another one will deliver less accurate results. The issues described call for integrative analyses that receive gene sets with an actual biological relevance and do not only rely on the statistical signals.

2.4 Annotating Gene Sets with Biological Knowledge

The results from traditional feature selection approaches are features, e.g. genes, that are selected as candidates because of their statistical relevance. Up to this point, however,

their biological context is ignored. Consequently, the biological and clinical relevance of biomarker candidates must be validated by annotating them with biological information. This can be achieved with automatic annotation and enrichment tools like EnrichR, DAVID, or Gene Set Enrichment Analysis (GSEA) [89, 215, 255]. These tools typically agglomerate knowledge from multiple biological knowledge bases.

Biological knowledge bases are online databases that provide curated biological knowledge. Biological knowledge comprises proven scientific findings about biological entities, e.g. genes, cells, or processes and their functions, interactions, and relationships to each other. In the context of feature selection on gene expression data, relevant biological knowledge focuses on genes and gene functions, signaling pathways, or gene-gene and gene-disease associations.

With the unremitting generation of biological knowledge, e.g. from large-scale studies, an increasing amount of public knowledge bases are now available. Even knowledge bases integrating information from multiple other knowledge bases are released at an increasing number. Depending on the type of information they provide, we group knowledge bases into *annotation*, *interaction*, and *meta knowledge bases*. However, not all knowledge bases are suitable for use in an automated fashion, especially not with regards to integrating them into the actual feature selection. Here, we provide a limited overview on available knowledge bases fulfilling the following criteria:

- *Acceptance*: The knowledge base is frequently used and well accepted by the general research community, e.g. for result set validation.
- *Data Access*: The knowledge base provides programmatic access or a data download.
- *Last Update*: The knowledge base contains the latest research results by being updated regularly, with the last update not older than 5 years.
- *Context*: The knowledge base contains genetic information.

2.4.1 Annotation Knowledge Bases

Annotation knowledge bases provide a comprehensive overview on genes and their products. They are typically organized in a structured format, e.g. as an ontology. Table 2.2 provides key characteristics of annotation knowledge bases that fulfill the aforementioned criteria. We group the annotation knowledge bases into those concentrating on functional information and trait associations.

Functional Information

Functional information about a biological entity, e.g. a gene, encompasses information that allows that entity and its characteristics to be described, e.g. structure, function, and interactions with others biological entities. Functional information can be used to identify functional similarities: for example, if two genes are annotated with terms of

Name	Content	Curation			Update
		man.	comp.	coll.	
Gene Ontology [226]	ontologies for cellular components, biological processes, molecular functions	•	•		monthly
UniProtKB [228]	functional information on proteins	•	•		monthly
Human Protein Atlas [237]	proteomic pathology, cell, tissue atlases	•		•	yearly
GWAS Catalog [128]	SNP-trait associations	•			weekly
COSMIC [60]	somatic mutations in cancer	•			3-monthly

Table 2.2: Online knowledge bases providing annotation information, e.g. gene functions. Abbreviations: manually (man.), computationally (comp.), collected (coll.).

similar semantics, they are likely to be similar in function and to participate in the same or similar biological process [12]. In addition, there is a strong correlation between the expression behavior of two genes and their functional similarity [246].

Amongst knowledge bases in general, *Gene Ontology (GO)* is a widely used knowledge base for annotating gene sets with functional information [10, 226]. GO provides a unified and machine readable representation of genes and their products. Three disjoint ontologies are provided: Cellular Component, Molecular Function, and Biological Process. Evidence for a relation between two genes or their gene product was identified either by human biocurators or a computational approach mimicking their behavior. A respective evidence code is assigned to the corresponding relation and therefore allows researchers to assess its reliability.

The *Human Protein Atlas (HPA)* maps all human proteins in cells, tissues, and organs [237]. HPA consists of three atlases that provide a comprehensive overview on genes and proteins: tissue, cell, and pathology atlas. The cell atlas depicts the subcellular localization of proteins in single cells. The pathology atlas describes the impact of protein levels in cells for survival of cancer patients. The tissue atlas shows a distribution of proteins across all major tissues and organs in the human body, which are most suitable for biomarker identification to incorporate tissue-selectiveness or -specificity of genes.

The *UniProt Knowledge Base (UniProtKB)* provides functional information on proteins and is frequently used for annotating gene sets [228]. UniProtKB aims to provide all known relevant information for a protein, e.g. gene and protein name, function, relevant protein-protein interactions, or expression patterns. UniProtKB consists of two parts: SwissProt and TrembL. SwissProt provides manually curated and reviewed records from

literature and computational analyses. TrembL, in turn, contains far more records, but without manual curation and peer review. As with GO, every evidence of a protein in UniProtKB is equipped with an indication of whether it was derived from experimental or computational analyses.

Trait Association

Trait associations encompass known relations of a biological entity, e.g. a gene, to specific traits, such as a disease. Trait information can be used to strengthen the statistical signal of genes that are known to be relevant in the particular context, e.g. by reducing noise in the data set. The manually curated catalogs on *Genome-Wide Association Studies (GWAS)* and *Somatic Mutations in Cancer (COSMIC)* provide gene-disease associations [60, 128]. Both have exhaustively reviewed literature on published GWAS and whole genome studies. While the GWAS catalog provides associations between SNPs and any traits in general, COSMIC concentrates on exploring the impact of somatic mutations in human cancer.

2.4.2 Interaction Knowledge Bases

Interaction knowledge bases contain information on any kind of interaction between genes, their products, or chemicals. These kind of knowledge bases are typically represented by graph-like structures. Table 2.3 provides key characteristics of interaction knowledge bases that fulfill the aforementioned criteria. We group the interaction knowledge bases presented here into those concentrating on gene co-expressions, pathways, and protein-protein or other interactions.

Gene Co-Expressions

Gene co-expressions are similar expression patterns across samples, e.g. a coordinated down- or upregulation. Genes that are co-expressed are likely to share similar functions and thus can be grouped together to identify redundancy. To date, there are two knowledge bases on gene co-expression in human tissue: *GeneFriends* and *COXPRESdb* [143, 172]. Both have created co-expression maps from public study data that show which gene (de-)activates other genes. GeneFriends additionally provides functional annotation and orthologous information on the regulation in other species, e.g. mice.

Biological Pathways

A biological pathway (pathway in the following) is a network of interactions among molecules that leads to a new molecular product or a change in a cellular state or process. Pathways play important roles in metabolism, gene expression, and signal transmission. Identifying altered pathways or submodules as biomarkers can reduce redundancy and increase robustness across data sets. Chowdhury et al. provide an in-depth discussion on

Name	Content	Curation			Update
		man.	comp.	coll.	
GeneFriends [44, 172]	gene co-expression		•		2021 (v5)
COXPRESdb [143]	gene co-expression		•		2021 (v8)
IntAct [85]	molecular interaction data	•			monthly
BioGRID [34, 208]	interactions of genes, proteins, chemicals	•			monthly
CTD [46]	associations of genes, proteins, chemicals to diseases	•			monthly
InnateDB [28]	interactions in innate immunity			•	weekly
REACTOME [43]	pathways and reactions, graph database	•			3-monthly
KEGG [101]	pathways	•			2-monthly
WikiPathways [202]	pathways	•			monthly

Table 2.3: Online knowledge bases providing interaction data, e.g. pathways or gene co-expressions. Abbreviations: manually (man.), computationally (comp.), collected (coll.).

existing human pathway knowledge bases [40]. Amongst these, the following knowledge bases appear to be the most suitable for feature selection:

The *Kyoto Encyclopedia of Genes and Genomes (KEGG)* and *REACTOME* are frequently used by the research community as they facilitate the understanding of signaling molecules and their reactions [43, 101]. Pathways of both knowledge bases are manually curated from literature reviews and grouped into specific sections, e.g. metabolisms or human diseases. KEGG's core module is the PATHWAY database that also contains a collection of disease pathways, e.g. for multiple cancer types. Both KEGG and REACTOME cross-refer to other online resources, e.g. UniProtKB.

WikiPathways is a collaborative approach for maintaining biological pathways [202]. Running on the same principles as Wikipedia, WikiPathways counts on community activity: instead of peer reviews, community members create, annotate, and change pathways. The fact that data entries are not peer reviewed makes its reliability questionable at first sight. However, WikiPathways is a specific knowledge base that is typically accessed by domain experts. Thus, it is likely to be less prone to false statements than Wikipedia, which has already proven to have a quality standard comparable to peer reviewed compendiums [67].

Interactions of Proteins and other Compounds

Protein-protein interactions (PPI) are physical contacts between single proteins that serve a specific function, e.g. as part of a pathway or by connecting two pathways. PPIs can be combined to form larger PPI networks that offer similar advantages as pathways. However, there are also other factors, such as environmental influences or chemicals, that can affect genes or proteins. Those interactions are typically curated manually from existing literature, e.g. study publications, and annotated with additional information from other knowledge bases to provide a complete view.

IntAct specializes in protein-protein interactions in humans and other species and provides further annotations [85]. *BioGRID* includes interaction data between genes, proteins, or chemicals in humans and also other model organisms [34, 208]. The *Comparative Toxicogenomic Database (CTD)* concentrates on studying the environmental components which influence human health and, therefore, provides associations between chemicals, genes, and diseases [46]. *InnateDB* originally focused on the innate immune response of humans and other organisms [28]. A full-time curator team integrates data from scientific publications to enable a system-level analysis by providing genes, proteins, experimentally-verified interactions, and signaling pathways.

2.4.3 Meta Knowledge Bases

Meta knowledge bases do not provide self-created information, but compile findings from multiple existing knowledge bases. Meta knowledge bases either gather information on a specific issue from similar knowledge bases, e.g. from existing pathway databases, or they aim to provide a system-level view of the respective topic by accessing heterogeneous data sources, e.g. from text, RNA, structured information, or interaction graphs. All of them provide some sort of ranking criteria for the knowledge bases they incorporate. Table 2.4 provides key characteristics of meta knowledge bases that fulfill the aforementioned criteria. We separate the currently existing meta knowledge bases into those providing information on pathway and protein interaction networks, and those providing target-disease associations.

Pathways and Protein Interaction Networks

PathwayCommons and *ConsensusPathDB* integrate a wide range of public pathway and interaction databases [33, 99]. Data in PathwayCommons represents biochemical reactions, gene regulatory networks, genetic interactions, transport and catalysis events, and physical interactions of components ranging from proteins to small molecules and complexes [47]. From the knowledge bases listed here, ConsensusPathDB integrates REACTOME, BioGRID, IntAct, KEGG, and WikiPathways, whereas PathwayCommons additionally includes CTD and UniProtKB.

The *STRING* database concentrates on protein-protein interactions [219]. The data originates from experiments, literature, and existing knowledge bases, but also from

Name	Content	Curation			Update
		man.	comp.	coll.	
Pathway-Commons [33]	KEGG, BioGRID, REACTOME, IntAct, WikiPathways, CTD, and others			•	2019 (v12)
Consensus-PathDB [99]	KEGG, BioGRID, REACTOME, IntAct, WikiPathways, InnateDB, and others			•	2021 (v35)
STRING [219]	GO, KEGG, REACTOME, IntAct, BioGRID, and others			•	2021 (v11.5)
DisGeNET [165]	UniProtKB, GWAS, CTD, and others			•	2020 (v7.0)
Open Targets [108]	UniProtKB, GWAS, REACTOME, IntAct, Cancer Gene Census, and others			•	2022 (v22_02)
Entrez Gene [129]	Entrez databases	•		•	2022

Table 2.4: Online meta knowledge bases that aggregate biological information from annotation and interaction knowlegde bases. Abbreviations: manually (man.), computationally (comp.), collected (coll.).

STRING-specific predictions. For every interaction, STRING offers a range of evidence scores, e.g. for text, expression, experiments, or knowledge bases, for prioritization. Amongst others, STRING currently integrates BioGRID, KEGG, REACTOME, IntAct, and GO.

Target-Disease Associations

DisGeNET provides scored associations between genes, variants, and diseases [165, 166]. Users can search and receive a ranked list of associated genes, diseases, variants, and their associations. DisGeNET incorporates knowledge from curated databases, some of which are presented here, but also from other databases on genetic variants, e.g. dbSNP, animal models, or literature mining tools [199]. DisGeNET defines scores for gene-disease and variant-disease associations, but additionally defines more specific indices for evidence, specificity, and disease pleiotropy.

Open Targets is an initiative from multiple institutions to identify targets for effectively treating diseases [108]. A target can be an RNA molecule, protein, or protein complex. The platform aims to help researchers to find and prioritize targets for further investigation. Starting with a disease or a target, Open Targets provides information on respective associations thereof, clearly differentiating evidence origins, e.g. text mining, RNA expressions, or genetic associations. Open Targets currently integrates a wide

range of knowledge bases, amongst them COSMIC, IntAct, REACTOME, GWAS, and UniProtKB.

2.4.4 Programmatic Access to Knowledge Bases

All knowledge bases provide bulk downloads of their data for offline processing in multiple basic data formats: character-separated, XML, or JSON. There are multiple formats based on the above for describing gene annotations and molecular pathways. For gene annotations, GO provides its own format with the Gene Annotation File (GAF), while other knowledge bases use the Gene Matrix Transposed (GMT) format [175, 225]. For molecular pathways, common standards are the Systems Biology Markup Language (SMBL), BioPAX, and the Proteomics Standards Initiative's (PSI) XML-based and tab-delimited formats for Molecular Interactions (PSI-MI/PSI-MITAB) [47, 59, 103, 213]. KEGG defines its own KEGG Markup Language (KGML) for describing pathways, WikiPathways applies the GenMAPP Pathway Markup Language (GPML); other knowledge bases use the Simple Interaction Format (SIF) [93, 102, 224].

Nearly all knowledge bases provide interfaces for programmatic access: most of them offer a RESTful Application Programming Interface (API). Some chose to provide SOAP/WISDL access. Few resources rely on SPARQL endpoints or provide custom solutions. Some knowledge bases provide example Python or R scripts and custom packages for data access, e.g. KEGGREST or UniProt.ws [32, 221]. Many interaction databases are accessible via PSICQUIC [9]. It offers standardized access through its own web service and query language to query multiple interaction databases at the same time and is accessible via corresponding R and Python packages. Python's Bioservices does not only integrate PSICQUIC but provides a unified interface to further external knowledge bases — including BioGRID, KEGG, OmniPath, Pathway Commons, REACTOME, and UniProtKB [42]. It can be further extended by any other RESTful API. BioMart also offers both a standalone installation and a Web Service for accessing data sets from multiple sources, e.g. COSMIC or UniProtKB [203]. OmniPathdb aims to provide a comprehensive collection of literature-curated human signaling pathways [235]. It contains descriptions and annotations on nearly every existing pathway database in order to assist researchers in selecting the most suitable database for their analysis. OmniPathdb comes with a Python module named pypath that provides a machine-readable representation for pathways [235]. Many R packages also provide indirect access to the respective knowledge bases by invoking functions for gene set enrichment or functional annotation[3].

[3] A general overview can be found at BioConductor's package overview, section *Annotation Data*: `http://bioconductor.org/packages/release/BiocViews.html`

Name	Data Type				Access		
	*SV	XML	RDF	JSON	downl.	REST	other
Gene Ontology [226]	GAF		•	•	•		BioLink, GOlr
UniProtKB/ SwissProt [228]	•	•	•		•	•	SPARQL
Human Protein Atlas [237]	•	•	•	•	•	•	
COSMIC [60]	•				•		
GWAS Catalog [128]	•		•		•	•	
GeneFriends [44, 172]	•				•		
COXPRESdb [143]	•				•	•	
IntAct [85]	PSI-MITAB	PSI-MI			•		PSICQUIC
BioGRID [34]	PSI-MITAB	PSI-MI		•	•	•	PSICQUIC
CTD [46]	•	•			•		batch queries
InnateDB [28]	PSI-MITAB	PSI-MI, XGML	BioPAX		•		PSICQUIC
REACTOME [43]	PSI-MITAB	SBML	BioPAX		•	•	PSICQUIC
KEGG [101]		KGML			•	•	
Wiki-Pathways [202]	GMT	GPML	•		•	•	SPARQL
Pathway-Commons [33]	SIF, GMT		BioPAX		•	•	SPARQL
Consensus-PathDB [99]	•	PSI-MI			•		SOAP/ WSDL
STRING [219]	PSI-MITAB	PSI-MI		•	•	•	
DisGeNET [165]	•		•		•	•	SPARQL
Open-Targets [142]	•	•		•	•		GraphQL, BigQuery
Entrez Gene [129]	•				•		E-Utilities

Table 2.5: Available data formats and endpoints for accessing online knowledge bases. Most of the knowledge bases allow to programmatically query their databases, e.g. via a REST API.

2.5 Summary

This chapter introduced the domain of biomarker detection on gene expression data. Gene expression is the process of reading a DNA sequence and building a functional product from it. A functional product can be either a protein, which drives cell processes, or tRNA, which is used to regulate the expression of other genes. In diseases such as cancer, the expression of many genes — and as such the produced functional products, are altered and consequently change cell processes, e.g. on cell growth. Therefore, research aims to identify disease-specific expression profiles, i.e. biomarkers, that allow to diagnose diseases or predict treatment outcomes.

In order to identify biomarkers, gene expression is measured either via RNAseq or Microarray to produce expression data sets in a computationally readable format. These data sets are then further processed, e.g. via normalization, and subsequently analyzed, typically via differential expression analysis or feature selection approaches, to find biomarkers. Finally, the biological relevance of biomarkers is only assessed at the very end of the analysis process, where proven biological knowledge from public data bases, i.e. knowledge bases, is used for annotation and enrichment. There is a wealth of — most often manually curated — knowledge bases that provide prior knowledge ranging from functional information, gene-disease associations, to topological information like protein-protein interaction networks. In the last years, many meta knowledge bases have evolved, which aggregate prior knowledge from many of the existing knowledge bases. However, existing approaches for biomarker detection, and feature selection in particular, have multiple shortcomings regarding redundancy in the identified biomarkers, their robustness across data sets, and the propagation of random noise due to small sample sizes — which research assumes to be solved by incorporating the biological context earlier into the analysis.

3

Related Work

This chapter presents work related to two research fields: biomarker detection with prior biological knowledge and the automated, flexible evaluation of analysis strategies for omics data sets. We review and discuss both existing prior knowledge approaches and benchmarking solutions for the analysis of omics data sets. For each research field, we first present an overview of existing approaches, and subsequently discuss current limitations.

3.1 Biomarker Detection Using Prior Knowledge

The approaches described in this section constitute a selective overview of the current state of the art when it comes to analyzing gene expression data sets and incorporating prior biological knowledge, e.g. from a knowledge base or similar sources. We particularly focus on feature selection and module extraction approaches. We do not consider multi-omics approaches that integrate multiple data artifacts from the same individuals, e.g. gene expression and methylation data.

3.1.1 Prior Knowledge Approaches — State of the Art

Figure 3.1 depicts which integration strategies are typically applied by prior knowledge approaches to incorporate prior biological knowledge into the analysis. Prior biological knowledge is used in three forms. First, as a simple list of entities, e.g. genes or annotations, that are in some way associated to a search term, e.g. a disease. Second, as a list of entities with corresponding relevance scores, e.g. genes and their association scores. Third, as some kind of topological interaction information, e.g. gene-gene interactions, pathways, or larger composed networks.

A *list of related entities*, is typically used to filter or extend a feature set. Filtering can be applied either before or after traditional feature selection [58, 98, 198]. For example, Shao and Conrad manually created an Epithelial Mesenchymal Transition (EMT) network from literature and used it for filtering a cancer data set before applying Lasso feature

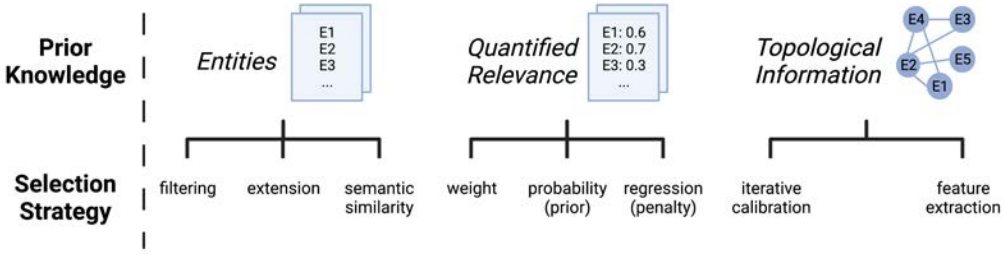


Fig. 3.1: Prior knowledge is generally available in three different forms, for which different integration strategies are typically applied by prior knowledge approaches. Figure created with BioRender.com.

selection [198]. Such filtering approaches are simple and can also reduce the runtime of subsequently applied traditional approaches. However, filtering approaches also prevent the discovery of previously unknown associations. To address this issue, approaches like Biological Pathway-based Feature Selection (BPFS), SoFoCles, or RelSim carry out traditional feature selection first and (iteratively) extend that list by features from prior knowledge afterwards [15, 132, 148]. If the partially-associated entitites are gene annotations rather than genes, then this information can be used to incorporate the notion of semantic similarity into the selection process. The actual integration strategies differ in their processing steps and range from computing scores for annotation terms from annotated genes, creating and ranking GO term paths, clustering GO terms, to adapting Google's PageRank algorithm [2, 3, 18, 133, 135, 169, 185, 234].

Available *relevance information* for features, e.g. association scores, can be incorporated using a mathematical framework. The simplest form of a mathematical framework applied in this context is a combined score from statistical data characteristics, e.g. expression behavior across samples, and biological relevance from prior knowledge, e.g. gene-disease association scores [75, 98, 171, 244]. More advanced approaches incorporate prior knowledge via Bayesian priors or penalty terms to apply regression or regularization methods, respectively [21, 68, 117, 153, 207, 209, 210, 216, 247, 257, 261, 265].

Topological information, such as gene-gene interactions or networks, can be intertwined with statistical data characteristics during an iterative selection process. As such, protein-protein interactions (PPIs) can be used to recalibrate gene relevance scores based on the scores of interaction partners or to compute correlation scores [127, 149]. Other approaches map genes, gene clusters (with similar expression values), gene correlations, or gene correlation clusters to PPI networks to identify hub genes [8, 131, 217, 218]. Feature extraction approaches carry the notion of integrating external knowledge even further by replacing genes by (sub-)networks as features. This requires two steps. First, identifying relevant pathways or (sub-)network modules as features. Second, mapping the original genes to the new feature space. Identifying pathways or (sub-)networks as potentially new features can be achieved by examining the expression levels of their mem-

ber genes. For example, pathways or network modules can be ranked based on the mean expression values of their member genes, or the correlation thereof, to class labels [113, 170]. Other approaches examine the enrichment with differentially expressed genes or compute principal components for a pathway or network module [38, 76]. New feature values are then computed via different strategies: some use the mean or normalized sum of the expression values of all genes that are contained in the feature [74, 76]. More advanced approaches use significance values, e.g. a p-value, for the correlation between the expression levels of member genes and class labels [41, 113, 262]. Other approaches use logistic regression or single-sample Gene Set Enrichment Analysis (ssGSEA) to compute new feature values [5, 16, 170].

3.1.2 Insights

We conducted a quantitative comparison of 47 existing prior knowledge approaches (see Table 10.12 in the appendix), which reveals multiple issues with the current state of the art and thus further motivates the contributions of this thesis. In the following, we summarize major findings and discuss practical implications.

Prior knowledge approaches are not generally available to the research community. From the approaches examined, only 16 percent provide the actual source code or an application. The sources of a further seven percent are no longer available, whilst pseudo-code is provided for seven percent of the approaches. For the remaining 70 percent, only their methodological approach description exists in the corresponding publication. Unavailable source code is a major barrier for a widespread application. On the one hand, results cannot be reproduced. On the other hand, the approach cannot be tested and compared to other approaches in a broader context. Ultimately, approaches with unavailable sources will not be adopted by the community for use in practice.

Prior knowledge approaches are not flexible. Only 16 percent of the examined approaches allow for custom adjustments, e.g. changing the applied knowledge base, statistical approach, or other configurations. Less than ten percent of the approaches were built with the intention of serving a general purpose and to be applied to different use cases. Nearly thirty percent of all approaches require considerable effort, either a) in advance, to transform the expression data or prior knowledge, or b) to adapt them for other use cases. Still, approaches integrating topological network information are the most flexible in regards to the prior knowledge input, as they typically only require some kind of network as input. The inflexibility of prior knowledge approaches makes it hard for them to be integrated seamlessly into other workflows, which is yet another obstacle to their broader application in practice.

Prior knowledge approaches are seldom cross-validated across data sets, although research suggests an improved robustness. From the approaches reviewed, less than 14 percent apply a cross-validation across data sets. Two thirds of the approaches predominantly use traditional cross-validation methods, e.g. k-fold or holdout

methods. But the remaining third of the approaches still does not apply any cross-validation strategy during evaluation. A thorough cross-validation strategy is essential for prior knowledge approaches — as it is in general for biomedical applications — to prove the robustness of approaches, as research expects integrative approaches to deliver more robust results [19, 151]. Consequently, the validation practice in the current research is insufficient, as it does not provide reliable insights into the actual robustness of the approaches.

Demonstrated improvements of prior knowledge approaches are often only relative assessments with limited validity. For three thirds of the approaches examined, their effectiveness is assessed by comparing them solely with traditional approaches. Around 27 percent of the approaches examined are compared with at least one other prior knowledge approach, while less than twenty percent are compared with both traditional and prior knowledge approaches. These results show that the findings drawn from these evaluations are only valid within a limited scope: demonstrated improvements are only relative assessments, e.g. proving that a prior knowledge approach that modifies a traditional approach supersedes it.

3.2 Benchmarking in Bioinformatics

Benchmarking is essential to show the effectiveness of existing and newly developed approaches in a broader context. Benchmarking results allow important conclusions to be drawn regarding the practicability, usefulness, reliability, and robustness of approaches. For both traditional approaches and differential expression analysis, there are already multiple benchmarking studies [7, 14, 48, 82, 138, 163, 173, 182]. However, no such benchmarking studies exist for prior knowledge approaches. This is mainly due to two issues with prior knowledge approaches: they are not available for practice and extensive effort is required to set up such a benchmarking experiment. Execution pipelines that include steps for preprocessing, feature selection, classification, and subsequent result set assessment must be implemented. Furthermore, for prior knowledge approaches, knowledge bases must be accessed and their output must be matched, e.g. to receive the correct gene identifier format.

3.2.1 Benchmarking Tools — State of the Art

The idea of reproducible benchmarking in bioinformatics has gained increasing attention recently. We have examined related works in this research area and summarized them according to selected criteria in Table 3.1.

Numerous efforts are being made to define data sets for benchmarking via high-quality curation or simulation strategies. VariBench and the Genome in a Bottle (GiaB) consortium provide gold standards from real-world data for analyzing genetic variants, e.g. for variant detection or effect prediction [139, 183, 267–269]. GeneSetBenchmark and

RNAontheBENCH constitute two alternatives for benchmarking gene expression data in the context of differential expression analysis and network derival [4, 66]. Other approaches like SimBA or GeneNetWeaver concentrate on simulating gene expression data sets, e.g. from gene regulatory networks or transcript annotations [11, 187].

Besides providing gold standards for benchmarking, the analysis results need to be assessed with standard evaluation metrics. Such tools typically compute standard evaluation metrics and create visualizations from the results; some of them also support the evaluation of the intermediate pipeline results [11, 187, 205, 212].

While the aforementioned approaches concentrate on benchmark data sets and assessing the analysis outcomes, they do not address the actual design and execution of the benchmark experiment. There are multiple R packages that can be applied to execute gene expression analysis pipelines automatically, some of them enabling the choice from a specified set of approaches, e.g. for normalization, feature selection, or classification [39, 49, 211, 253]. Frameworks of more general scope allow experiment pipelines to be flexibly designed and individually configured, and even enable the comparison of results from nested pipeline variations [65, 105, 214]. Such frameworks only provide the underlying pipeline orchestration, which makes them applicable to nearly any use case. They are, therefore, the most flexible and allow custom implementations to be included. However, the functionality for all pipeline steps must be provided and dealt with by the user, e.g. identifier mapping or visualizations.

There are only a few benchmarking suites that cover all aspects of benchmarking, ranging from providing gold standards to pipeline execution and evaluation [20, 64, 167, 231]. They all provide a choice of state of the art methods and support the full benchmarking process.

3.2.2 Insights

The publication years of the reviewed approaches clearly indicate that reproducible benchmarking is an emerging topic of research, as all approaches have been published since 2015. However, there are still opportunities for improvement.

Benchmarking biomarker detection approaches for gene expression data sets is not fully supported. While there are predominantly approaches for gene regulatory network inference methods that support the full benchmarking process, the complete benchmarking process is sparsely covered in other areas. General purpose approaches like SummarizedBenchmark, pipeComp, or CellBench are the most flexible in their application, although they do not provide the necessary administrative functionality, e.g. identifier mapping or visualizations [65, 105, 214]. Instead, users need to provide the complete code of applied methods on their own. Only few approaches are actually extensible and allow custom approaches to be added [20, 105, 167].

Name	**Year**	**Domain**	**Supported Aspects**				**Extensibility**
			Gold Standard	Pipeline Design	Pipeline Execution	Result Evaluation	
RNAontheBENCH [66]	2016	RNAseq and DEA	•				
Genome in a Bottle [267]	ongoing	genetic variation	•				
VariBench [139]	2013	genetic variation	•				
GeneSetBenchmark [4]	2014	gene expression, pathways	•				
SimBA [11]	2017	RNAseq	•			•	
GeneNetWeaver [187]	2011	gene expression, regulatory networks	•			•	
QPEP [212]	2017	mass spectometry data preprocessing				•	
iCOBRA [205]	2016	ranking comparison, binary assignments, e.g. DEA				•	
ClassifyR [211]	2015	gene expression feature selection and classification		(•)	•	•	
DaMiRseq [39]	2017	gene expression feature selection and classification			•	•	
OmicsMarkeR [49]	2015	omics feature selection and classification		(•)	•	•	
NormalyzerDE [253]	2015	gene expression normalization, DEA		(•)	•	•	
SummarizedBenchmarks [105]	2019	general purpose		•	•	•	•
CellBench [214]	2020	general purpose		•	•	•	
pipeComp [65]	2020	general purpose		•	•	•	
GeneSPIDER [231]	2015	GRNI methods	•		•	•	
BEELINE [167]	2020	GRNI methods	•		•	•	•
NetBenchmark [20]	2015	GRNI methods	•		•	•	•
GSEAbenchmarkeR [64]	2020	GSEA methods	•		•	•	

Table 3.1: Review of existing approaches related to benchmarking omics data sets. Approaches cover a) data set curation or generation to provide benchmarking data sets, b) flexible pipeline design, c) automated pipeline execution, and d) assessment of evaluation results. Few approaches were designed to be extended by custom functionality. Approaches with (•) at *Pipeline Design* mean a limited flexibility, i.e. that users can choose between a limited number of approaches. Abbreviations: Gene Regulatory Network Inference (GRNI), Differential Expression Analysis (DEA), single-cell RNAseq (scRNAseq), Gene Set Enrichment Analysis (GSEA).

The specific needs of benchmarking prior knowledge approaches are currently not addressed. Prior knowledge approaches imply higher implementation efforts, as they require additional external information. None of the benchmarking approaches provide access to any kind of biological knowledge, but rather concentrate on quantitative performance metrics, e.g. classification accuracy. Only one of the approaches presented supports general cross-validation strategies [49]. However, unified access to multiple knowledge bases and cross-validation strategies, especially across data sets, are essential for benchmarking prior knowledge approaches.

3.2.3 Implications for the Contributions of this Thesis

The insights derived from the quantitative reviews of both prior knowledge approaches and benchmarking tools are correlated: an extensible benchmarking tool that addresses the specific needs for prior knowledge approaches can improve their general availability and enables the design of flexible solutions. Increased availability and flexibility of prior knowledge approaches will facilitate their broader application, which will subsequently promote more benchmarking studies. These can be executed with a corresponding benchmarking tool, which allows asking broader questions, e.g. regarding the impact of the chosen knowledge base and integration strategy on biomarker robustness. The contributions of this thesis address these needs in three steps. First, we identify common integration concepts that are applied by prior knowledge approaches. Second, we implement selected integration concepts in a benchmarking tool. Third, we use the developed tool for a benchmarking study to examine the effectivity of prior knowledge approaches.

3.3 Summary

This chapter described relevant work on both prior knowledge approaches and benchmarking systems for assessing the effectiveness of omics analysis methods. Prior knowledge approaches generally apply different strategies to incorporate prior knowledge. As such, they can use a) lists of entities, e.g. genes or annotation terms, to adapt an existing gene set or use similarity measures to group semantically similar genes, b) relevance information, e.g. gene-disease associations, to use them as additional weight, prior, or penalty term, and c) topological information, e.g. protein-protein interaction networks, for iterative score recalibration or feature extraction strategies. However, most of the approaches presented here are either not available to the general public or are not flexible enough to adapt for different use cases. This leads to a currently insufficient assessment of the effectiveness of prior knowledge approaches that leaves many questions unanswered, e.g. on the actual robustness. In turn, there are many tools for benchmarking omics data sets and computational analysis methods, e.g. by providing benchmark data sets or built-in standard evaluation measures. However, no such tool exists that a) supports the complete analysis process for biomarker detection on gene expression data sets and b) explicitly meets the needs for prior knowledge approaches.

4

Making Prior Knowledge Approaches Flexible: Defining Prior Knowledge and Integration Strategies

This chapter deals with the conceptual description of both prior biological knowledge and prior knowledge approaches. In order to be able to incorporate prior knowledge into feature selection, we need to identify what kind of prior knowledge can be combined with gene expression data and subsequently specify a joint definition thereof. Consequently, we derive different levels of prior knowledge from reviewing existing knowledge bases. We then describe common integration strategies for the different levels, and further clarify under what assumptions prior knowledge can be transformed to a level that is suitable for a particular integration strategy. Finally, we present novel generalized concepts of prior knowledge approaches that apply the defined prior knowledge levels.

4.1 A Conceptual Definition of Prior Biological Knowledge for Gene Expression Data

Prior biological knowledge has already been applied in a multitude of feature selection approaches on gene expression data [98, 148, 171]. Most of these approaches incorporate prior knowledge that is manually prepared for a particular use case and consequently do not provide a definition of prior knowledge that is more generally applicable. The few approaches that provide a more formal description define and distinguish prior knowledge primarily based on the type of resource that is provided [68, 207, 264]. For example, Zhao et al. define three types of prior biological knowledge that can be incorporated into feature selection [264]: gene similarities, gene functions, and gene interactions. However, the focus of such definitions is rather on differentiating prior knowledge into the kind of provided information instead of finding a joint definition for automated prior knowledge retrieval from online resources.

In the scope of this work, we therefore define prior biological knowledge as curated information on biological entities and processes and particularly concentrate on prior knowledge that is available in public online knowledge bases and can be automatically retrieved from there. Prior knowledge, however, can only bring added value to an analysis if it connects to the limited biological context inherently present in a gene expression

data set: this includes the features, i.e. genes, and the use case context, e.g. if we want to classify samples into particular disease (sub-) types. We therefore focus exclusively on the kinds of prior knowledge that provide such connection points. In Chapter 2, we already presented online knowledge bases whose prior knowledge appears suitable for integration into feature selection on gene expression data. Subsequently, we analyze these knowledge bases regarding their expected input and delivered output to derive a conceptual notion of prior knowledge (see also Table 10.1 in the appendix for an overview).

In general, every knowledge base requires some kind of input for which related prior knowledge is provided. Thus, we define the process of prior knowledge retrieval as an individual function per knowledge base kb

$$f_{kb} : qt \rightarrow pk_{kb}^{qt}, \text{ for } qt \in QT, pk_{kb}^{qt} \in PK \tag{4.1}$$

that maps a given query term qt to prior knowledge pk_{kb}^{qt} that is part of PK, which we will define in more detail later. In the scope of this work, the set of applied query terms QT can contain gene identifiers from the set of human genes G, but also other biological search terms, e.g. a disease name. The retrieved prior knowledge pk_{kb}^{qt} contains biological information from knowledge base kb that shows — according to the internal search strategy of kb — a relation to the query term qt. While pk_{kb}^{qt} can contain a lot of information that does not necessarily connect to the restricted biological context in a gene expression data set, we limit our considerations here to the kind of prior knowledge that can be used for our purposes. We also do not examine the heterogeneous data formats in which prior knowledge can be returned by a knowledge base (see also our review in Table 2.5), as this issue is related to implementation instead of conceptualization.

We define prior knowledge PK as a set of biological information pk that was retrieved from a knowledge base $kb \in KB$ for a query term $qt \in QT$

$$PK = \{pk_{kb}^{qt} : qt \in QT \text{ and } kb \in KB\} \tag{4.2}$$

From a conceptual point of view, pk_{kb}^{qt} can take three forms based on the provided information content:

$$pk_{kb}^{qt} \begin{cases} pk1_{kb}^{qt}, \text{ a set of entities} \\ pk2_{kb}^{qt}, \text{ a set of scored entities} \\ pk3_{kb}^{qt}, \text{ a set of gene-gene interaction networks} \end{cases} \tag{4.3}$$

$pk1_{kb}^{qt}$ constitutes first-level prior knowledge, which is a **set of entities** a_i retrieved for a query term qt

$$pk1_{kb}^{qt} = \{a_i\}_{i=1...|M|}, \text{ with } a_i \in M, M \subseteq G \text{ or } M \subseteq D \text{ or } M \subseteq A \tag{4.4}$$

The entities a_i are either a set of genes from G, diseases from D, or annotation terms from A.

$pk2_{kb}^{qt}$ constitutes second-level prior knowledge, which is a **set of scored entities**

$$pk2_{kb}^{qt} = \{(a_i, rs_i)\}_{i=1...|M|}, \text{ with } a_i \in M, M \subseteq G \text{ or } M \subseteq D \text{ or } M \subseteq A, rs_i \in \mathbb{R}^+ \tag{4.5}$$

that contains tuples of an entity a_i and a relevance score rs_i. a_i can be of the same type as first-level prior knowledge, e.g. genes. The relevance score rs_i for an entity a_i quantifies how strongly it is related to the query term qt. rs_i is provided by the knowledge base and typically based on the number and reliability of evidences found.

$pk3_{kb}^{qt}$ constitutes third-level prior knowledge and is a **set of gene-gene interaction networks**

$$pk3_{kb}^{qt} = \{n_i\}_{i=1...|N|}, n_i \in N, n_i = (V_i, E_i), V_i \subseteq G, E_i \subseteq G \times G \tag{4.6}$$

where each network n_i from the set of retrieved networks N consists of a set of genes as nodes V_i and a set of gene-gene interactions as edges E_i.

The level of prior knowledge correlates with the provided information content: while first-level prior knowledge contains a list of entities that are generally related to the query term, second-level prior knowledge additionally provides scores that quantify these relationships. Third-level prior knowledge, i.e. networks, even provides topological information and relations between single entities. Consequently, higher levels of prior knowledge have a higher information content and can be applied to more complex integration strategies during biomarker detection.

4.2 Transforming Prior Knowledge Levels

Typically, a knowledge base provides only one level of prior knowledge. For example, an annotation knowledge base like COSMIC provides disease genes as first-level prior knowledge, a meta knowledge base like Open Targets provides gene-disease associations as second-level prior knowledge, and an interaction knowledge base like PathwayCommons provides pathways as third-level prior knowledge [33, 60, 108]. Consequently, not all knowledge bases are suitable by default for all kinds of prior knowledge approaches. It is, however, possible to transform their default level of prior knowledge into a higher or lower level. To make knowledge bases usable for a wide range of prior knowledge approaches, we describe transformation strategies for the different levels of prior knowledge and discuss potential constraints in the following subsections.

4.2.1 Transforming Higher-Level Prior Knowledge into Lower-Level Prior Knowledge

Transforming prior knowledge of a higher level into a lower level always means a reduction of the information content in prior knowledge. Consequently, the transformation process mainly involves reduction or aggregation strategies. There are three cases of transformation that we describe here: transforming second- into first-level prior knowledge, transforming third- into second-level prior knowledge, and transforming third- into first-level prior knowledge.

The transformation of second-level prior knowledge, e.g. gene-disease associations, into first-level prior knowledge, e.g. genes, is straightforward. We define a transformation function

$$tr_{2\to1} : \{(a_i, rs_i)\}_{i=1\ldots|M|} \to \{a_i\}_{i=1\ldots|M|}, \quad \text{with } a_i \in M \text{ and } M \subseteq D \text{ or } M \subseteq A \text{ or } M \subseteq G \tag{4.7}$$

that takes second-level prior knowledge as input and returns the corresponding first-level prior knowledge. As both types of prior knowledge can contain entities of the same type, it is sufficient to remove all relevance scores rs_i from second-level prior knowledge and keep the remaining entities a_i to form first-level prior knowledge.

When transforming third-level prior knowledge, e.g. networks, to second-level prior knowledge, e.g. gene-disease associations, we can apply a similar extraction strategy as shown in Equation (4.7) by defining a transformation function

$$tr_{3\to2} : \{n_i\}_{i=1\ldots p} \to \{(a_j, rs_j)\}_{j=1\ldots o}, \text{with } a_i \in V_{all}, o = |V_{all}|, V_{all} = \bigcup_{i=1}^{p} V_i \tag{4.8}$$

that extracts the entirety of nodes V_{all} from every network n_i that was contained in third-level prior knowledge to build second-level prior knowledge. As a network n_i is composed of gene-gene interactions, i.e. $V_i \subseteq G$, we consequently can only derive genes as entities for first-level prior knowledge. In addition, we need to find a quantification of relevance rs_j for each extracted entity a_j. There are many options for computing rs_j based on the topographical information in the network, e.g. the number of interaction partners [117, 247, 265]. We choose to assess the relevance of a gene a_j based on the number of its interaction partners in every network by computing a relevance score rs_j for gene a_j from the sum of its percentile connectedness ranks $pr_{j,i}$, normalized by the overall number of pathways N_{a_j} that contains gene a_j:

$$rs_j = \frac{\sum_{i=1}^{|N_{a_j}|} pr_{j,i}}{100 \times |N_{a_j}|} \tag{4.9}$$

$pr_{j,i}$ is the percentile connectedness rank of a_j in a network n_i, where we rank each gene a_j by its number of interaction partners in n_i. While $pr_{j,i}$ ranges between 0 and 100, we add 100 to the calculation to rescale the final relevance score rs_j to a value between 0 and 1. Our method favors genes that have many interaction partners, i.e. serve as hub genes, and favors them even more if they serve as hub genes in multiple networks.

If third-level prior knowledge is transformed into first-level prior knowledge, we can apply the same transformation schema as in Equation (4.8), but leave out the computation of a relevance score

$$tr_{3\to1} : \{n_i\}_{i=1\ldots p} \to \{(a_j\}_{j=1\ldots o}, \quad \text{with} \quad a_i \in V_{all}, o = |V_{all}|, V_{all} = \bigcup_{i=1}^{p} V_i \tag{4.10}$$

4.2.2 Transforming Lower-Level Prior Knowledge into Higher-Level Prior Knowledge

Transforming prior knowledge from a lower to a higher level involves increasing the provided information content. This, however, can only take place if particular constraints are met.

First-level prior knowledge can be transformed into second-level prior knowledge via a transformation function

$$tr_{1\to2} : \{\{a_1, \ldots a_p\}_j\}_{j=1\ldots|KB_{query}|} \to \{(a_k, rs_k)\}_{k=1\ldots|M|},$$
$$\text{with } M = \bigcup_{j=1}^{|KB_{query}|} \{a_1, \ldots, a_p\}_j, M \subseteq D \text{ or } M \subseteq A \text{ or } M \subseteq G, rs_k \in \mathbb{R}^+ \tag{4.11}$$

that takes as input multiple sets of first-level prior knowledge $\{a_1, ..., a_p\}_j$ retrieved for the same input entity from multiple knowledge bases KB_{query}. Furthermore, all a_i must be of the same type, e.g. only contain genes. The output of $tr_{1\to2}$ is then second-level prior knowledge which contains a tuple for every entity a_i that occurs in the sets of first-level prior knowledge. These tuples consist of the entity a_k and an assigned relevance score rs_k that reflects the evidence level for a_k across the queried knowledge bases KB_{query}. This scheme is internally applied in a more sophisticated manner by meta knowledge bases like Open Targets and DisGeNET, which provide cumulative relevance scores based on the evidences provided by the integrated knowledge bases [142, 166].

The transformation processes for first- and second-level prior knowledge to third-level prior knowledge follow the same scheme, except that the query term qt_i used for prior knowledge retrieval must be provided as well.

For transforming first-level prior knowledge, e.g. genes, to third-level prior knowledge, e.g. networks, we define a transformation function

$$tr_{1\to3} : \{(qt_i, \{a_j\}_{j=1\ldots p})_k\}_{k=1\ldots|QT|\cdot|KB_{query}|} \to \{n_l\}_{l=1\ldots|N|} \quad \text{with } \forall(qt_i, \{a_j\}_{j=1\ldots p}) : qt_i \in G \text{ or } a_j \in M, M \subseteq G \quad (4.12)$$

that transforms a set of query terms and their correspondingly retrieved prior knowledge to a set of networks N. The transformation function $tr_{2\to3}$ for transforming second-level prior knowledge, e.g. gene-disease associations, to third-level prior knowledge, e.g. networks, works analogously. However, a transformation for both first- and second-level prior knowledge is only possible if in $(qt_i, \{a_j\}_{j=1\ldots p})_k$ at least the query term qt_i or the retrieved entities a_j are genes.

If both the query term qt_i and the entities a_j are genes, every a_j can be interpreted as an interaction partner of qt_i. As such, one set of first- or second-level prior knowledge is already sufficient to create a star-shaped network n_l with qt_i and every retrieved a_j as nodes and edges drawn between qt_i and every a_j. If multiple sets of first- or second-level prior knowledge are provided, these star-shaped networks can be merged on joint nodes to a larger network. In case second-level prior knowledge is used, the relevance score rs_j can additionally serve as edge weight for the edge between qt_i and a_j.

If only the query term qt_i is a gene, the retrieved entities a_j can be interpreted as annotation terms of gene qt_i. From this kind of information, it is possible to create a network n_l by using all $qt_i \in QT$ as network nodes and drawing edges between them if they share the same annotation term a_j in their retrieved prior knowledge [135]. To create such a network, however, we need prior knowledge retrieved for at least two different query terms, i.e. $|QT| > 1$.

If only the retrieved entities a_j are genes and qt_i is some other biological term, e.g. a disease name, the retrieved entities a_j share the same annotation term qt_i and can thus be considered as interaction partners. As such, we can already construct a densely coupled network n_l from a single set of prior knowledge where all entities a_j are nodes that are connected with each other. If multiple sets of prior knowledge that share at least one entity a_j are provided, we can join these entities to a larger set from which the densely coupled network is then created. When transforming first- or second-level prior knowledge, i.e. entities or scored entities, to third-level prior knowledge, i.e. networks, it is likely that the result will be a single large network, i.e. $|N| = 1$.

4.3 Strategies for Integrating Prior Knowledge into Feature Selection

Feature selection is the process of identifying the most relevant features in a data set. In the context of classifying gene expression samples, a *relevant* feature is a gene that

enables the distinction of samples into their separate classes. Traditional feature selection approaches assess the relevance of a gene based on its signals in a given data set [7]. The input data set is a gene expression matrix GE of size $q \times r$, in which an entry $ge_{i,j}$ captures the expression level of a gene $g_i \in G_{GE}$ for a sample $sa_j \in S$, with $|G_{GE}| = r$ and $|S| = q$. In addition, there exists a set of class labels L and a label assignment LA that assigns each sample sa_j to a class label $l \in L$.

We define a traditional feature selection approach as a function f_{trad} that processes the given input into a set of candidate genes $G_{trad} \subseteq G_{GE}$

$$f_{trad} : (GE, G_{GE}, LA) \rightarrow G_{trad} \tag{4.13}$$

or

$$f_{trad} : (GE, G_{GE}, LA) \rightarrow (G_{trad}, R_{trad}) \tag{4.14}$$

with or without returning a feature ranking R_{trad}, respectively. G_{trad} contains the candidate genes, which are either a fixed set that was determined by the feature selection approach, e.g. Lasso, or the top k genes, with $1 \leq k \leq |G_{GE}|$, which have the highest score in an accompanied ranking. The ranking

$$R_{trad} := \{(g_i, s_i^{trad})\}_{i=1\ldots|G_{GE}|} \tag{4.15}$$

that consists of tuples of a gene g_i and its relevance score s_i^{trad}.

Prior knowledge approaches obtain the relevance of a gene g_i from both its signals in the data and its biological importance according to prior knowledge. We thus adapt the initial definition for feature selection to also include prior knowledge PK

$$f_{pk} : (PK, GE, G_{GE}, LA) \rightarrow G_{pk} \tag{4.16}$$

or

$$f_{pk} : (PK, GE, G_{GE}, LA) \rightarrow (G_{pk}, R_{pk}) \tag{4.17}$$

that again produces a set of candidate genes G_{pk} with or without a ranking

$$R_{pk} := \{(g_i, s_i^{joint})\}_{i=1\ldots|G_{GE}|} \tag{4.18}$$

where s_i^{joint} is based on both the data signals and biological relevance of the gene.

The actual integration strategy of a prior knowledge approach and the required level of prior knowledge can differ between approaches. However, we can still classify prior knowledge approaches into distinct types of *modifying*, *combining*, and *network approaches* based on how thorough that prior knowledge is integrated into the selection process. In the following subsections, we describe the general integration strategies of the differ-

ent types. Additionally, we provide a corresponding classification of related approaches presented from Section 3.1 in Tables 10.8 to 10.11 in the appendix.

4.3.1 Modifying Approaches

Modifying approaches typically work in combination with a traditional feature selection approach, by forming a two-level selection process. In a modifying approach, a feature set that was retrieved either from a knowledge base, e.g. first- or second-level prior knowledge, or a traditional feature selection approach is subsequently adapted in a filtering or extending manner. Figure 4.1 depicts the two general processing schemes that are followed by modifying approaches. Depending on the point in time a feature set is adapted by prior knowledge, we split up modifying approaches into those applying pre- and post-modification.

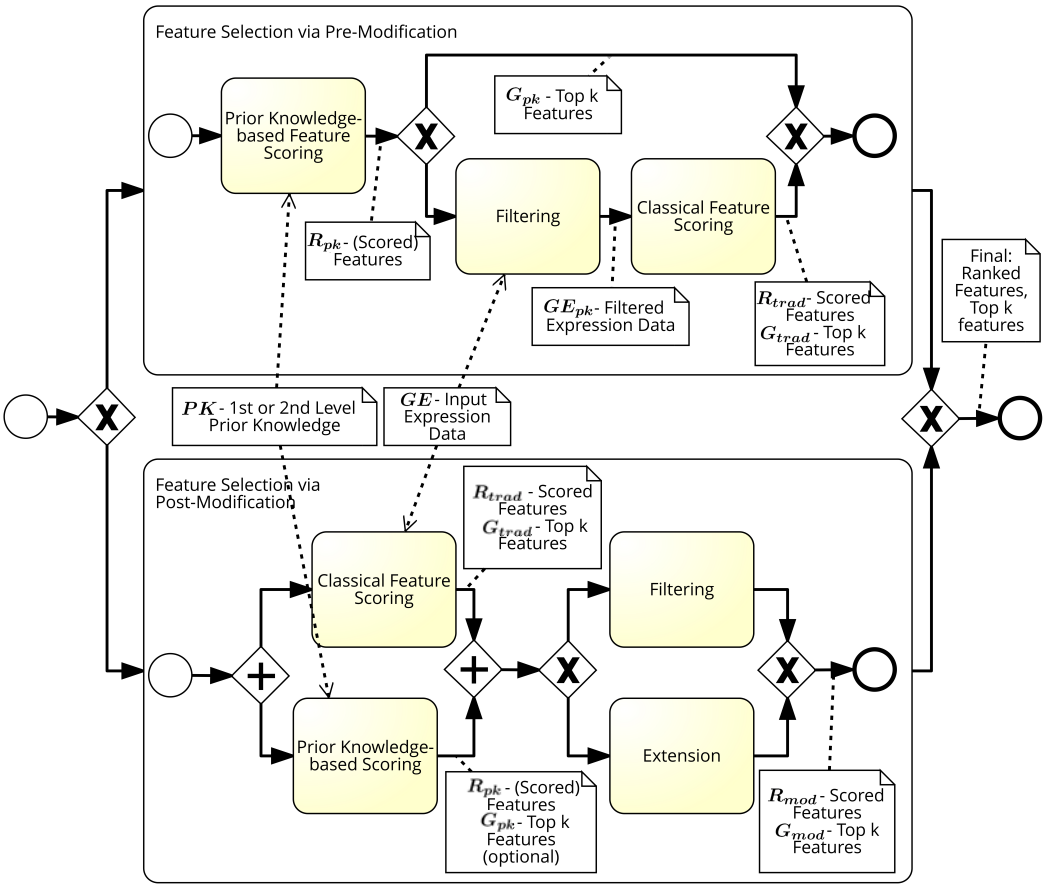


Fig. 4.1: The two possible process flows for modifying approaches, modeled in BPMN 2.0. Prior knowledge can be incorporated into feature selection either at the very beginning for filtering (upper part of the process) or at the end, to filter or extend an existing feature set retrieved by traditional feature selection (lower part of the process).

Pre-modification

A pre-modification approach as depicted in the upper part of Figure 4.1 uses prior knowledge directly on the input data set before any further data analysis is carried out [58, 98]. If a ranking R_{pk} can already be produced, e.g. from the relevance scores of second-level prior knowledge, feature selection can stop after this step. However, this does not take the statistical information present in the data set into account. A two-level approach instead uses the derived feature set G_{pk} to first filter the original input matrix GE to a new matrix GE_{pk} whose columns correspond to G_{pk} and then apply any desired feature selection strategy afterwards.

When prior knowledge is used for prefiltering in such a two-level approach, it can reduce the computational runtime of advanced feature selection approaches that have a high computational complexity. However, prefiltering also prevents the discovery of previously unknown relationships that are present in the data set.

Post-modification

If prior knowledge is applied in a post-modification manner as depicted in the lower part of Figure 4.1, both traditional feature selection and prior knowledge-based scoring can be carried out in parallel to produce two separate feature sets G_{trad} and G_{pk} with or without a ranking R_{trad} and R_{pk}, respectively. These are subsequently combined in a filtering or extension manner to form a new feature set G_{mod}, e.g. via $G_{trad} \cup G_{pk}$. If rankings R_{trad} and R_{pk} exist, G_{mod} is based on the ranking

$$R_{mod} := \{(g_i, s_i^{mod})\}_{i=1\ldots|G_{GE}|}, \text{with } s_i^{mod} \in R_{trad} \text{ or } s_i^{mod} \in R_{pk} \tag{4.19}$$

where the score s_i^{mod} of a gene g_i comes from either traditional feature selection or from prior knowledge-based scoring, depending on the chosen approach.

Filtering a feature set, however, can prevent the discovery of unknown associations, which is sometimes not desirable. In contrast, using both feature sets to extend each other enables the detection of relevant features that have weak signals in the data, e.g. due to processing errors.

4.3.2 Combining Approaches

Unlike modifying approaches, combining approaches incorporate both data signals and prior knowledge in a joint processing step to produce a combined feature set G_{comb} with or without a ranking R_{comb}, as depicted in Figure 4.2. This way, genes showing weak signals in the data but strong evidence found in a knowledge base can still be selected for G_{comb}. The same is valid for genes showing strong signals in the data but having no evidence in knowledge bases. Combining approaches thus allow genes that have both well- and unknown relations to the use case context to be detected.

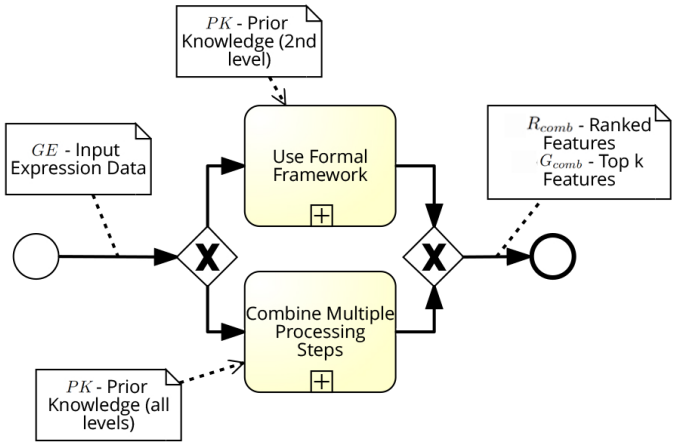


Fig. 4.2: Process flow of combining prior knowledge approaches. Both statistical data characteristics and biological relevance are considered in a joint processing step to compute a feature set G_{comb} and (optionally) a corresponding ranking R_{comb} whose scores reflect both aspects.

The actual selection strategy for G_{comb} varies for the individual approaches. Still, combining approaches can be separated into those applying a formal framework to include prior knowledge and those following a multi-stage processing that periodically blends in prior knowledge, e.g. via multiple clustering stages. We thus separate combining approaches into the two subtypes of formal frameworks and process-oriented combinations.

Formal Frameworks

Combining approaches that apply a formal framework to select a set of candidate genes G_{comb} require a definition of feature relevance regarding both its statistical characteristics in the data set and its biological importance from prior knowledge. While such formal definitions are inherently available for statistical characteristics from traditional feature selection approaches, the biological relevance of a feature must be provided separately. From the three levels of prior knowledge, only second-level prior knowledge, i.e. a list of scored entities, quantifies the biological relevance of an entity. As such, combining approaches that apply a formal framework are restricted to work with second-level prior knowledge only. This biological relevance is typically incorporated as an extra weight into the computation process, e.g. as Bayesian prior, regression weight, or penalty term to use in regression or regularization methods(see also Tables 10.8 to 10.11 in the appendix) [68, 153, 207].

Process-oriented Combination

In contrast to formal frameworks, a process-oriented combination of prior knowledge and statistical characteristics interleaves both information types in multiple processing steps. The expected level of prior knowledge cannot be generalized here, as the required

level of prior knowledge strongly depends on the actual processing steps, which can vary widely from iterative score recalibration, clustering of the data or of prior knowledge, network construction and mapping, to correlation computations [8, 131, 169].

4.3.3 Network Approaches

Network approaches are the most complex form of integrating prior knowledge, as they select networks or subnetworks — in the following, uniformly referred to as networks — from third-level prior knowledge, e.g. pathways, as features. Consequently, network approaches are rather a feature extraction strategy, as they replace the original feature space of genes by networks. Fundamental for network approaches is the assumption of network locality of genes. Genes that are linked via the same network or pathway participate in the same biological process and share similar functions [117]. Sharing similar functions results in similar expression levels and patterns, i.e. co-expression [54, 249]. The effect increases the closer two genes are in the network [244]. As such, disease genes, in particular, tend to build densely coupled subnetworks [132].

Consequently, using networks as features enhances analysis in multiple aspects. First, it reduces the chance of noise propagation: while expression levels of single genes can correlate by chance with an outcome, this is far more unlikely for expression levels of a majority of genes in a network [257]. Second, it reduces redundancy in the selected feature set: genes in a network typically show joint expression behavior. When selecting genes as features, it can thus happen that multiple genes showing strong statistical signals and belonging to the same underlying pathway are selected for the final feature set. This leads to a particular process being overrepresented in the feature set, thus introducing redundancy. Third, networks likely include important marker genes that could be mistakenly ignored in gene-based feature selection, e.g. because they are not differentially expressed.

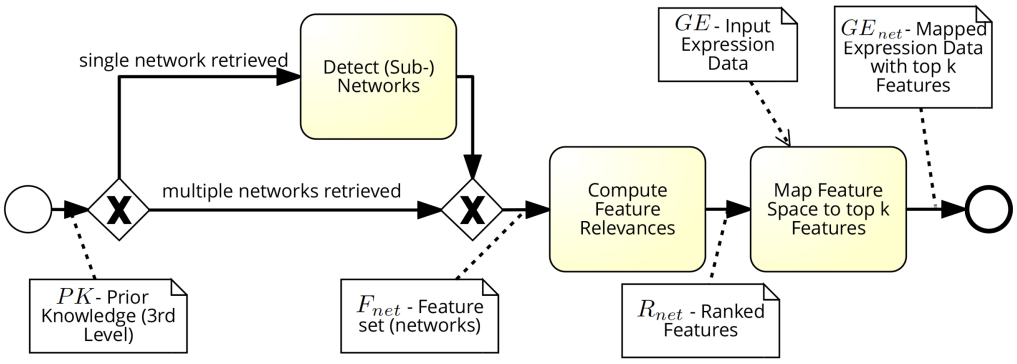


Fig. 4.3: Process overview for network approaches. Features are optionally derived first from third-level prior knowledge and ranked subsequently. In addition, the original feature space of the input data set must be transformed to the new feature space.

Figure 4.3 depicts the processing steps for network approaches. At first, third-level prior knowledge is used to compute a new set of features F_{net} and a corresponding ranking

$$R_{net} := \{(f_i, s_i^{net})\}_{i=1...|F_{net}|}, f_i \in F_{net} \tag{4.20}$$

If the retrieved third-level prior knowledge consists of a single network n_1, the set of features F_{net} must be derived from that network first, e.g. as subnetworks or network motifs. These can be detected via standard search strategies, e.g. greedy search, or by correlating the expression values of interacting genes, e.g. via single-sample Gene Set Enrichment (in which genes are coordinately up- and downregulated) [16]. If a set of networks N is provided, a network n_i can be directly used as feature, i.e. $n_i = f_i$ with $N = F_{net}$, for which a relevance score s_i^{net} will be computed. The relevance scores of network-based features like pathways are typically based on the expression levels of their member genes, e.g. the mean expression values, and their correlation with sample classes, e.g. via differential expression or significance tests.

In order to allow for further processing after feature selection, e.g. classification, network approaches need to carry out an additional processing step after the actual feature selection: transferring the original feature space G into the new feature space F_{net}. For this purpose, new feature values — so-called *pathway activity scores* — must be computed from the input data set GE to achieve the mapped data set GE_{net}. Such a computation typically incorporates the expression levels of all or of a particular subset of member genes, also by correlating these to sample class labels [41, 76, 113].

4.4 Generalized Approaches to Flexibly Integrate Prior Knowledge into Feature Selection

While existing prior knowledge approaches can be categorized into the aforementioned classes, they still face the challenges described in Section 3.1.2: existing approaches are custom solutions that are not flexible regarding the applied knowledge base and — if they are modifying or combining approaches — the applied traditional feature selection approach. In the following, we describe our concepts for modifying, combining, and network approaches that address these issues. We assume that prior knowledge is retrieved for a set of predefined query terms, i.e. $qt_j \in QT$, and that the entities of first- and second-level prior knowledge correspond to genes, i.e. $a_i \in G$.

4.4.1 Modifying Prior Knowledge Approaches

We define concepts for three modifying approaches that are generalized to work with any given traditional feature selection approach — as long as it produces a feature ranking — and first- or second-level prior knowledge with genes as entities. For every approach, we first extract all entities a_i from the retrieved prior knowledge to build G_{pk}. In the

following, G_{pk} refers to the set of genes derived from the prior knowledge retrieved for the individual approaches.

4.4.2 Prefiltering Approach

For the first modifying approach, we apply a function

$$f_{filter} : (G_{pk}, GE) \rightarrow (GE'_{q' \times r}), \text{with } q' = |G'_{GE}|, G'_{GE} = G_{pk} \cup G_{GE} \tag{4.21}$$

that filters the features, i.e. genes G_{GE}, of an input matrix GE to only contain features that were present in the provided first-level prior knowledge G_{pk}. The reduced matrix GE', the reduced feature set G'_{GE}, and the label assignment LA are then provided to traditional feature selection

$$f_{trad}(GE', G'_{GE}, LA) \rightarrow (G_{trad}, R_{trad}) \tag{4.22}$$

to receive a ranking of the remaining features G'_{GE} that is based on the statistical characteristics of the data. As all genes that were not provided in prior knowledge are removed in this approach, it prevents unknown interactions and relations between genes from being detected.

4.4.3 Postfiltering Approach

The second filtering approach first applies any desired traditional feature selection strategy as described by Equation (4.13) and then filters the produced ranking R_{trad} afterwards to receive a new ranking R_{mod} that contains only those genes G_{pk} that were retrieved from first-level prior knowledge:

$$R_{mod} := \{(g_i, s_i^{trad})\}_{i=1\ldots|G'_{GE}|}, \text{with } g_i \in G'_{GE}, G'_{GE} = G_{pk} \cup G_{GE} \tag{4.23}$$

For univariate feature selection approaches that do not require an additional correction for multiple statistical testing, e.g. t-tests using Bonferroni correction, results of both pre- and postfiltering approaches will be the same. However, results of both pre- and postfiltering approaches will differ for multivariate feature selection approaches that incorporate gene-gene dependencies, e.g. Information Gain, or apply additional corrections for multiple testing.

4.4.4 Extension Approach

The last modifying approach extends a feature set that is retrieved with a traditional feature selection approach by genes that are contained in the retrieved prior knowledge. The aim of this extension approach is that every feature set of arbitrary size selected

should consist of nearly equal parts of a) genes highly ranked from traditional feature selection and b) genes retrieved from prior knowledge. For that we define a function

$$f_{ext} : (G_{pk}, GE, G_{GE}, LA) \rightarrow (G_{trad} \cup G_{ext}, R_{trad}) \tag{4.24}$$

that receives as additional input the set of genes G_{pk} contained in the retrieved prior knowledge, and returns the traditional ranking R_{trad}. In addition, f_{ext} returns an adapted feature set in which half of the k features are selected based on their traditional score, i.e. originate from G_{trad} with $|G_{trad}| = 2/k$, and the other half is selected from $G_{ext} = G_{GE} \cup G_{pk}$ with $|G_{ext}| = 2/k$. If second-level prior knowledge is available, G_{ext} contains those genes g_i with the highest relevance score rs_i assigned.

4.4.5 Combining Approach

Our combining approach applies a formal framework to compute the relevance score of a feature. Consequently, it requires second-level prior knowledge. It can be applied to any traditional feature selection approach that produces feature relevance scores and be combined with any knowledge base providing second-level prior knowledge. The objective behind this approach is to consider both the data characteristics and the biological relevance of a gene for score computation. By this, it should be possible to select genes that do not show a strong signal in the data, e.g. due to processing errors, but are known to play an important role in the disease at hand. Vice versa, our approach should also favor genes that show a strong signal in the data, even if no evidence is provided in prior knowledge. Consequently, our approach computes a relevance score

$$s_i(g_i) = rs_i \cdot s_i^{trad} \tag{4.25}$$

for a gene g_i by introducing its biological relevance rs_i as additional weight to its statistical relevance s_i^{trad}. If no second-level prior knowledge exists for g_i, we assign a minimum default value of $rs_i = 0.1^{-5}$. s_i^{trad} can be provided by any feature selection approach that computes a statistical relevance, e.g. variance or Lasso feature coefficients.

4.4.6 Network Approach

The third approach we present selects features from a given set of networks N_{PK} that was retrieved as third-level prior knowledge from a knowledge base. As we do not apply any motif detection algorithm in advance, we require $|N_{PK}| > 1$ to be able to select multiple features. In order to produce a ranking of networks as described in Equation (4.20), we compute a score s_i for every network $n_i \in N_{PK}$ that extends the strategy described by Tian et al. [229]. The authors consider a network n_i to be relevant if its member genes g_j show a coordinated expression behavior with the sample classes. To retrieve a relevance score for a network n_i, Tian et al. compute the average from the t-test statistic

scores of its member genes, which assesses the probability whether n_i is altered in the disease. However, as t-tests can only be used for binary comparisons, this method is not applicable when comparing more than two classes, e.g. disease subtypes. Building on this, we extend the method of Tian et al. to be applicable to such use cases by using F-test statistic scores instead of t-test. Hence, we compute the relevance score s_i for a network n_i as the average F-test statistic score from its member genes V_i. F_{g_j}, which is the individual F-test statistic score for a gene $g_j \in V_i$, denotes if the expression values of g_j across all samples are correlated to the samples' class assignments.

$$s_i(n_i) = \frac{\sum_{j=1}^{|V_i|} F_{g_j}}{|V_i|}, \text{with } g_j \in V_i \tag{4.26}$$

The lower the resulting score s_i, the more likely is a correlation of a network n_i and overall sample classes. The produced ranking R_{net} then has a reverse order: networks with a lower score s_i have a higher rank, whereas networks with a higher score s_i are less likely to correlate and therefore receive a lower rank.

We select the top-scored networks as features and compute a new feature value $fv_{i,k}$ for every feature n_i and sample sa_k based on its member genes' expression values and Vert's and Kanehisa's definition of *relevance* and *smoothness* [244]: a gene as feature is considered relevant if it shows a high variance across samples; it is considered smooth if it is co-expressed with its neighbor genes.

While Vert and Kanehisa define these attributes for individual genes in a large joint network, we adapt this meaning to networks themselves. The relevance and smoothness of a network n_i is thus based on the relevance and smoothness of its member genes. Highly variant genes are suspected to play more important roles in processes than genes that show the same expression level across conditions. Thus, if a network contains many genes of high variance, it is likely to be *activated* and *deactivated* across conditions. The smoothness of a gene g_j is based on how coordinatedly up- and downregulated it is with its direct interaction partners in the network. As genes which share the same pathway are supposed to participate in successive cell reactions, they likely share a similar expression pattern in clusters, which is even more likely for characteristic disease genes [132]. We thus compute a feature value

$$fv_{i,k}(n_i, sa_k) = \frac{\sum_{j=1}^{|V_i|}(expr(g_j)_{sa_k} \cdot rel(g_j) \cdot smooth(g_j))}{|V_i|} \tag{4.27}$$

for a network n_i and sample sa_k from the average of weighted expression values of its member genes g_j, where the weights incorporate the relevance and smoothness of g_j. The relevance $rel(g_j)$ of a gene g_j corresponds to the variance of its expression levels $expr(g_j)$ across all samples

$$rel(g_j) = Var(expr(g_j)) = \mathbb{E}\left[(expr(g_j) - \mathbb{E}(expr(g_j)))^2\right] \tag{4.28}$$

The smoothness of a gene g_j is computed from the co-expression between g_j and its interaction partners I_{g_j} in the network, where $|\rho_{g_j,g_i}|$ is the absolute Pearson correlation between the expression values of g_j and its interaction partners g_i

$$smooth(g_j) = \frac{\sum_{i=1}^{|I_{g_j}|} |\rho_{g_j,g_i}|}{|I_{g_j}|} \tag{4.29}$$

Thus, the feature value not only depicts the average expression value of genes in the network, but aims to emphasize the expression levels of those genes that are smooth or relevant and ignore the others. If a gene is not coordinately expressed with its neighbor genes, i.e. $smooth(g_j) \approx 0$, or shows no variance across samples, i.e. $rel(g_j) \approx 0$, its expression value will practically be ignored during feature value computation. Consequently, networks having few genes that are both relevant and smooth will generally receive a lower feature value than networks with many relevant and smooth genes.

4.5 Summary

In this chapter, we defined the concepts of prior biological knowledge and prior knowledge approaches in the context of feature selection on gene expression data sets. By studying the available knowledge bases, we identified that prior knowledge can take on three levels that vary in their provided information content: first-level prior knowledge, which are lists of entities like genes; second-level prior knowledge, which are lists of scored entities; and third-level prior knowledge, which are lists of gene-gene interaction networks. Not every knowledge base provides all three levels of prior knowledge, and not every prior knowledge approach can cope with all kinds of levels for integration. However, it is still possible to transform prior knowledge from a higher to a lower level and vice versa, if particular assumptions and constraints are met.

Prior knowledge approaches that incorporate the aforementioned prior knowledge can further be classified into three types based on how thoroughly prior knowledge is integrated into the selection process: modifying approaches simply adapt a feature set with prior knowledge via filtering or extension and can be combined with any traditional feature selection approach. Combining approaches incorporate prior knowledge into the computation of the final relevance score of a feature, e.g. as additional weight in a formula. Network approaches do not select genes but networks as features, which were retrieved from prior knowledge. Finally, we also described novel modifying, combining, and network approaches that are flexible regarding the applied knowledge base and combination with traditional feature selection approaches.

5

Comprior: A Software Tool to Effortlessly Implement and Benchmark Prior Knowledge Approaches

This chapter deals with the technical realization of the concepts we defined in the preceding chapter for both prior knowledge and prior knowledge approaches. The resulting implementations are bundled in a software tool called *Comprior*, which addresses the previously described issues regarding the practical applicability and comparability of prior knowledge approaches. Comprior provides the technical infrastructure for both the rapid implementation and evaluation of prior knowledge approaches. In the following, we describe Comprior's key functionalities, discuss its architecture design, and provide the implementation details of selected aspects to also achieve flexibility, extensibility, and uniform access to prior knowledge.

5.1 General Description of Comprior

Comprior supports large parts of the classical feature selection workflow, at the same time complying with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [252].

5.1.1 Supported Processing Functionality

Figure 5.1 illustrates the general concept of Comprior and its key functionalities. The classical processing workflow for feature selection typically involves preprocessing, feature selection, and evaluation [23]. Consequently, Comprior provides core functionalities across these three workflow steps.

For **preprocessing**, Comprior provides automated data cleansing, identifier mapping, and data labeling. The input expression data set, which is assumed to contain normalized data, can contain any column orientation, i.e. features in rows or columns, and use any identifier format, e.g. microarray probes or Ensembl identifier. If necessary, the gene expression data can be filtered for samples and features that have missing values above a specified threshold. If the user wants to retrieve features in a different format than the original one, e.g. Human Gene Nomenclature (HGNC) format instead of microarray probes, Comprior takes care of automated identifier mapping throughout the

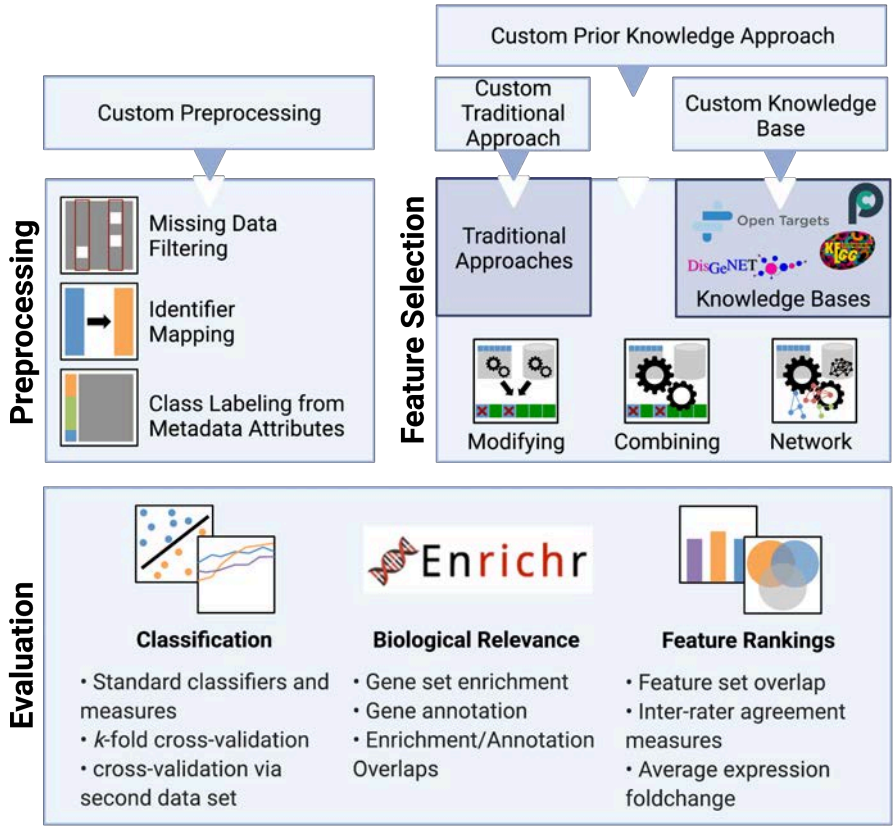


Fig. 5.1: Overview of Comprior's functionality, covering preprocessing, feature selection, and evaluation. Comprior provides both traditional feature selection approaches, e.g. filter or wrapper, and prior knowledge approaches, e.g. modifying or combining, that can be combined with prior knowledge from Open Targets, DisGeNET, KEGG, or PathwayCommons. For evaluation, Comprior provides classification functionality with standard measures and cross-validation on a second data set, gene set enrichment and annotation with EnrichR, and multiple measures to compare the actual feature sets. Comprior can also be extended by custom functionality, e.g. own prior knowledge approaches.

complete workflow. In order to create a labeled data set for subsequent classification, Comprior automatically labels the input data set based on a user-defined metadata attribute. Comprior can also be extended by custom preprocessing functionality, e.g. normalization. For **feature selection**, Comprior provides both traditional feature selection and prior knowledge approaches. Users can configure Comprior to run multiple feature selection approaches in parallel. Available traditional approaches cover filter, wrapper, and embedded approaches. Available prior knowledge approaches cover modifying, combining, and network approaches, all of which can be flexibly combined with any of the currently available knowledge bases: KEGG, Open Targets, DisGeNET, and PathwayCommons [101, 108, 165, 180]. Comprior currently provides the three prefiltering, postfiltering, and extending modifying approaches that were described in detail in Section 4.4.1. All of these three types allow for combining any of the available statistical

approaches with any of the available knowledge bases. Comprior currently provides two combining approaches. First, we implemented the approach presented in Section 4.4.5 by weighting the statistical relevance of a feature, e.g. computed by any available statistical selection method, by an association score retrieved from a knowledge base. The second combining approach introduces prior knowledge as a feature-specific penalty score during Lasso computation and was implemented by Zeng et al. and integrated into Comprior by us [261]. As for network approaches, Comprior currently provides our own approach as described in Section 4.4.6 and the approach by Lee et al, which we reimplemented based on their descriptions in the publication [113]. Lee et al. use the same selection strategy as our approach that was first described by Tian et al. [229]: a network is considered relevant if the gene expression profiles of its member genes correlate with the data set classes. However, Lee et al. follow a different approach to map the feature space, i.e. by defining a pathway activity score based on the average expression values of condition-responsive genes (CORGs) in the network.

For **evaluation**, Comprior provides functionality to assess input data quality, knowledge base coverage, and effectiveness of feature selection approaches. All plots generated by Comprior use a consistent coloring scheme. From the given gene expression data sets, Comprior creates density plots, distribution box plots, and multi-dimensional scaling (MDS) plots for quality assessment. Comprior further computes summary statistics for the available prior knowledge in the applied knowledge base and visualizes these in corresponding plots. For assessing the effectiveness of the approaches, Comprior provides automated functionality for classification, enrichment, and runtime measurements. For classification, users can select multiple standard classifiers for k-fold cross-validation. Classification results are then assessed with standard measures, e.g. accuracy, F_1, area under ROC curve (AUROC), or Matthew's correlation coefficient (MCC), for which Comprior automatically creates corresponding plots. In addition, users can provide a second data set for robustness evaluation. This data set can be related to the original input data set in a traditional train-test manner, but can also be completely unrelated and even use different identifiers — Comprior automatically carries out cross-validation of the selected features on this data set. Comprior also measures runtime performances of all feature selection approaches and breaks down the amount of time needed for prior knowledge retrieval, traditional feature selection approaches, and feature mapping. To assess the biological relevance of feature sets, Comprior uses Enrichr for automated gene set annotation and enrichment [35, 255]. Feature sets, annotations, and enrichments are further compared to each other by creating plots that depict their overlaps.

5.1.2 Tool FAIRness

While the FAIR principles, as introduced by Wilkonson et al., were originally intended for the management of data sets, recent efforts are aiming at transferring and adapting them to software as well [252]. Based on guidelines summarized by Gruenpeter et al., we discuss the software FAIRness of Comprior [73]. The complete software is licensed under

the MIT license and freely accessible in a public GitHub repository that also provides a limited version control (F, A, R) [155]. Comprior can be installed from source in a semi-automated process or can be directly executed in a Docker container that automatically resolves all installation dependencies (I, R). Comprehensive online material provides full code documentation, architecture description, tutorials, and troubleshooting help (F, A, R) [155]. Together with Comprior's modular architecture with clearly defined interfaces, it supports and encourages researchers to integrate custom extensions into Comprior (A, I, R). In addition, Comprior also returns intermediate data artifacts during the analysis, e.g. the transformed input data set or feature rankings, which can be reused for any other custom workflows (I).

5.2 Architecture Design

Comprior consists of multiple system components that correspond to distinct functionality and interact with each other via predefined interfaces. Figure 5.2 depicts the system architecture of Comprior in a UML 2.0 components diagram, showing components for pipeline execution, preprocessing, prior knowledge retrieval, evaluation, and administrative tasks.

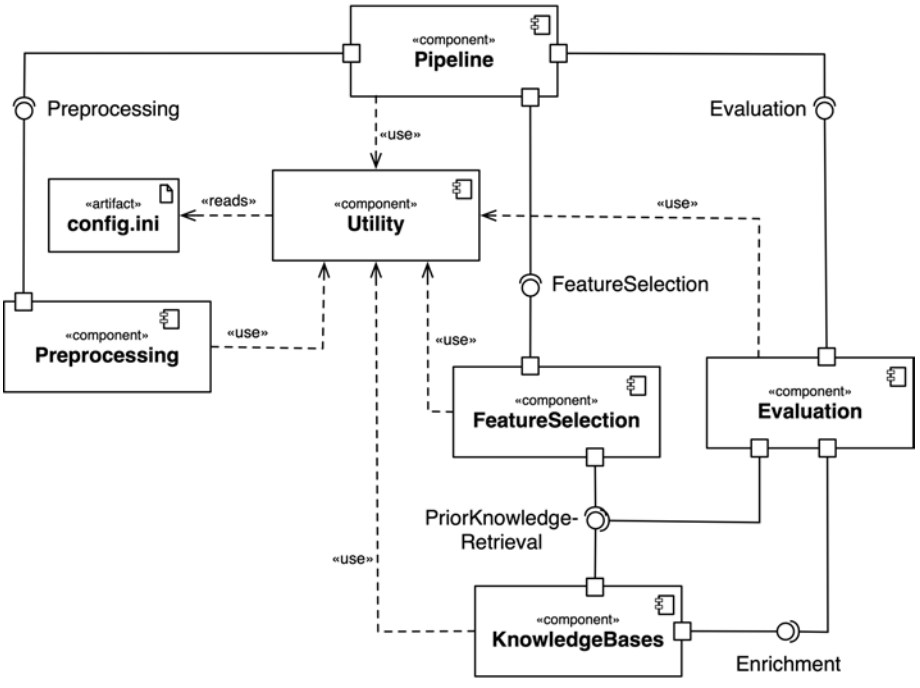


Fig. 5.2: Overview of Comprior's system components. The Pipeline component is responsible for benchmark orchestration, while specific functionality is implemented in dedicated components. Communication between the components is realized via corresponding interfaces.

The **Pipeline** component constitutes the starting point of Comprior and orchestrates the complete benchmark execution. It defines the logic for benchmark execution based on the user-defined configuration: preprocessing the input data, running feature selection approaches, and executing evaluation strategies. It also performs administrational tasks, e.g. preparing output directories or handling the parallel execution of feature selection strategies.

The **Utility** component provides general functionality that is needed throughout the whole benchmarking process and is therefore accessed by all other components. It stores the user-defined configuration parameters and offers functionality for directory and file management, logging, and identifier mapping. It also provides wrappers for invoking R and Java code.

The **Preprocessing** component is responsible for preprocessing and transforming the input data sets, e.g. missing value filtering or identifier mapping. Preprocessing functionality is invoked and orchestrated by the Pipeline component via the *Preprocessing* interface.

The **FeatureSelection** component provides the approaches for feature selection. We have implemented functionality to:

- provide baseline selection strategies, e.g. selecting at random or using only prior knowledge,
- import traditional approaches from existing packages, e.g. ANOVA,
- provide wrapper selectors that invoke non-Python implementations,
- combine traditional approaches with knowledge bases, and
- select networks, pathways, or submodules as features.

Feature selectors are invoked and orchestrated by the Pipeline component via their *FeatureSelection* interface. Feature selectors retrieve prior knowledge by accessing the KnowledgeBase component via the interface *PriorKnowledgeRetrieval*.

The **KnowledgeBase** component encapsulates access to knowledge bases. We consider a knowledge base to be any online resource that provides scientific biological information. Most of the knowledge bases available in Comprior are used for prior knowledge retrieval during feature selection. However, the KnowledgeBase component also provides feature set enrichment and annotation functionality for the Evaluation component, and identifier mapping that is accessed by the Utility component.

The **Evaluation** component contains all functionality that is related to quality and performance assessment and plot creation. This covers a) the assessment of knowledge base coverage for the provided search terms, b) the inspection of input data set quality, c) the classification of the input data sets based on the selected feature sets, d) the assessment of the feature sets based on classification performance metrics, feature set annotations

and enrichments, and their subsequent comparisons. For this, the Evaluation component accesses the KnowledgeBase component either via the *PriorKnowledgeRetrieval* or *Enrichment* interfaces. All of Comprior's output plots are generated within the Evaluation component, whose functionality is invoked and orchestrated by the Pipeline component via the *Evaluation* interface.

5.3 Ensuring Extensibility by Custom Functionality

Comprior was designed to be extensible and to facilitate a straightforward implementation of custom approaches. This is achieved by a standardized communication between system components and wrapper functionality to include custom code from programming languages other than Python.

5.3.1 Standardized Interfaces between Components

To preserve extensibility, Comprior defines interfaces for interactions between components. These interfaces are enabled by a certain inheritance structure that is implemented in each of Comprior's components. New functionality can therefore be easily integrated into Comprior by following the inheritance structure and implementing the required interface methods.

Figure 5.3 exemplifies the inheritance structure by depicting the class structure of the Preprocessing component. The top of the hierarchy is the main abstract class *Preprocessor*, which enforces its inheriting classes to realize the *Preprocessing* interface. This interface consists of a single method called *preprocess()*. Interface methods like *preprocess()* do not require any parameters — instead, necessary parameters are set at object creation. Actual preprocessing functionality is realized in distinct classes that inherit from *Preprocessor* and then implement the interface method *preprocess()*. For example, class *MappingPreprocessor* is responsible for mapping the input data sets to the desired format, while *FilterPreprocessor* filters the input data for samples and genes with missing values.

This inheritance principle was analogously applied in the modules FeatureSelection, KnowledgeBases, and Evaluation. Developers can then integrate new functionality by implementing it in a custom class that inherits from the main super class — or one of its descendants — and implements the interface methods.

5.3.2 Including External Code

The majority of Comprior is implemented in Python. However, custom functionality must sometimes be implemented in a different programming language, e.g. because the developer is more familiar with it or because an efficient implementation is already available. In particular, many bioinformatics methods are provided by R packages. To

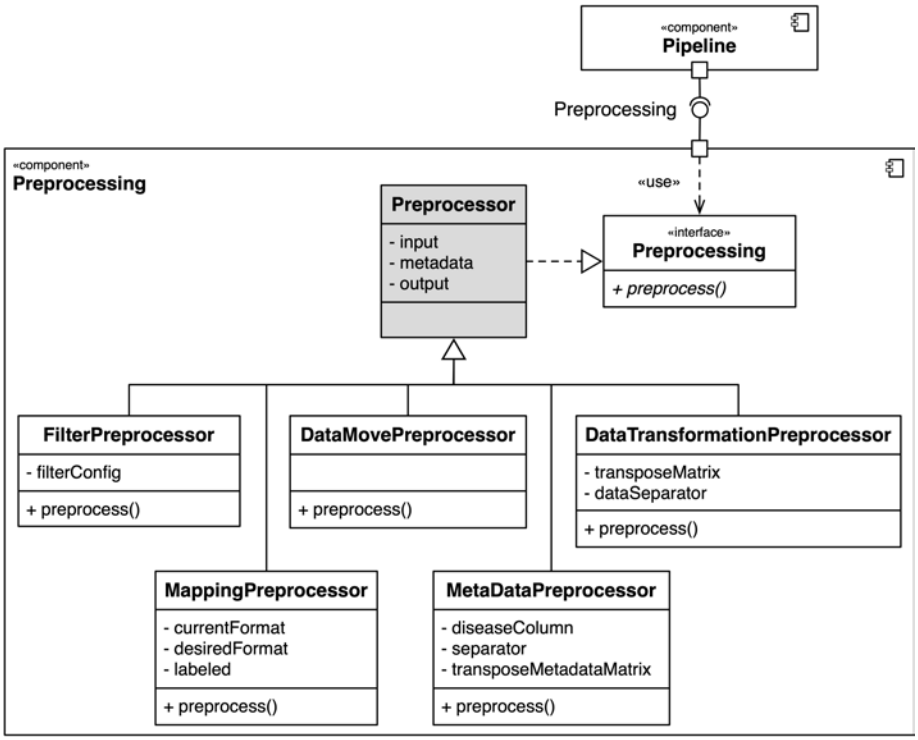


Fig. 5.3: Class structure of the preprocessing module. The top of the hierarchy is an abstract Preprocessor class that forces its inheriting classes to implement the Preprocessing interface. Actual functionality for data preprocessing, e.g. identifier mapping or filtering, is realized in corresponding subclasses.

allow users to make use of the full spectrum of the available bioinformatics functionality, Comprior can also run R and Java code. For this, the Utility component implements interface functions to call external code. These functions can be invoked from anywhere within Comprior, e.g. they are used by *RSelector* and *JavaSelector* to invoke corresponding code for feature selection. When implementing new functionality, developers can therefore decide on their own if they wish to implement it in Python, R, or Java.

5.4 Reducing the Implementation Effort for Comprehensive Benchmark Experiments

One of Comprior's key objectives is to reduce the effort of setting up comprehensive evaluation experiments for feature selection approaches. This covers an effortless experiment configuration, low requirements on the format of input data sets, flexible selection of feature selection approaches and knowledge bases, and comprehensive measures to assess the performance of feature selection approaches. In the following, we elaborate on how these aspects are realized within Comprior.

5.4.1 Experiment Configuration

As a benchmark spans a lot of processing steps, there are many options for adjustment. These range from high-level experiment design decisions, e.g. which feature selectors or classifiers to use, to more fine-grained parameters like identifier format or parameters of a particular feature selection approach. To enable a flexible pipeline design, Comprior uses *.ini* configuration files that are processed via Python's built-in ConfigParser component. These configuration files contain all parameters that Comprior needs for functioning properly, including access points to knowledge base web services and output folder structure. There is a main configuration file that specifies default parameters and is always used by Comprior. On top of that main configuration file, users can specify their own configuration files that contain only those parameters they want to overwrite from the main configuration, e.g. where the input data is located or what feature selectors to apply.

```
[Dataset]
input = /path/to/example.csv
metadata = /path/to/example_metadata.csv
genesInColumns = true
dataSeparator = ,
currentGeneIDFormat = ENSG
finalGeneIDFormat = HGNC

[Gene Selection - Methods]
traditional_methods = ANOVA RandomForest

[Classification]
classifiers =  NB LR SMO RF
metrics = sensitivity specificity accuracy kappa F1 AUROC
```

Listing 5.1: Excerpt from an example configuration file as used by Comprior. Configuration parameters are separated into sections, e.g. for specifying the input data format, that contain key-value pairs.

Listing 5.1 shows an excerpt of a configuration file that applies ANOVA and RandomForest feature selection approaches and validates their performance by classifying them on four different classifiers: Naive Bayes (NB), linear regression (LR), Support Vector Machine (SMO), and RandomForest (RF). The parameters in Comprior's configuration files are grouped into different categories. For example, the *Dataset* category enables specification of the paths to the input files, their column orientation, data separators used, and whether the currently used identifier format should be mapped to another, e.g. Human Gene Nomenclature (HGNC).

5.4.2 Enabling a Flexible Combination of Feature Selection Approaches

One aim of Comprior is to enable users to flexibly combine a feature selection approach with any of the available knowledge bases. This functionality is mainly realized in the FeatureSelection component, whose class architecture is depicted in Figure 5.4. Abstract classes that do not implement a specific feature selection approach are highlighted in grey. For the sake of clarity, Figure 5.4 only shows the most important classes, omitting some classes implementing concrete feature selection approaches. Any feature selection approach is realized in a separate class that inherits from the abstract *FeatureSelector* class — or one of its inheriting abstract classes. Class *FeatureSelector* realizes the *FeatureSelection* interface, which consists of the single method *selectFeatures()* and must be implemented by a class inheriting from *FeatureSelector*. This method is invoked by the Pipeline component to start the dedicated feature selection. The object creation of a FeatureSelector instance is encapsulated by a *FeatureSelectorFactory* class that corresponds to the Factory Method Pattern as described by Gamma et al [62]. When given a particular keyword from the Pipeline component, the FeatureSelectorFactory automatically recognizes what kind of feature selection object to create and if a knowledge base must be added. For example, an ANOVA feature selector can be combined with the KEGG knowledge base in a prefiltering manner as described in Section 4.4.1. For that, the FeatureSelectorFactory receives a configuration that contains up to three keywords, separated by an underscore, specifying the main selection strategy and, optionally, a knowledge base to use for prior knowledge retrieval and a second selection strategy to combine. For example, the configuration *Prefilter_ANOVA_KEGG* means that the FeatureSelectorFactory first creates an object of class *ANOVASelector* that implements ANOVA feature selection. It then uses the *KnowledgeBaseFactory* class — which encapsulates the creation logic of a knowledge base — to create the corresponding KEGG knowledge base object. These two objects are then handed over to a new instance of class *PreFilterSelector* that implements the actual prefiltering strategy.

We have set up an inheritance structure that splits up into multiple types of feature selectors that provide specialized functionality: the *JavaSelector* and *RSelector* classes provide functionality to invoke R and Java code, respectively. Class *PythonSelector* provides the functionality to implement statistical feature selection strategies that make use of Python's *scikit-learn* package, e.g. for using RandomForest [152]. Class *PriorKnowledgeSelector* constitutes the super class for any feature selection approach using prior knowledge. For example, the *LassoPenaltySelector* inherits from both *PriorKnowledgeSelector* and *RSelector*. In this way it is able to hold a knowledge base object to retrieve prior knowledge and forward it to an external R implementation that applies the actual feature selection strategy. Prior knowledge approaches are further refined to those inheriting from *CombiningSelector*, which combines an object of another feature selection implementation with a particular knowledgebase, e.g. as described above for *Prefilter_ANOVA_KEGG*, and *NetworkSelector*, which select networks, e.g. pathways, as features instead of the original genes. A class inheriting from *NetworkSelector*

additionally requires a *FeatureMapper* class that maps the feature space from genes to networks. The actual mapping strategies that generate feature values, e.g. pathway activity scores, are implemented in dedicated classes by inheriting from the original FeatureMapper class.

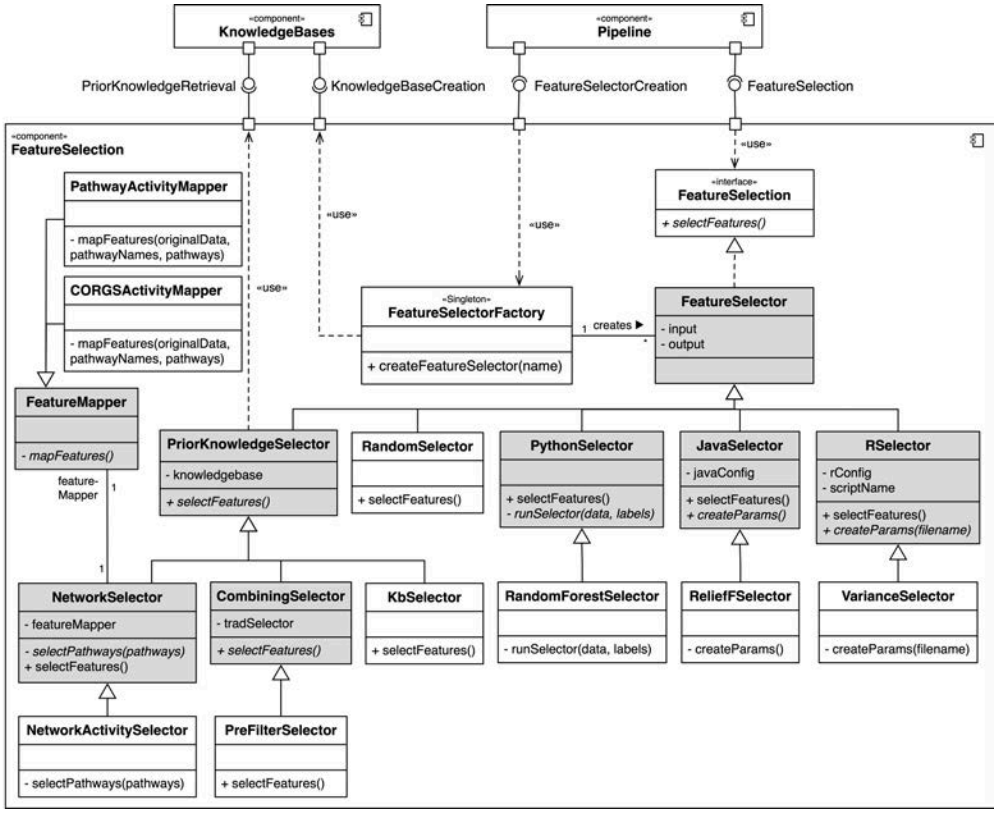


Fig. 5.4: Class structure of the FeatureSelection module, with abstract classes in grey. Every feature selection approach is implemented in a distinct class, which inherits from the main *FeatureSelector* class or one of its inheriting abstract classes and implements the *FeatureSelection* interface. This method is invoked during pipeline execution. A *FeatureSelectorFactory* is responsible for creating new *FeatureSelector* objects from a given keyword.

5.4.3 Enabling a Comprehensive Result Assessment

Comprior allows for a flexible pipeline design by offering a broad range of functionality for assessing input data quality, knowledge base coverage, and feature selection results. Figure 5.5 depicts the class architecture of the evaluation component. The external *Pipeline* component creates and orchestrates objects of type *Evaluator*. Each evaluation aspect is implemented in a dedicated class that inherits from the main abstract *Evaluator* class. Any added functionality must be incorporated by one of the existing classes if it fits to the type of evaluation, e.g. a metric for feature relevance should be added to class

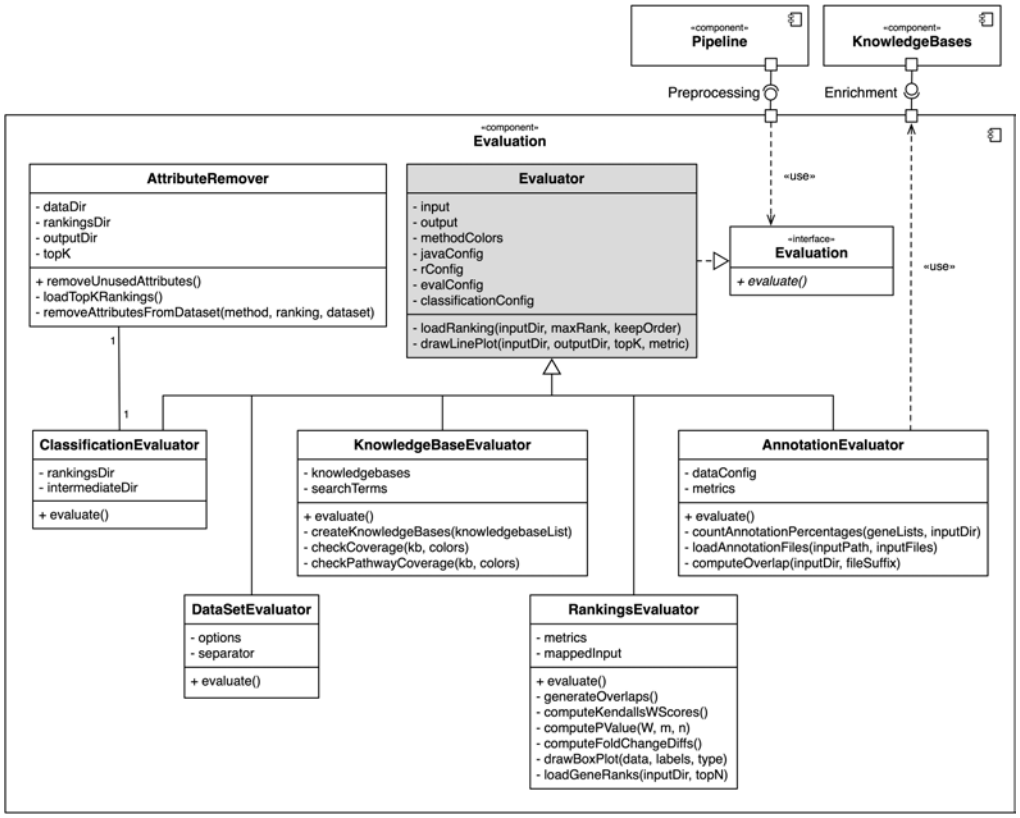


Fig. 5.5: Class structure of the Evaluation component (abstract classes in grey). On top of the hierarchy is an abstract *Evaluator* class that forces its inheriting classes to realize the interface *Evaluation*, which is invoked during pipeline execution. Actual evaluation functionality is realized in distinct inheriting classes.

RankingsEvaluator that provides functionality for assessing and comparing the feature sets directly.

The *DataSetEvaluator* class visualizes aspects on the input data quality. The *KnowledgeBaseEvaluator* class queries each knowledge base used for feature selection with the provided search terms, and summarizes the amount of returned genes — or pathways — and their association scores in an overview plot. The *AnnotationEvaluator* class is responsible for annotating the retrieved feature sets with additional biological information and creates overlap plots for visualization. The AnnotationEvaluator uses EnrichR and one of its corresponding libraries selected by the user to a) annotate the feature sets with terms and b) search for terms that are enriched in the feature set. The *RankingsEvaluator* class provides measures to assess the actual feature rankings, e.g. by comparing them to each other, and creates corresponding plots. The *ClassificationEvaluator* class is responsible for carrying out the full classification procedure for a given input data set and feature rankings. Assigned to it is an object of class *AttributeRemover*, which reduces the input data set's features to those of only the given feature rankings. The evaluation

component for classification uses WEKA functionality for k-fold cross validation and computation of the performance metrics [79].

5.4.4 Handling Multiple Identifier Formats

Identifier handling is a recurrent and cumbersome necessity when processing biological data. There are numerous identifier formats for genes and microarray probes. Comprior provides built-in mapping strategies that can be applied throughout the complete workflow when necessary, e.g. mapping the input data sets, but also prior knowledge retrieved from a knowledge base. Users of Comprior only specify the current identifier formats of their input data sets, e.g. Ensembl gene identifiers, and select a desired output format, e.g. Human Gene Nomenclature (HGNC). In the following, we describe the mapping strategies that are applied by Comprior.

It cannot be ruled out that identifiers have a many-to-many mapping, i.e. $m : n$ cardinality. Consequently, an identifier of the original format can be mapped to multiple new identifiers in the desired format ($1 : n$). Vice versa, multiple identifiers of the original format can be mapped to the same identifier of the desired format($m : 1$). Both cases can introduce redundancy that can be problematic for downstream analyses. Depending on the use case, Comprior must deal with such multiple-cardinality mappings. The mapping functionality is located in Comprior's Utility component. However, to remain lightweight and flexible, the actual identifier mapping is not carried out by Comprior itself. Instead, Comprior accesses the online service of g:Profiler via a corresponding implementation in the KnowledgeBase component [174]. g:Profiler is an online tool that offers a range of services to characterize lists of biological identifiers, for example via enrichment analysis. g:Convert, as part of g:Profiler, allows identifier conversion and uses the Ensembl databases, which covers all popular identifier formats and many more from different species and experimental platforms [87]. This allows Comprior to flexibly handle input data from basically any format and map it to any of the formats available in the Ensembl database.

There are two cases in which identifier mapping is necessary during a benchmarking process in Comprior. First, when the user has selected an output identifier format that is different from the one of the input data sets. Second, when prior knowledge is retrieved from a knowledge base that mismatches the format of the input data. In the following, we describe how Comprior handles many-to-many mappings for the described cases by separating into $m : 1$ and $1 : n$ mappings. For simplicity, we assume a gene expression data set to have gene identifiers in the columns.

$m : 1$ Mappings

Figure 5.6 describes how Comprior deals with $m : 1$ mappings for two data artifacts: the input data and prior knowledge. In the first case — an $m : 1$ mapping occurs when mapping a gene expression matrix — the mapped data set would contain m columns with

the same identifier but different expression levels. As expression matrices require unique column identifiers, these m columns must be reduced to one. There is no best practice regarding which of the columns should be kept, or if their values should be merged. Thus, Comprior keeps the first column and removes all other columns with a duplicate column identifier. Here, we assume that all m columns show a similar expression profile. As the original identifiers are converted to the same new identifier, they seem to represent the same entity and their expression profiles should therefore be very similar. If one of the original identifiers would be selected during feature selection, it is likely that the other identifiers would be selected as well.

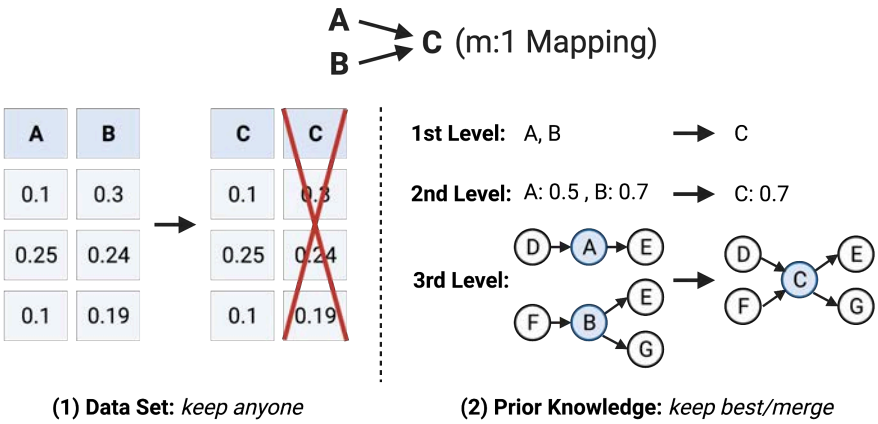


Fig. 5.6: Comprior internally deals with $m : 1$ mappings differently, depending on whether these occur when mapping (1) an expression data set or (2) prior knowledge. In the expression data set, Comprior keeps any one column of the m duplicated identifiers, removing the rest. Depending on the level of prior knowledge for which identifier mapping takes place, Comprior keeps highest entries or merges networks.

In the second case, an $m : 1$ mapping occurs when mapping prior knowledge retrieved from a knowledge base to a new identifier format. As we have different levels of prior knowledge — list of identifiers, list of scored identifiers, and networks — Comprior must address this issue for all levels. For first-level prior knowledge, an $m : 1$ mapping results in a list with m duplicate entries. Comprior will keep the first occurrence of the identifier and remove the others. For second-level prior knowledge, an $m : 1$ mapping results in a list with m tuples containing the same identifier, but different relevance scores. Comprior will keep the tuple with the highest score and remove all others. For third-level prior knowledge, an $m : 1$ mapping leads to m nodes in a network that must be merged. As a result, Comprior creates a new single node in the mapped network and assigns to it all the incoming and outcoming edges of the m original nodes.

$1 : n$ Mappings

Figure 5.7 describes how Comprior deals with $1 : n$ mappings for both the input data and prior knowledge. For gene expression data sets, an $1 : n$ identifier mapping results in unique column identifiers, but n columns will have the exact same expression levels. Keeping all of these columns introduces redundancy into the data and causes problems in downstream analysis because some tools require columns with unique expression profiles. Again, there is no best practice regarding how to deal with $1 : n$ mappings in gene expression data sets. Thus, Comprior keeps the first column in the data set and removes duplicates.

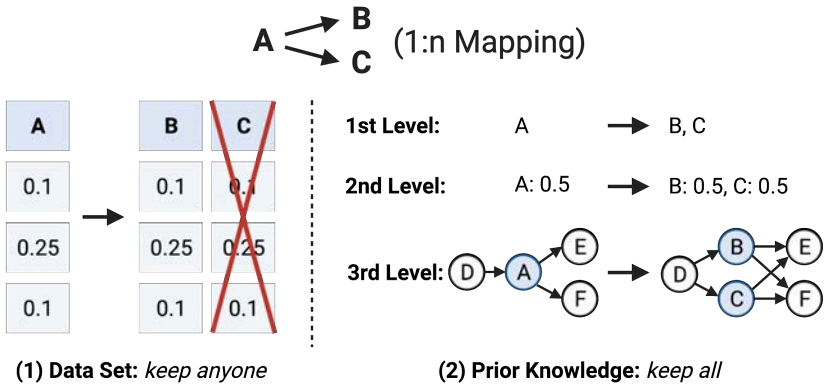


Fig. 5.7: Comprior deals with $1 : n$ mappings which occur when mapping (1) an expression data set and (2) prior knowledge. For expression data sets, Comprior keeps only one identifier and removes remaining columns. For prior knowledge, Comprior keeps all n mapped identifiers.

If a $1 : n$ mapping occurs for prior knowledge, Comprior keeps all n mapped identifiers. This way, we can ensure that a gene is assigned the related prior knowledge and that no relationships are accidentally removed. For first-level prior knowledge, a $1 : n$ mapping has no implications except that the list of identifiers grows. The same applies for second-level prior knowledge, where n identifiers will have the same score assigned. For third-level prior knowledge, i.e. networks, Comprior has to create n new nodes in the network with the same incoming and outgoing edges.

5.5 Uniform Access to Prior Knowledge

One of Comprior's key features is the uniform access to prior knowledge. Researchers that want to implement a new prior knowledge approach and make use of the available knowledge bases should not have to deal with the issue of accessing them and converting their result to a uniform format. Consequently, Comprior encapsulates prior knowledge retrieval and provides clear interfaces that return prior knowledge of the different levels

as defined in Section 4.1. In the following, we describe the technical realization of this uniform access in Comprior.

5.5.1 Knowledge Base Concept

Figure 5.8 depicts how the concept of a knowledge base is realized in Comprior. For the sake of clarity, we model an example knowledge base here instead of the actually implemented classes. Figure 10.1 in the appendix provides the complete class architecture of the KnowledgeBase component. As with feature selectors, the creation logic for knowledge bases is encapsulated into a factory class called *KnowledgeBaseFactory*. Given a particular keyword, this class creates the corresponding knowledge base object and assigns it a web service query class and — if the knowledge base provides network information — a pathway parsing class. Conceptually, a knowledge base consists of at least two classes. The first class inherits from the abstract *KnowledgeBase* class and realizes the interface *PriorKnowledgeRetrieval*, which is accessed by Comprior's components to retrieve prior knowledge. The interface *PriorKnowledgeRetrieval* contains three methods that correspond to the three levels of prior knowledge we defined in Section 4.1: *getRelevantGenes()* returns first-level prior knowledge, *getGeneScores()* returns second-level prior knowledge, and *getRelevantPathways()* returns third-level prior knowledge.

The second class — generally written in upper case letters — inherits from Bioservices' REST class and retrieves the actual prior knowledge from the corresponding web service [42]. Bioservices offers web service query implementations for many biological knowledge bases. If such an implementation is not yet available for a knowledge base, it can be implemented correspondingly. If a knowledge base provides network information, e.g. KEGG and PathwayCommons, it requires additional functionality to transform the network information, which can be provided in many different formats, into a uniform format that can be used throughout Comprior. For this, a knowledge base gets assigned an additional class that inherits from the abstract *PathwayParser* class and implements a transformation strategy for the individual knowledge base.

5.5.2 Processing Pathways

As already stated, Comprior needs to transform network information from many heterogenous formats into a uniform format that can be used throughout Comprior. For this, a knowledge base gets assigned an additional class that inherits from the abstract *PathwayParser* class, e.g. *KEGGPathwayParser*, and implements a transformation strategy for the individual knowledge base. The uniform data structure used for network information is provided by Pypath [235]. Pypath is a flexible Python package that provides multiple administrative methods, e.g. for retrieving interaction partners, and even allows a single network from multiple input networks to be constructed. We chose this package as it already provides a well-thought data structure with lots of additional functionality for pathway analysis, e.g. to easily retrieve interaction partners and add annotation

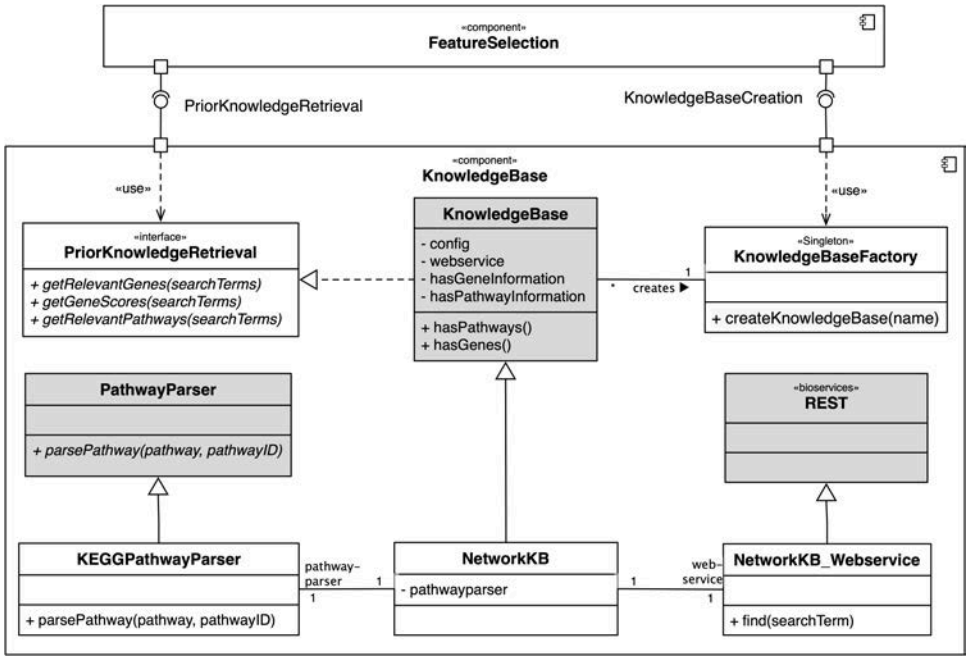


Fig. 5.8: Example class structure for a knowledge base providing network information. Class *NetworkKB* inherits from the abstract *KnowledgeBases* class and implements the interface *PriorKnowledgeRetrieval*. Class *NetworkKB_Webservice* retrieves the actual prior knowledge from a web service via its REST API. Class *NetworkKB_PathwayParser* parses the network information and transforms it into a uniform format.

information, and is flexible enough to be useful for future implementations, e.g. when compiling a network from multiple other networks.

5.5.3 Mapping Prior Knowledge Levels

Comprior currently provides uniform access to prior knowledge from Open Targets, DisGeNET, KEGG, and PathwayCommons. While Open Targets and DisGeNET provide gene associations, i.e. first- and second-level prior knowledge, KEGG and PathwayCommons provide pathway information, i.e. third-level prior knowledge, only. Nevertheless, all of them must implement the interface *PriorKnowledgeRetrieval* and, with this, provide all three levels of prior knowledge as described in Section 4.1.

Comprior implements mapping of higher-level prior knowledge to a lower level by applying the mapping strategies described in Section 4.2.1. To map third-level prior knowledge from KEGG and PathwayCommons, i.e. pathways, to first-level prior knowledge, i.e. a list of entities, Comprior extracts the nodes of all retrieved pathways and joins them to a final set. To map third-level prior knowledge from KEGG and PathwayCommons to second-level prior knowledge, i.e. a list of scored entities, Comprior computes a gene score s_i for a gene i, as described in Section 4.2.1, based on the connectedness of a gene in the networks it participates in.

In Section 4.2.1 we discussed that there are certain restrictions when mapping lower-level prior knowledge to a higher level. As the majority of currently implemented prior knowledge approaches do not require third-level prior knowledge, we have postponed an actual implementation of these strategies to a future version of Comprior and refer users to use the offered network approaches in combination with KEGG and PathwayCommons.

5.6 Summary

In this chapter, we presented Comprior as a benchmarking tool for feature selection approaches that specifically addresses the needs of prior knowledge approaches. Comprior supports the complete benchmarking process from pipeline design to execution and result set visualization. Furthermore, it unifies access to prior knowledge and allows for comprehensively assessing both prior knowledge and traditional feature selection approaches regarding their quantitative performance and biological relevance. Instead of being constrained by heterogeneous knowledge base information, data harmonization, and complex benchmark setups, researchers can now concentrate on the development of novel feature selection approaches and flexibly combine them with multiple knowledge bases or traditional approaches.

6

Assessment of the Impact of Prior Knowledge Integration During Biomarker Detection

This chapter provides a comprehensive case study on prior knowledge approaches that examines their performance under different aspects. We use six gene expression data sets from three disease domains to evaluate both selected traditional and prior knowledge approaches. We combine the applied prior knowledge approaches with multiple knowledge bases to examine their impact on result sets. We assess the effectiveness of the tested approaches based on their classification performance, agreement of feature sets, and semantic similarity of the retrieved enrichments. We also examine the aforementioned aspects in a cross-validation setting, which provides insights into the robustness of the tested approaches and biological relevance of the identified feature sets. In the following, we describe how we prepared our data sets, outline the experiment setup, and present our evaluation results.

6.1 Data Sets

We have selected both microarray and RNAseq data sets from Alzheimer's disease, glioma, and breast cancer. We selected our data sets based on their dimensionality, i.e. more than 100 samples and 15.000 features, the availability of disease subtype information, and the availability of a second, complementary data set with matching subtype information, as we want to use the subtype information for sample classification. Table 6.1 shows an overview of the data sets, available sample classes, their sources, and final dimensions after preprocessing.

For Alzheimer's disease, we downloaded the AddNeuroMed cohorts I and II from the Gene Expression Omnibus (GEO, accession numbers *GSE63060* and *GSE63061*) [17, 206]. Based on the available metadata, we grouped samples of both data sets into classes of *Alzheimer's disease*, *mild cognitive impairment*, and *normal*. We removed all samples that had been assigned other group information or had no metadata information available. For glioma, we downloaded data that was acquired by the Cancer Genome Atlas (TCGA) and the REMBRANDT project from the Genomic Data Commons (GDC) and GEO (accession number *GSE108474*), respectively [71, 77]. For the TCGA data set,

	Subtypes/ Classes	Data Set	Source	Technology	#Samples	#Genes/ Probes
Alzheimer's Disease	Alzheimer's disease, mild cognitive impairment, normal	AddNeuroMedI [206]	GEO (*GSE63060*)	Microarray	329	26,325
		AddNeuroMedII [206]	GEO (*GSE63061*)	Microarray	383	32,049
Glioma	glioblastoma multiforme, astrocytoma, oligodendroglioma	TCGA-GBM/ LGG [27, 223]	GDC (tumor primary)	RNAseq	496	19,301
		REMBRANDT [77]	GEO (*GSE108474*)	Microarray	436	31,442
Breast Cancer	luminal A, luminal B, HER2-enriched, basal-like, normal-like	TCGA-BRCA [222]	GDC (tumor primary)	RNAseq	1,090	20,950
		SCAN-B [30]	GEO (*GSE96058*)	RNAseq	378	15,011

Table 6.1: Overview of the data sets used and their available subtype information, sources, technology of origin, and final data set dimensions after preprocessing. Abbreviations: Gene Expression Omnibus (GEO), Genomic Data Commons (GDC).

we combined tumor primary samples from lower-grade glioma (LGG) and glioblastoma multiforme (GBM) to a final data set (TCGA-GBM/LGG) [27, 223]. Based on the metadata, we grouped samples of both TCGA-GBM/LGG and REMBRANDT data sets by their histological subtypes into classes *glioblastoma multiforme*, *astrocytoma*, and *oligodendroglioma*. Analogously to the AddNeuroMedI and AddNeuroMedII data sets, we removed all samples that had no or other histological subtypes assigned. For breast cancer, we downloaded data acquired by the TCGA breast cancer (TCGA-BRCA) project via GDC and the Sweden Canceromics Analysis Network - Breast (SCAN-B) initiative via GEO (accession number *GSE96058*) [30, 222]. Based on the metadata, we grouped samples of both data sets into classes corresponding to the PAM50 breast cancer subtypes of *luminal A*, *luminal B*, *HER2-enriched*, *basal-like*, and *normal-like* and removed samples of other subtypes [150].

All data sets were provided at different preprocessing stages, e.g. raw or normalized counts, and thus required us to apply additional measures for preprocessing. Table 6.2 describes how the data sets were processed. Purple cells denote that this preprocessing step was already conducted by the data sets' authors. Light purple cells denote that we had to extend the preprocessing conducted by the original authors. The corresponding R scripts for downloading and preprocessing the data sets are provided in a public GitHub repository [157].

The expression levels of the AddNeuroMedI and AddNeuroMedII data sets were previously transformed by the original authors with variance-stabilizing transformation and normalized by upper quantiles (VST-UQ) [121]. As no filtering was applied beforehand, we removed lowly expressed genes by filtering those that had expression levels below

Disease	Data Set	Filtering	Normalization	Outlier Removal	Batch Effects?
Alzheimer's Disease	AddNeuroMedI	genes with expr. levels <6.0 in >30% of samples	VST-UQ	-	PCA (see Figure 10.2c in appendix)
	AddNeuroMedII	genes with expr. levels <6.0 in >30% of samples	VST-UQ	-	PCA (see Figure 10.2f in appendix)
Glioma	TCGA-GBM/LGG	genes <70 counts in >30% of class samples	TMM	-	PCA (see Figure 10.2b in appendix)
	REMBRANDT	genes with expr. level <7.5 in >30% of samples	MAS5	-	authors controlled for batch effects, PCA (see Figure 10.2e in appendix)
Breast Cancer	TCGA-BRCA	genes <60 counts in >30% of class samples	TMM	-	PCA (see Figure 10.2a in appendix)
	SCAN-B	genes with median zFPKM <-3.0	zFPKM	6 samples based on PCA (PC1 <-800)	PCA (see Figure 10.2d in appendix)

Table 6.2: Overview of the used data sets and applied preprocessing measures. Purple cells denote preprocessing was conducted by the data sets' original authors; light purple cells denote that we had to extend these preprocessing steps by additional measures. Abbreviations: Variance-Stabilized Transformation with Upper Quantile normalization (VST-UQ), Trimmed Mean of M-values (TMM), Fragments Per Kilobase Million (FPKM), Principal Component Analysis (PCA).

6.0 in more than 30% of the samples. We additionally applied Principal Components Analysis (PCA) on the data to look for batch effects (see also Figures 10.2c and 10.2f in the appendix). However, we could not identify abnormal clusters in the data. The glioma and breast cancer data sets provided by TCGA were available as raw counts. We first removed lowly expressed genes with few counts in more than 30% of the samples. We then applied Trimmed Mean of M-values (TMM) normalization with subsequent log transformation. An applied PCA did not show any batch effects (see also Figures 10.2a and 10.2b in the appendix). The glioma data set from the REMBRANDT study was available with MAS5 normalized and log2 transformed expression levels [91]. As the data set was not filtered for lowly expressed genes before, we applied a soft filter by removing genes with expression levels below 7.5 in more than 30% of the samples. According to the authors, they accounted for batch effects during data processing [77]. We could confirm this in our own PCA (see also Figures 10.2b and 10.2e in the appendix). The SCAN-B data set was originally available as log2 transformed Fragments Per Kilobase Million (FPKM) normalized counts. However, FPKM does not account for library size — which is necessary for inter-sample comparisons — and is not considered to be a robust normalization method anymore [51]. We thus retransformed the data into FPKM values and applied zFPKM normalization, which applies a z-score normalization on the FPKM values and thus allows for inter-sample comparisons [81]. We additionally applied PCA and removed outlier samples as described in Table 6.2; however, we could not detect any batch effects (see also Figure 10.2d in the appendix).

6.2 Experiment Setup

We conduct the complete case study with Comprior, whose implementation details were presented in Chapter 5. As such, we limit our description of the experiment setup to only those details that have not been discussed previously. The complete configuration files used for running the case study with Comprior are provided in the online supplementary material on GitHub [157].

6.2.1 System Specifications

We carry out all experiments on a virtual machine equipped with eight Intel® Xeon® X7560 CPUs running at 2.27GHz clock speed, 24MB L3 cache size, and 64GB of main memory capacity. Each CPU consists of two physical cores running a 64-bit instruction set. The virtual machine uses disk space of 105GB capacity that is a mix of both hard and solid state drives (SSD), which are combined via RAID6 and managed in an EMC VNX 7500 unified storage system. The machine is connected to the internet via a glass fibre cable of 1 GBit/s bandwidth for uploads and downloads, shared across all users of the institution. The virtual machine runs on a Linux 18.04.5 LTS distribution, with installations of Python 3.6.9, R 4.0.2, and openJDK's Java Runtime Environment 11.0.8. The

complete descriptions on Python, Java, and R installations are provided in the online supplementary material on GitHub [157].

6.2.2 Feature Selection Approaches

For feature selection, we apply both traditional and prior knowledge feature selection approaches. Table 6.3 provides an overview of the traditional feature selection approaches applied, which cover filter, wrapper, and embedded approaches. As a baseline approach for traditional approaches, we apply randomly selected genes (*Random*). *ANOVA*, *ReliefF*, *RandomForest*, and *Lasso* are provided by the sci-kit learn Python package and integrated within Comprior [25, 106, 152, 230]. With the exception of *Lasso*, which runs with iterative parameter fitting via 10-fold cross-validation in steps of 0.01, all approaches are used with their default settings. SVM-RFE is available in the WEKA tool suite (Java) and used with polynomial kernel and default settings [78, 79].

Class	**Traditional Approaches**
Baseline	Random
Filter	ANOVA
	ReliefF [106]
Wrapper	SVM-RFE (*SVMpRFE*) [78]
Embedded	Lasso [230]
	RandomForest [25]

Table 6.3: Overview of the applied traditional feature selection approaches. As a baseline approach, we use randomly selected features.

Table 6.4 provides an overview of the applied prior knowledge approaches and their combinations with traditional approaches and knowledge bases. As a baseline approach, we exclusively apply feature selection based on the retrieved prior knowledge (*KBonly*). Utilized prior knowledge approaches cover all three complexity levels of modifying, combining, and network approaches. As modifying approaches, we use a prefiltering and extension approach (*Pref* and *Ext* prefixes, respectively) as described in Section 4.4.1 in combination with ANOVA, SVM-RFE (*SVMpRFE*), and Lasso. As combining approaches, we use our weighted approach (*Weight*) as described in Section 4.4.5, again in combination with ANOVA, SVM-RFE, and Lasso. We additionally use the method developed and implemented in R by Zeng et al. (*LassoPenalty*), where second-level prior knowledge, e.g. gene-disease associations, is incorporated as a feature-specific penalty term during Lasso computation [261]. As network approaches, we use both our own NetworkActivity (*NetAct*) approach as described in Section 4.4.6 and the approach described by Lee et al., which we reimplemented in Comprior (*CORGS*) [113]. All modifying and combining approaches are carried out in combination with each of the available knowl-

edge bases, namely DisGeNET, Open Targets, KEGG, and PathwayCommons. The final approach names as used in the subsequent results section are constructed from combining the individual approach names or abbreviations (in brackets), e.g. a prefiltering approach using ANOVA in comination with DisGeNET is named *Pref_ANOVA_DG*. However, as Comprior currently does not provide strategies to transform prior knowledge from a lower to a higher level, we carried out the network approaches in combination only with KEGG and PathwayCommons as these are the only knowledge bases providing third-level prior knowledge.

Class	**Approach**	**... Combined with Trad. Approach**	**... Combined with Knowledge Base**
Baseline	KBonly	—	DisGeNET (*DG*) Open Targets (*OT*) KEGG PathwayCommons (*PC*)
Modifying	Prefilter (*Pref*)	ANOVA	
	Extension (*Ext*)	SVM-RFE(*SVMpRFE*)	
Combining	Weighted (*Weight*)	Lasso	
	LassoPenalty [261]	—	
Network	CORGS [113]	—	KEGG PathwayCommons (*PC*)
	NetworkActivity (*NetAct*)	—	

Table 6.4: Overview of applied prior knowledge approaches and their combinations with traditional approaches and knowledge bases. The final method names are constructed from combining the individual approach names or abbreviations (in brackets), e.g. a prefiltering approach using ANOVA in comination with DisGeNET is named *Pref_ANOVA_DG*.

6.2.3 Identifier Mapping

Identifiers from all data sets and retrieved prior knowledge are mapped from their original format to the Human Gene Nomenclature (HGNC) format. Occuring $n : m$ mappings are resolved by Comprior internally as described in Section 5.4.4. However, we prepare data sets used for cross-validation differently: when a $1 : n$ mapping occurs, i.e. one identifier in the original format is mapped to multiple identifiers in the desired format, we keep all n mappings. This way, we ensure that no feature from a feature set, when applied on the cross-validation data set, is accidentally ignored because its mapped identifier was previously removed from the cross-validation data set.

6.2.4 Prior Knowledge Retrieval

Prior knowledge is retrieved with particular search terms via Comprior from KEGG, Open Targets, DisGeNET, and PathwayCommons. As search terms, we use the main disease name, class labels, and their corresponding synonyms as searched for in the National Cancer Institute's (NCI) metathesaurus browser [140]. The complete lists of the applied search terms per disease are provided in Tables 10.5 to 10.7 in the appendix.

6.2.5 Classification

We use the identified feature sets to classify samples of both the original and cross-validation data sets into their disease subtypes. For example, we select feature sets from the TCGA-BRCA data set and use these features to classify samples of both the TCGA-BRCA (original) and SCAN-B (cross-validation) data sets into their PAM50 subtypes. For every classification, we apply ten-fold cross-validation, which we identified to be one of the standard cross-validation methods applied in related work (see also Table 10.12 in the appendix). In order not to give preference to a feature selection approach for a single classifier, we apply the feature sets to five different classifiers which, according to Tabares et al., are among those most commonly used [220]: Naive Bayes, Linear Regression, Support Vector Machines, Random Forest, and k-Nearest neighbor ($k = 3$). Unless stated otherwise, the classification performance depicted by subsequent plots corresponds to the average classification performance of these classifiers. As most of our data sets are imbalanced (see also Tables 10.2 to 10.4 in the appendix providing class sizes), we measure classification performance via Matthew's Correlation Coefficient (MCC), as this measure — in contrast to the popularly used classification accuracy or F_1 measure — is more reliable for imbalanced data sets.

6.2.6 Enrichment Analysis

For assessing the biological relevance of a feature set, we use Enrichr to annotate feature sets and retrieve enriched terms [35, 255]. It is important to note that we only conduct the subsequently described enrichment analysis via Enrichr for modifying and combining prior knowledge approaches, wheras we proceed differently for network approaches. We filter out terms with an adjusted p-value above 0.05 and sort the remaining terms in descending order by their combined score, which is provided by Enrichr. The complete annotation and enrichment functionality is also implemented in Comprior. However, the prior knowledge used during feature selection must not overlap with the biological information used for annotation and enrichment, as this would introduce a bias towards prior knowledge approaches in subsequent assessments. Thus, for the Alzheimer's disease data sets, we use Gene Ontology's Biological Processes (gene set library *GO_Biological_Process_2018*), as they also contain annotations specific to Alzheimer's disease from the Alzheimer's Gene Ontology Annotation (GOA) initiative [109, 226]. For the glioma data sets, we also use *GO_Biological_Process_2018*, as GO is frequently used for annotation and this library covers many more genes than most of the other gene set libraries available in Enrichr. For the breast cancer data sets, we use the oncogenic signatures of MSigDB (*MSigDB_Oncogenic_Signatures*) to also have a more cancer-specific annotation [118]. Both GO and MSigDB are not included in any of the knowledge bases applied by a prior knowledge approach during feature selection. As network approaches have pathways instead of genes as features, a gene-based enrichment analysis as provided by Enrichr is not possible. Instead, we consider the pathways retrieved as prior knowledge to be an enrichment if their relevance score

computed during feature selection is below 0.05, i.e. the member genes of the pathway show an expression behavior that is related to the disease subtypes.

6.2.7 Feature Set and Enrichment Robustness

We assess the robustness of individual feature rankings by comparing feature rankings retrieved by a feature selection approach on two data sets of the same disease domain. We compare these feature rankings by using the rank-biased overlap (RBO), which takes into account feature overlaps and further allows us to give more weight to highly ranked features [248]. The RBO of two feature rankings ranges between 0 and 1, meaning the two feature sets are completely disjunct or identical, respectively.

We assess the robustness of enrichments by comparing enrichments retrieved for an approach on two data sets of the same disease domain. For this, we compute a similarity score for two sets of enriched terms and then applying the Best Matching Average (BMA) method, which was demonstrated to be a useful method in the MegaGO package [197, 243]. For GO terms, we use Lin's semantic similarity measure [120]. For both MSigDB oncogenic signatures and pathways, we compute similarities based on overlapping member genes using the Dice coefficient. The similarity between two sets of enriched terms ranges between 0 and 1, which indicates whether the two sets are completely unrelated or identical, respectively.

6.3 Results

This section presents the results we obtained from running the case study in the aforementioned setting on all six data sets. Results cover an examination of the information content in knowledge bases for the applied search terms, execution runtimes, classification performances, feature sets, and enrichments; the latter three were further evaluated in a cross-validation setting for robustness assessment. Central to our case study are the following questions, which we answer on the basis of our obtained results in the next subsections:

Q1: *Do we already have a sufficient coverage of knowledge bases?*

Q2: *Are prior knowledge approaches feasible in terms of runtime?*

Q3: *How do prior knowledge approaches compare to traditional approaches?*

Q4: *Compared to a highly complex traditional approach, is it sufficient to use a low-complexity prior knowledge approach?*

Q5: *What are the benefits of applying a dense integration of prior knowledge compared to simple filtering strategies?*

Q6: *How much does the choice of a knowledge base affect results?*

In the following subsections, we first present evaluation results by individual criteria and subsume the major findings at the end of each subsection.

6.3.1 Coverage of Diseases in Knowledge Bases

In this subsection, we examine how much biological information related to our chosen disease domains is currently provided by knowledge bases, which addresses the earlier posed question: *Q1: Do we already have a sufficient coverage of knowledge bases?* We assess the coverage of a disease in a knowledge base on the basis of the available information, i.e. prior knowledge, for a set of disease-specific search terms. For every disease domain, we examine how much prior knowledge is returned for the applied disease-specific search terms by a knowledge base. We conclude the key insights at the end of this subsection.

Figure 6.1 depicts summary statistics on prior knowledge retrieved from our four knowledge bases for breast cancer, glioma, and Alzheimer's disease. For DisGeNET and Open Targets, bars show how many genes were returned for a search term (left-hand y axis) and boxes show the distributions of association scores that were assigned to these genes in second-level prior knowledge (right-hand y axis). For PathwayCommons and KEGG, bars show the number of pathways retrieved per search term (left-hand y axis) and boxes show the number of contained genes (right-hand y axis). We group search terms for every disease into main disease and its subtypes, e.g. "Breast Cancer" containing more general terms like "Breast Carcinoma" and "HER-2" comprising subtype-specific search terms like "ERBB2 Overexpressing Subtype of Breast Carcinoma". For the sake of clarity, we use numerical identifiers for the search terms. The actual search terms for the distinct diseases are provided in Tables 10.5 to 10.7 in the appendix. It is important to note that the prior knowledge retrieved for different search terms is not disjunct, i.e. prior knowledge retrieved for different search terms can contain the same entities.

From Figure 6.1 we observe a generally low coverage for DisGeNET. While there are multiple thousand genes returned for the general disease names, the overall relevance scores remain low, with the majority of genes showing a score far below 0.1 across all three diseases. When computing the relevance score, DisGeNET takes into account the number and type of original sources and the number of publications supporting the association [165]. Consequently, a score close to zero means that there were only a few original sources and publications found that support an association between a particular gene and the applied search term. Furthermore, the subtype-specific search terms seem to always acquire the same prior knowledge for a subtype, which contains a few hundred genes with scores close to zero.

For Open Targets, we observe an improved coverage. While there is still identical prior knowledge retrieved for particular search terms that fall into the same category, e.g. for breast cancer search terms 27, 28, 30, and 31, the retrieved prior knowledge still contains multiple thousand genes for search terms of both the main disease and the subtypes.

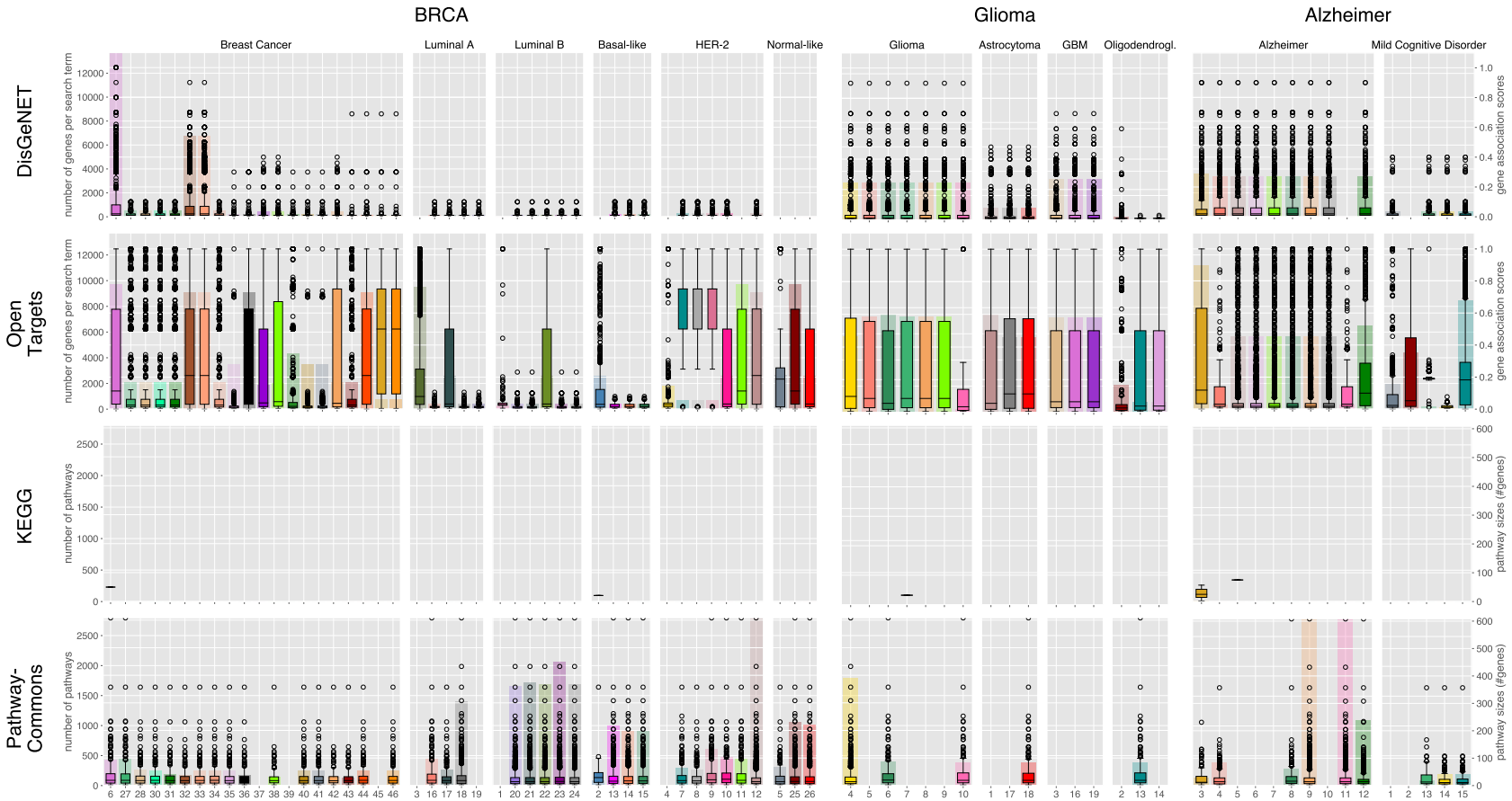


Fig. 6.1: Summary on prior knowledge coverage across the four knowledge bases for breast cancer, glioma, Alzheimer's disease and their respective subtypes. For every search term, bars show the number of returned genes (or pathways for KEGG and PathwayCommons), whereas boxes show the retrieved association scores (or pathway sizes for KEGG and PathwayCommons). DisGeNET provides results, although likely being duplicates, for nearly every search term across all diseases. However, most often association scores are below 0.1. Open Targets shows highest coverage for all diseases, returning results for every search term applied and reaching highest association scores. KEGG shows lowest coverage for all three diseases with few pathways returned for only single search terms. PathwayCommons returns multiple hundreds up to thousands of pathways for most of the applied search terms, where most of the pathways contain up to 50 genes.

Furthermore, the genes contained in prior knowledge have higher relevance scores, with upper quartiles often reaching a relevance score of 0.4 or higher, especially for breast cancer and glioma related search terms (including their subtypes). As Open Targets is a meta knowledge base which integrates many different sources, a moderate relevance score indicates that associations were found in multiple original sources.

For KEGG, we observe a coverage worse than for DisGeNET. Prior knowledge can only be retrieved for single search terms, with each containing few pathways with 20 to 80 member genes respectively. Only for Alzheimer's disease search terms was KEGG able to provide more than a single pathway for a particular search term. While the availability of prior knowledge for only a few tens of genes in few, single pathways can become problematic for prior knowledge approaches in general, it can become especially problematic for network approaches that select their features from retrieved pathways.

For PathwayCommons, on the contrary, we observe a better coverage than KEGG, especially for the individual subtypes. The majority of prior knowledge retrieved contains rather small-sized pathways, often containing less than 50 genes. However, search terms for most of the disease subtypes acquire prior knowledge comprised of multiple hundreds to thousands of pathways, with the highest numbers for Luminal B and HER-2 search terms.

Open Targets and PathwayCommons Have the Highest Coverage, Whereas DisGeNET and KEGG Provide Limited Prior Knowledge

From the four knowledge bases examined, we find that Open Targets and PathwayCommons provide the highest information content to be used as prior knowledge, whereas KEGG, in particular, provides limited prior knowledge and thus does not seem to be suitable for automated prior knowledge retrieval. We observe that Open Targets shows the highest coverage for all of our three diseases and their subtypes, generally returning many genes with moderate relevance scores. Despite also being a meta knowledge base, coverage in DisGeNET cannot keep up with Open Targets, returning many genes with low scores. It remains to be seen if and how these differences in score level affect feature selection results. PathwayCommons has proven to generally return many small-sized, subtype-specific pathways. KEGG, on the contrary, always provides only a handful of small-sized pathways related to general disease terms. KEGG thus provides a good example for examining how a low and unspecific prior knowledge coverage will affect feature selection results. The observations described are consistent across all of the three examined diseases, where the highest coverages were found for breast cancer, followed by glioma and then Alzheimer's disease.

6.3.2 Runtime Performance

In this subsection, we examine the runtime performances of both traditional feature selection and prior knowledge approaches to assess their applicability, which addresses

the earlier posed question: *Q2: Are prior knowledge approaches feasible in terms of runtime?* For this, we measure the average time needed for prior knowledge retrieval and compare absolute runtimes of both traditional feature selection and prior knowledge approaches. We conclude this subsection by summarizing our key findings.

Runtimes for Prior Knowledge Retrieval

We measure the time needed for retrieving prior knowledge from a knowledge base across every data set and prior knowledge approach used (see also Table 6.4). Figure 6.2 shows the corresponding average runtimes as a bar plot, with error bars representing overall variance. Runtimes are grouped per knowledge base, colors indicate the disease for which prior knowledge was retrieved.

In general, runtimes for prior knowledge retrieval from DisGeNET, KEGG, and Open Targets are on an acceptable level, with average runtimes between 24 and 66 seconds for DisGeNET, 27 and 70 seconds for KEGG, and 63 and 152 seconds for Open Targets. Strikingly, retrieval times for prior knowledge from PathwayCommons differ noticably from those of the other knowledge bases, with runtimes between 1,050 and 1,770 seconds, which correspond to 17 to 30 minutes. The high runtimes are caused by parsing the retrieved pathways into a uniform format. If KEGG would return more than only a few pathways, we would observe higher runtimes as well.

Figure 6.2 also shows that retrieval times for a particular knowledge base differ across diseases. These differences can be explained by a) the number of search terms applied for prior knowledge retrieval and b) the amount of prior knowledge retrieved. For example, KEGG requires the longest runtimes of 70 seconds for retrieving prior knowledge for breast cancer, although, compared to the other diseases, it does not return many pathways. However, KEGG does not allow bulk queries, which results in single queries for every search term. As we use 45 search terms for breast cancer — compared to 15 and 19 search terms used for Alzheimer's disease and glioma, respectively — runtime increases compared to the other diseases. In contrast, retrieval of prior knowledge from KEGG for Alzheimer's disease takes longer than for glioma, although less search terms were used. This is caused by the number of pathways retrieved; while KEGG returns only two pathways for glioma search terms, it returns 15 for Alzheimer's disease search terms. These pathways must be parsed into Comprior's internal format, which is computationally expensive. We can also observe the described relationships between runtime, the number of search terms, and the returned amount of prior knowledge for PathwayCommons and Open Targets, which also do not allow for bulk queries. In contrast, DisGeNET is the only knowledge base allowing for such queries, which is why the runtimes correspond directly to the amount of prior knowledge retrieved.

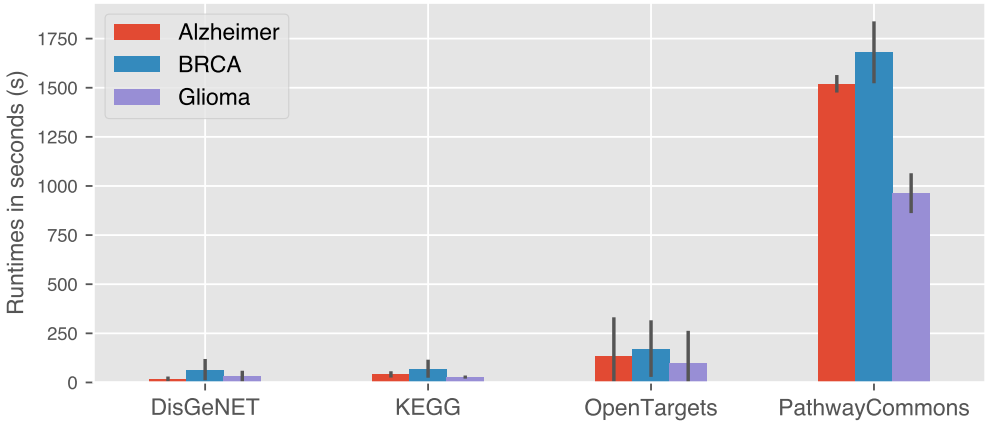


Fig. 6.2: Average time needed to retrieve prior knowledge for a particular disease from the different knowledge bases. Black lines represent variances.

Comparing Absolute Runtimes

During our experiments, we also measured the runtimes of different steps of the feature selection process, namely traditional feature selection, prior knowledge retrieval, feature mapping, and remaining tasks. Figure 6.3 depicts runtime performances or the applied feature selection approaches on the TCGA-BRCA data set grouped by approach category, i.e. traditional, modifying, combining, and network approaches. For reasons of space, we only show runtimes for computations on the TCGA-BRCA data set and selected prior knowledge approaches here. Runtime performances of these prior knowledge approaches on the other data sets are provided in Figures 10.3 to 10.5 in the appendix. This is sufficient, as Figure 6.3 conveys the overall trends we observe for other data sets and other prior knowledge approaches. In Figure 6.3, modifying and combining approaches retrieve prior knowledge from Open Targets, whereas network approaches retrieve prior knowledge from PathwayCommons. Red parts of the bars correspond to runtimes of traditional feature selection approaches, blue parts to runtimes for prior knowledge retrieval, purple parts to runtimes for feature mapping, and grey parts to runtimes of all remaining tasks, e.g. computing joint scores.

In general, most of the prior knowledge approaches shown in Figure 6.3 demonstrate a higher runtime performance than traditional approaches. While we expect a performance gain for prefiltering approaches, we can only observe these for prefiltering prior knowledge approaches using Lasso and SVM-RFE, which are generally computationally intensive. This effect is stronger for data sets with high sample and feature sizes, e.g. TCGA-BRCA, AddNeuroMedII, and REMBRANDT. Therefore, Lasso and SVM-RFE consistently show the highest runtimes amongst the traditional approaches. In contrast, other traditional approaches like RandomForest and ANOVA show runtimes below 60 seconds. Figure 6.2 already showed that runtimes for retrieving prior knowledge are often higher than these 60 seconds. Consequently, prefiltering prior knowledge approaches

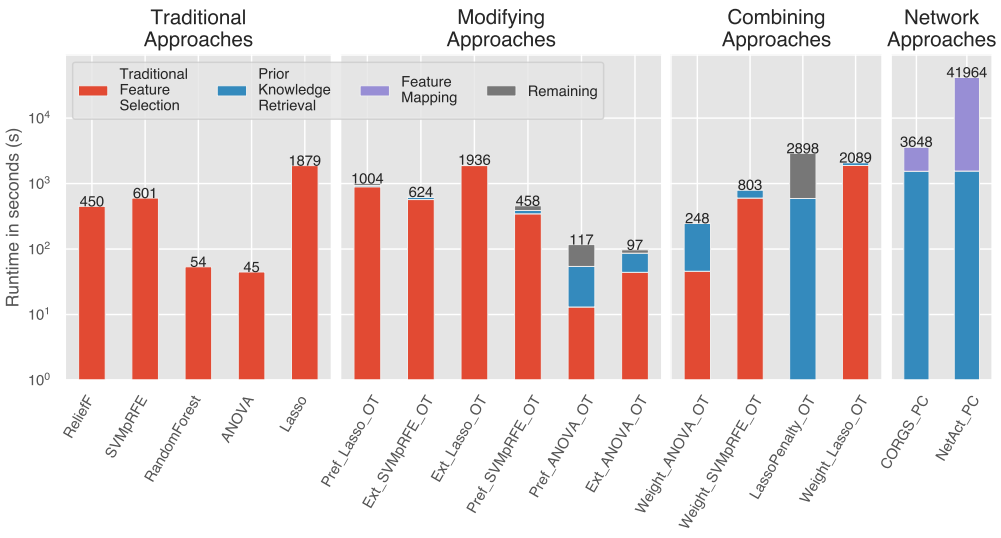


Fig. 6.3: Runtime performances of traditional feature selection and prior knowledge approaches on the TCGA-BRCA data set (logarithmic scale), with modifying and combining approaches using Open Targets and network approaches using PathwayCommons. Numbers on top of the bars correspond to overall runtimes in seconds.

cannot show large performance improvements, even though they reduce runtimes of the subsequently applied traditional approaches to a large extent. Apparently, more complex prior knowledge approaches like LassoPenalty and both network approaches require significantly higher runtimes than all other approaches. LassoPenalty, on the one hand, requires most of the computation time for the actual feature selection[4] and shows increasing runtimes for larger samples sizes up to a couple of days. Network approaches, on the other hand, spend the majority of their runtimes on a) prior knowledge retrieval and b) feature mapping, with the actual feature selection requiring only a fraction of the overall runtime.

Prior Knowledge Approaches are Slower than Traditional Approaches and Require Particularly High Runtimes When Processing Network Information

In summary, runtimes of prior knowledge approaches are often higher than runtimes of traditional approaches, which is primarily a result of the additional time needed for prior knowledge retrieval. While first- and second-level prior knowledge, e.g. genes and gene-disease associations, can be retrieved from knowledge bases within a reasonable time, third-level prior knowledge, e.g. pathways, requires considerably more time than the runtimes of traditional approaches. Furthermore, due to the additional retrieval time for prior knowledge, a prefiltering prior knowledge approach has only improved performance compared to a traditional approach on data sets with high sample and feature

[4] The grey part corresponds to the actual computation of LassoPenalty, which is carried out outside Comprior in R.

dimensionalities. There are, however, still multiple options for improving runtime performances by implementing advanced processing strategies, e.g. caching or parallelization of single queries. With these measures in place, runtimes can be considerably reduced, especially for retrieving first- and second-level prior knowledge. This can promote overall runtimes of prior knowledge approaches to be even closer to those of traditional approaches.

6.3.3 Performances on Data Sets from Different Disease Domains

In this subsection, we examine performances of both traditional feature selection and prior knowledge approaches on data sets of the three disease domains, which addresses the earlier posed question: *Q3: How do prior knowledge approaches compare to traditional approaches on the different disease data sets?* We examine the performance of both traditional and prior knowledge approaches in the context of the three disease domains of our selected data sets, namely Alzheimer's disease, glioma, and breast cancer. We present performance results per disease domain and focus on comparing a) classification performance, b) feature set agreements, and c) the similarity of enrichments. Each of these aspects is further assessed in the context of both an original and a cross-validation data set. While we discuss results from both data sets of a disease domain, we only provide diagrams for one of the two data sets, e.g. TCGA-BRCA, here. Corresponding diagrams for the second data set of the same domain, e.g. SCAN-B, are provided in the appendix (denoted by the A prefix in the figure references). For reasons of clarity, we group the tested feature selection approaches into traditional, modifying, combining, and network approaches. As traditional feature selection approaches, we apply *ANOVA*, *ReliefF*, *Lasso*, *SVM-RFE* (with polynomial kernel), and *RandomForest*. As modifying approaches, we apply prefiltering (*Pref*) and extending (*Ext*) prior knowledge approaches that are combined with *ANOVA*, *Lasso*, and *SVM-RFE*. As combining approaches, we apply weighting (*Weight*) prior knowledge approaches, again combined with *ANOVA*, *Lasso*, and *SVM-RFE*, and the prior knowledge approach of Zeng et al., which integrates prior knowledge as a feature-specific penalty score into Lasso (*LassoPenalty*) [261]. As network approaches, we apply our own approach as described in Section 4.4.6 (*NetAct*) and the approach developed by Lee et al. (*CORGS*), which both apply the same feature selection strategy but use different pathway mapping approaches [113]. We combine the described approaches with each of the available knowledge bases, i.e. KEGG, DisGeNET (*DG*), Open Targets (*OT*), and PathwayCommons (*PC*). We do not provide performance results for all tested combinations with knowledge bases, but instead limit the comparison to the best-classifying combination. In this way, all knowledge base combinations not provided in the subsequent diagrams can be considered to perform either equally well as or worse than the chosen approach.

Classification Performances

Figures 6.4 to 6.6 depict MCC classification performances of feature sets of increasing size (1 to 25) selected for the TCGA-BRCA, TCGA-GBM/LGG, and AddNeuroMedI data sets, respectively. Classification performances of feature sets selected on the other data sets are provided in Figures 10.6 to 10.8 in the appendix. In each of the figures, the upper rows show MCC scores of feature sets selected on a particular data set and used for classifying that same data set, whereas the lower rows show MCC scores of the same feature sets when used for classification on an independent cross-validation data set from the same disease domain. We group the tested approaches into separate diagrams for traditional, modifying, combining, and network approaches.

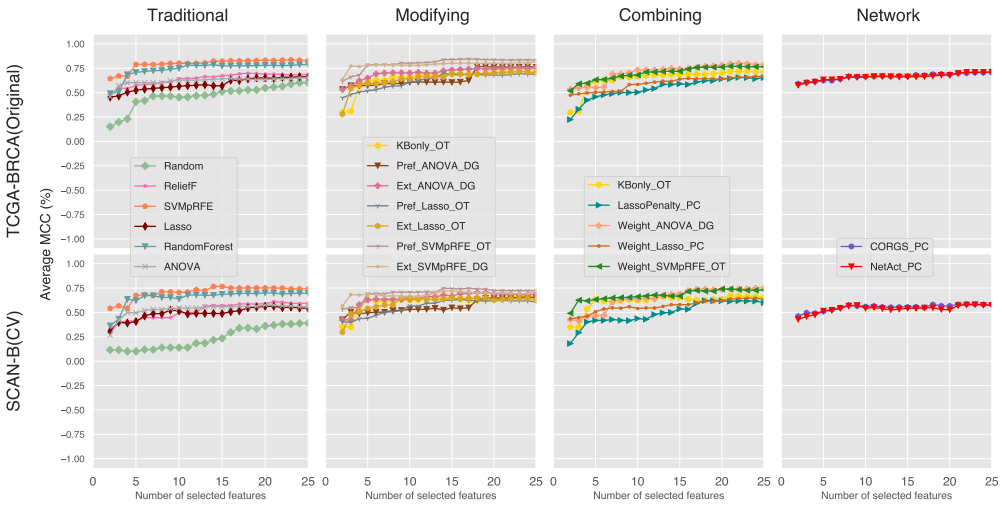


Fig. 6.4: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by the tested approaches on the TCGA-BRCA data set, grouped into traditional, modifying, combining, and network approaches (from left to right). Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the SCAN-B data set for cross-validation (CV). RandomForest and SVM-RFE (*SVMpRFE*) show the highest MCC scores, however most of the tested approaches achieve similarly high MCC scores.

We observe that the classification performances of the tested approaches vary across data sets. While MCC scores of around 0.30 and 0.25 are the lowest for Alzheimer's disease data sets, they are higher for glioma (around 0.75 and 0.5) and breast cancer (around 0.75 and 0.60) data sets. For all data sets we observe that the tested approaches generally perform on a similar level, with the largest variability on the SCAN-B data set (see Figure 10.6 in the appendix). For all but the AddNeuroMedI data set, we recognize that SVM-RFE (*SVMpRFE*) and its prior knowledge approaches, which adapt it in an extending or prefiltering manner (*Ext* and *Pref*), show classification performances that outperform all other approaches. From the modifying approaches, *LassoPenalty*

and prior knowledge approaches that adapt SVM-RFE and Lasso in a weighting manner (*Weight*) often show the worst classification performances, with *Weight_Lasso* approaches performing lowest on breast cancer data sets, *LassoPenalty* performing lowest on both breast cancer and Alzheimer's data sets, *Weight_SVMpRFE* approaches performing lowest on glioma data sets. Network approaches generally perform at the lower end and instead show a constant classification performance in spite of increasing feature set sizes. The baseline approaches *Random* and *KBonly* typically perform worse than the other tested approaches. However, particularly for glioma and breast cancer data sets, both of these baseline approaches exhibit a high and robust classification performance that can outperform some modifying approaches, *LassoPenalty*, and network approaches.

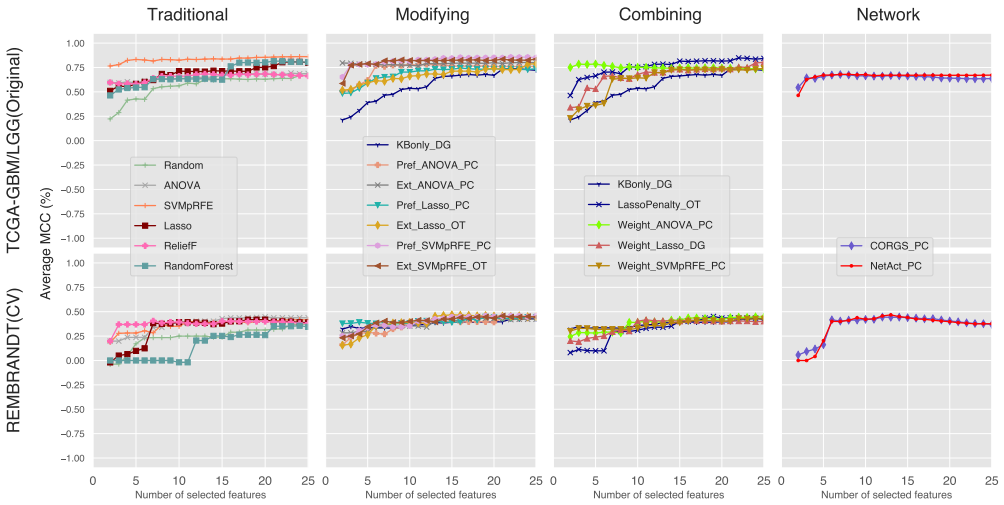


Fig. 6.5: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by traditional, modifying, combining, and network approaches (from left to right) on the TCGA-GBM/LGG data set. Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the REMBRANDT data set for cross-validation (CV). Differences in classification performances observed on the TCGA-GBM/LGG cannot be maintained on the REMBRANDT data set, where all approaches fall back to the same level.

For the cross-validation data sets from Alzheimer's disease and glioma, we observe that the classification performances of both traditional and prior knowledge approaches decline to roughly the same level with narrow ranges. In particular, the performance differences seen on the original data sets — in particular for SVM-RFE (*SVMpRFE*) and its prefiltering (*Pref*) and extending (*Ext*) adaptations — are not recognizable anymore. Only for breast cancer data sets are the performance differences observed on the original data sets robust in terms of remaining visible on the cross-validation data sets. From the modifying approaches, *LassoPenalty* shows the worst classification performances on

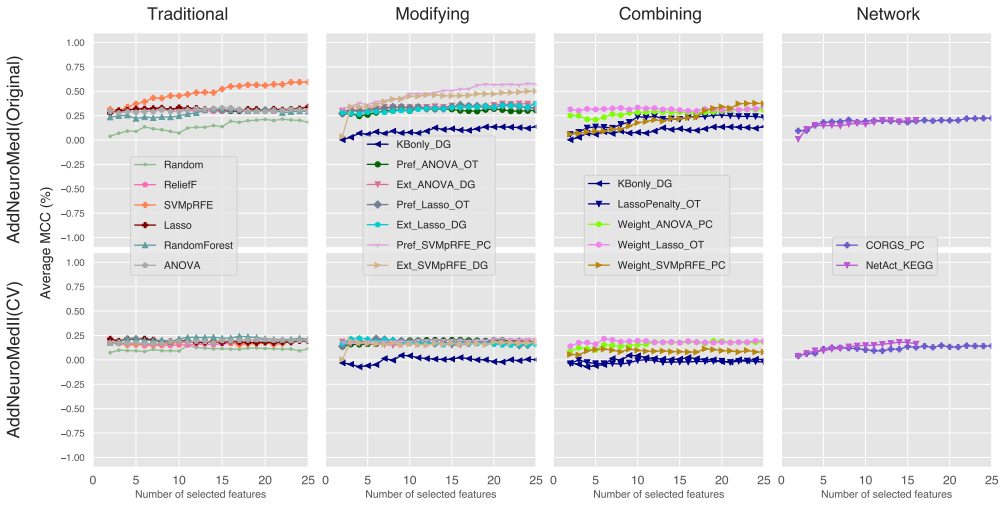


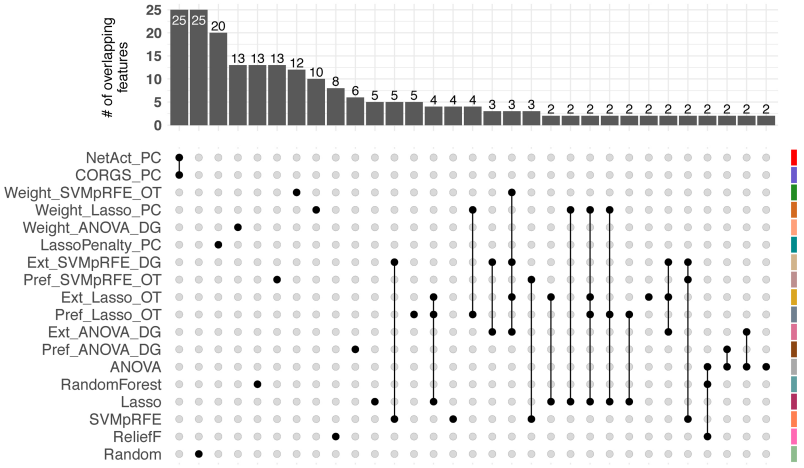
Fig. 6.6: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by traditional, modifying, combining, and network approaches (from left to right) on the AddNeuroMedI data set. Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the AddNeuroMedII data set for cross-validation (CV). Only SVM-RFE (*SVMpRFE*) and its prefiltering (*Pref*) and extending (*Ext*) adaptations reach MCC scores above 0.5 on AddNeuroMedI. However, they fall back onto the same level (MCC below 0.25) as the other approaches on the AdDNeuroMedII data set.

Alzheimer's disease and breast cancer data sets — even falling short of both *Random* and *KBonly* baseline approaches. However, it exhibits a robust classification performance on both glioma data sets. For breast cancer and glioma data sets, feature sets selected exclusively based on the associations retrieved from a knowledge base (*KBonly* baseline approach) show a robust classification performance that is better than randomly selected genes and, in some cases, even outperforms traditional and prior knowledge approaches. In contrast, *KBonly* approaches do not exhibit robust classification performance on Alzheimer's disease data sets, as their MCC scores are lowest on both cross-validation data sets. Randomly selected genes (*Random*) cannot keep up with the performances shown by *KBonly* and only exhibit a robust performance on glioma data sets.

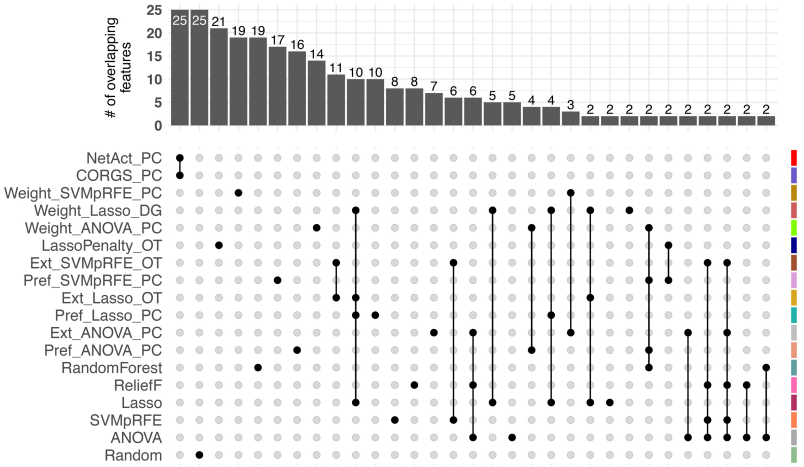
Feature Sets

Figures 6.7 and 10.9 depict overlaps of feature sets between feature sets selected by the tested approaches whose classification performances were already shown in Figures 6.4 to 6.6 and Figures 10.6 to 10.8, respectively.

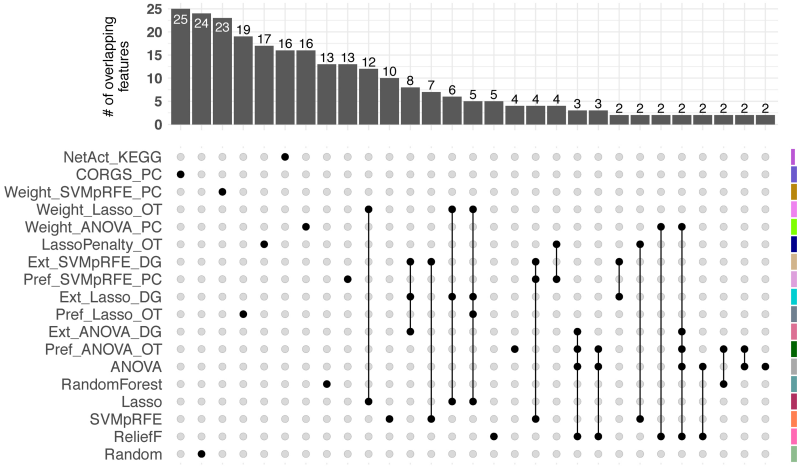
We observe that the majority of feature sets only overlap by individual features, whereas most of the features selected by the tested approaches remain distinct. The highest overlap can still be observed between feature sets of modifying approaches. However, it is worth mentioning that a high proportion of these overlaps exist particularly for



(a)



(b)



(c)

Fig. 6.7: Overlaps of gene sets selected with the same approaches as used for classification, for data sets a) TCGA-BRCA, b) TCGA-GBM/LGG, and c) AddNeuroMedI. We omit overlaps of single features. Often up to half of the features selected by the tested approaches are distinct, with fewest overlaps observed for TCGA-BRCA.

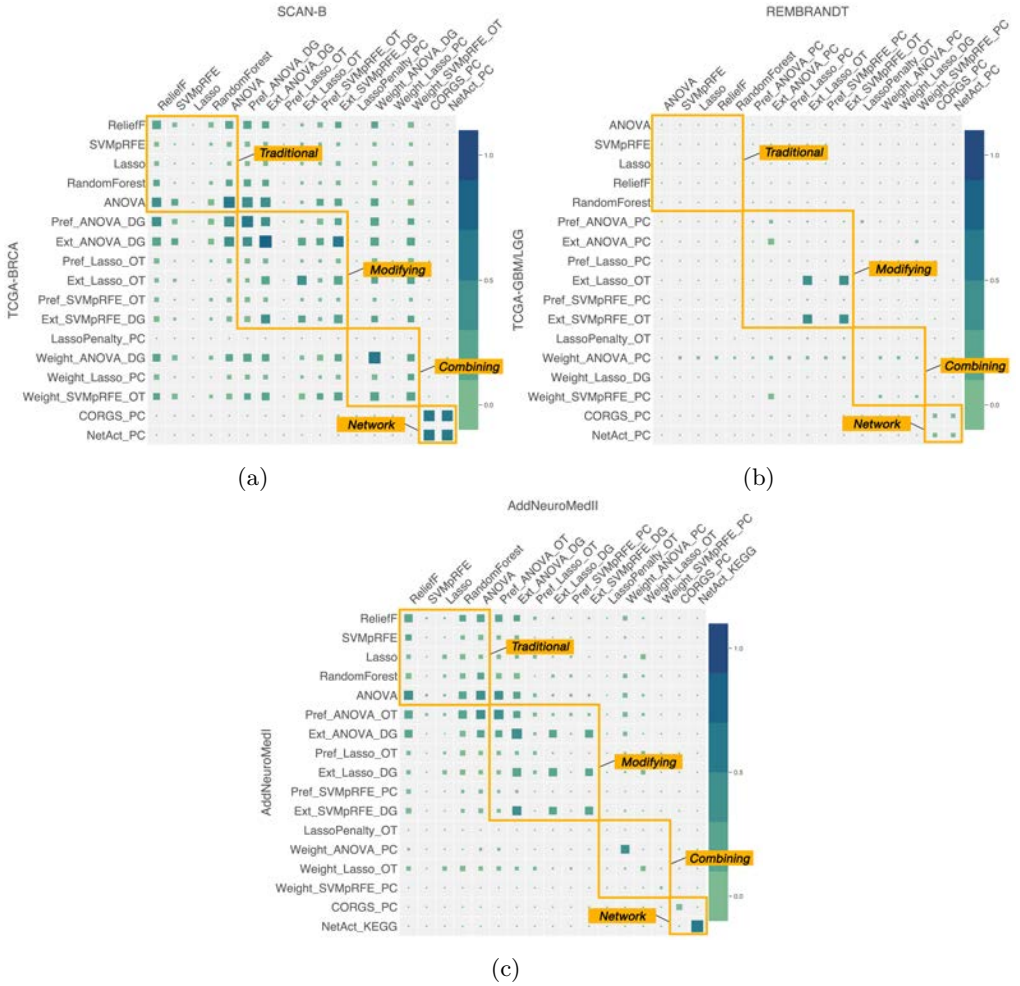


Fig. 6.8: Robustness across data sets of the same disease domain for feature sets selected by the best-classifying approaches from a) TCGA-BRCA, b) TCGA-GBM/LGG, and c) AddNeuroMedI. Robustness is measured by comparing feature rankings via ranked-biased (RBO) overlap. A high RBO score means high agreement of feature rankings across both data sets, and as such indicates a high robustness.

prior knowledge approaches that either combine the same traditional feature selection approach or apply the same knowledge base. This effect is particularly prominent for prior knowledge approaches that extend (*Ext*) the feature sets selected by traditional approaches . The feature sets of these approaches, across all data sets, do not have a single individual feature, but instead share all of their selected features with other approaches. This is expected, as these prior knowledge approaches select half of their feature set exclusively based on prior knowledge and the other half exclusively based on a traditional approach. As such, if the same knowledge base is used, or if the same traditional approach is applied, there is generally a high chance that these feature sets will overlap. Network approaches do not share any features with the other approaches, which is a result of them having pathways instead of genes as features. If network

approaches are further using the same knowledge base, they show complete overlaps due to the same applied selection strategy. In contrast, network approaches using different knowledge bases agree to a far less extent. However, the overlap here rather focusses on feature identity, i.e. compares feature identifiers, and does not attempt to compare pathways based on their member genes. As knowledge bases can use different identifiers for their pathways, it is thus possible that the same pathway listed in both knowledge bases is not recognized in the overlap.

Figures 6.8 and 10.10 depict feature set robustness across data sets by showing rank-biased overlap (RBO) scores of feature rankings selected by approaches on both data sets of the same disease domain, with Alzheimer's disease data sets in the first row, glioma data sets in the second, and breast cancer data sets in the third row. Orange surroundings mark tested approaches of a particular group, i.e. traditional, modifying, combining, and network approaches. A high RBO score indicates that feature rankings are similar and, therefore, serves as measure for feature set robustness across data sets.

Comparing robustness results for all data sets in Figures 6.8 and 10.10, we observe that the tested feature selection approaches show a similar robustness for data sets from the same disease domain — even if they originate from different experiments. Robustness of feature selection approaches, however, differs across disease domains: while we observe a moderate to high robustness for most approaches on breast cancer data sets, robustness decreases for Alzheimer's disease data sets and is worst on glioma data sets. Besides, *LassoPenalty*, SVM-RFE (*SVMpRFE*), and most of the prior knowledge approaches that combine the latter are generally not robust, irrespective of the disease domain. In contrast, approaches like *ANOVA*, *ReliefF*, prior knowledge approaches that extend feature sets from traditional approaches (*Ext*), and network approaches exhibit distinctive robustness which is higher than for other approaches on all data sets and disease domains.

Enrichments

Figures 6.9 to 6.11 depict enrichments retrieved by the tested approaches on the TCGA-BRCA, TCGA-GBM/LGG, and AddNeuroMedI data sets as bubble plots that group semantically similar enrichments into clusters(enrichments for the other data sets are provided in Figures 10.11 to 10.13). The sizes of the bubbles determine with how many enriched terms a tested approach contributes to a cluster. Except for network approaches, where we consider the pathways retrieved as features to be an enrichment, we enriched feature sets selected on Alzheimer's disease and glioma data sets with GO terms and feature sets selected on breast cancer data sets with MSigDB oncogenic signatures. The most general GO term — or gene signature/pathway with the lowest p-value — is selected as cluster representative, respectively.

In general, not many approaches retrieve any enrichments. However we observe that, analogously to feature rankings, enrichments of prior knowledge approaches that use

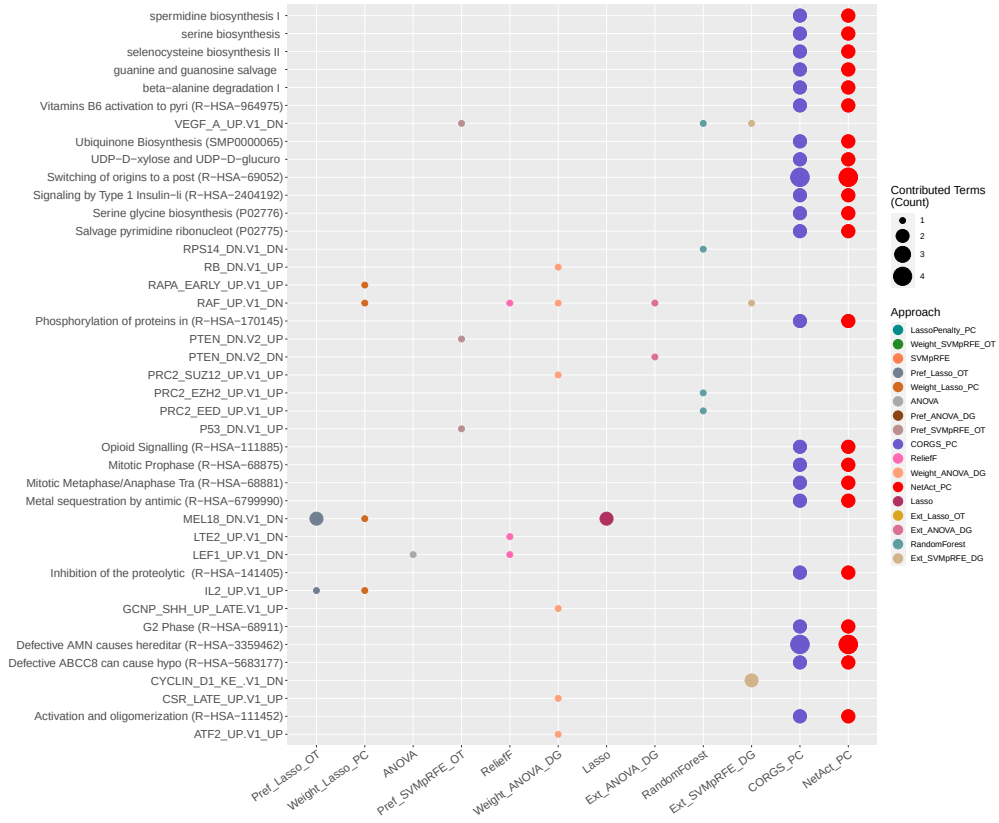


Fig. 6.9: Semantic similarities of enriched MSigDB oncogenic signatures (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the TCGA-BRCA data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. Clusters are nearly identical for network approaches, whereas there is only few similarity between enrichments of the other approaches.

the same knowledge base or adapt the same traditional approach are often grouped into the same semantic category as enrichments of other approaches. Prior knowledge approaches further retrieve many more enrichments in total than traditional approaches, which often only retrieve single enriched terms. For enrichments on Alzheimer's disease data sets, however, we observe more semantically-similar grouped enrichments compared to enrichments of glioma or breast cancer data sets. Traditional approaches, in particular *ANOVA*, *ReliefF*, and *RandomForest*, seem to have retrieved highly similar enrichments. On glioma data sets, we observe such high semantic similarities between enrichments only for *Lasso* and for prior knowledge approaches which incorporate it. On breast cancer data sets, enrichments of *ANOVA*, *ReliefF*, and *RandomForest* again show more semantically similar enrichments, whereas the enrichments of the remaining approaches are most often distinct. For Alzheimer's disease, approaches using DisGeNET as a knowledge base retrieved — outstandingly — many enrichments, whereas this was

the case for approaches using Open Targets on glioma data sets. As we consider pathways with a feature score less than 0.05 as enriched for network approaches, there is no overlap with enrichments of other approaches. As the computation strategy for the feature scores is the same across both *NetAct* and *CORGS*, we observe an exact match in enrichments if both approaches apply the same knowledge base.

Figures 6.12 and 10.14 depict similarity scores of enrichments retrieved for the tested approaches on two independent data sets of a particular disease domain, with Alzheimer's disease data sets in the first row, glioma data sets in the second row, and breast cancer data sets in the third row. A high similarity score retrieved for enrichments of the same approach on two independent data sets indicates a high robustness and, as such, an actual biological relevance of the retrieved enrichments.

We generally observe that the robustness of tested approaches varies across disease domains. As such, robustness is highest on breast cancer data sets, decreases for Alzheimer's disease data sets, and is lowest for glioma data sets. On data sets of the latter disease domain, most of the tested approaches do not retrieve enrichments on both data sets that can be compared. Only few individual approaches, e.g. *Lasso* or *Ext_SVMpRFE_OT*, retrieve enrichments on both glioma data sets. However, most of these enrichments have a low semantic similarity score and are, therefore, not robust. Only prior knowledge approaches using Open Targets as a knowledge base are able to achieve a reasonably high similarity score. In contrast, enrichments retrieved by the tested approaches on breast cancer data sets show a high semantic similarity even across different approaches. From the traditional approaches, enrichments of *ANOVA*, *ReliefF*, and *RandomForest* in particular, show a high semantic similarity and, as such, robustness on data sets from both Alzheimer's disease and breast cancer domains. We also observe a higher robustness particularly for prior knowledge approaches that extend feature sets of a traditional approach (*Ext*). However, this is naturally due to the fact that some of these features are selected independently of the data set at hand. Combining approaches show varying levels of robustness except for *LassoPenalty*, which never retrieves any enrichments that can be compared across two independent data sets. In contrast, network approaches using PathwayCommons achieve high similarity scores for their enrichments, i.e. pathways, on nearly all data sets irrespective of the disease domain. Furthermore, while network approaches do not necessarily select the same pathways across data sets (see also Figure 6.8), they do select semantically similar pathways which might be involved in the same overall processes.

Best Performances on Breast Cancer Data Sets

We consistently observe performance differences of both traditional and prior knowledge approaches on data sets of different disease domains. Data sets of Alzheimer's disease are the hardest to classify while, in contrast, the tested approaches achieve higher classification performances on glioma data sets, but could not identify robust and biologically relevant features. It is only on breast cancer data sets that the tested approaches achieve

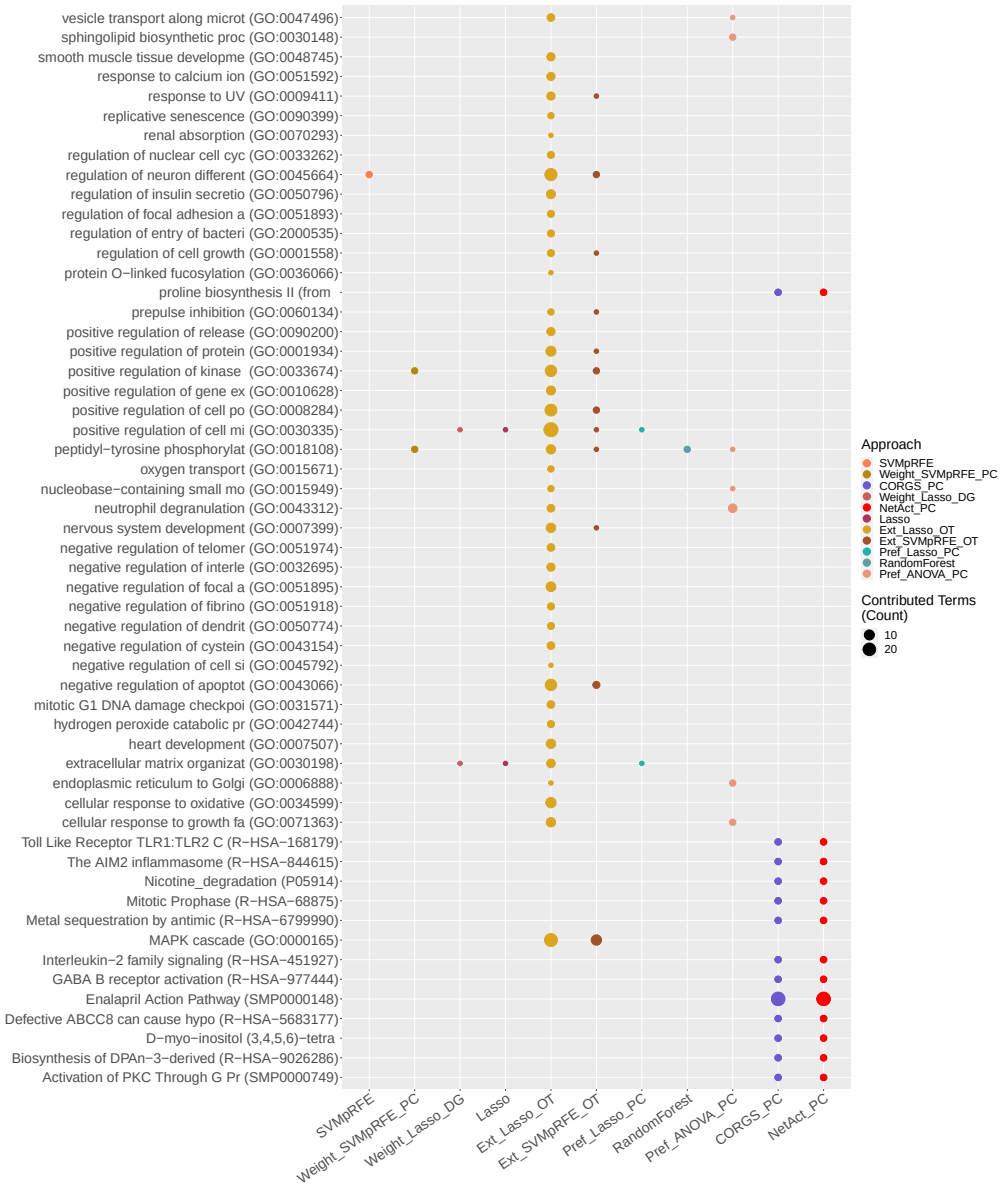


Fig. 6.10: Semantic similarities of enriched GO terms (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the TCGA-GBM/LGG data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. There is a high overlap between *Ext_Lasso_OT* *Ext_SVMpRFE_OT* and both network approaches, respectively.

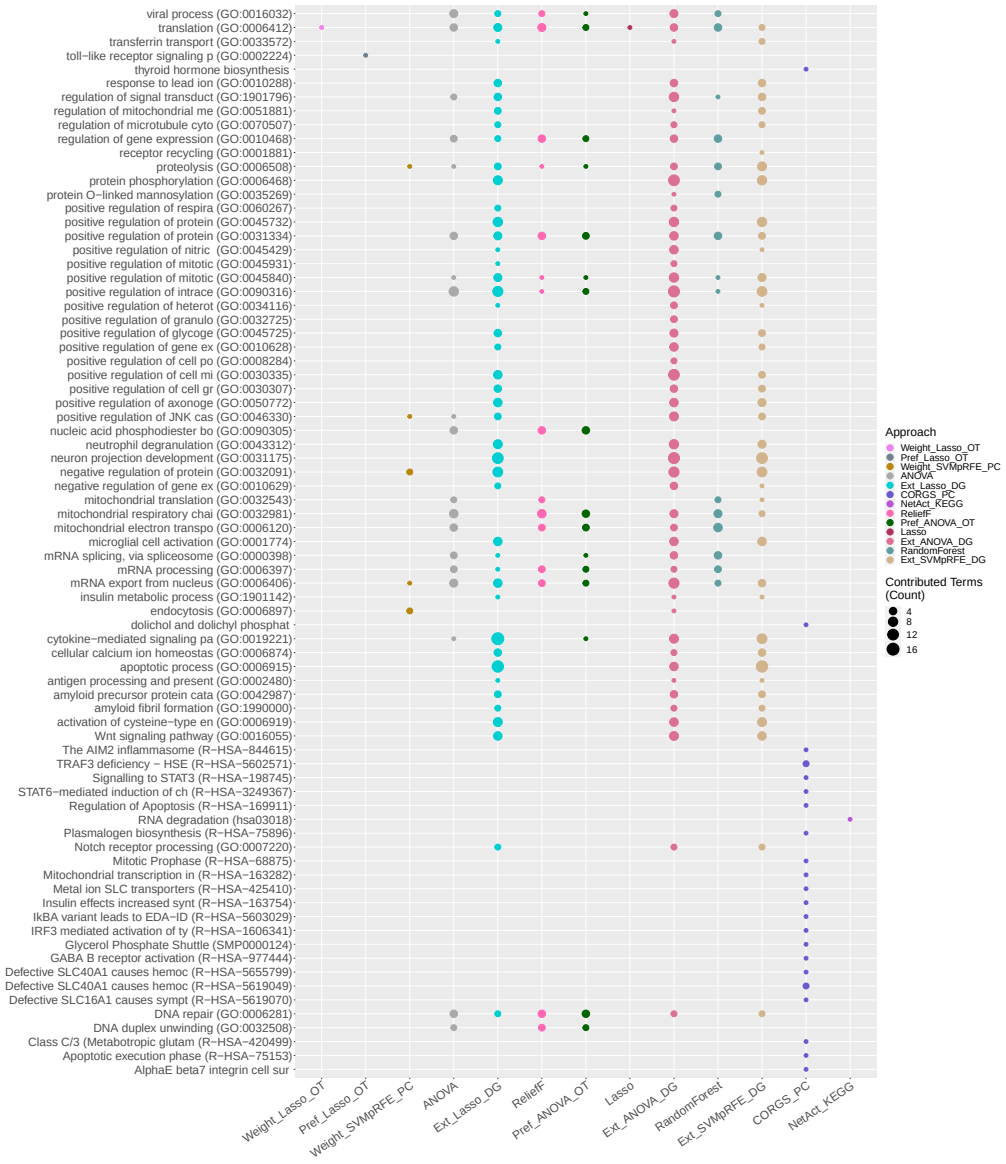


Fig. 6.11: Semantic similarities of of enriched GO terms (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the AddNeuroMedI data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. There is a high overlap between approaches using DisGeNET (*DG*) as knowledge base and between *ANOVA*, *Pref_ANOVA_OT*, *ReliefF*, and *RandomForest*.

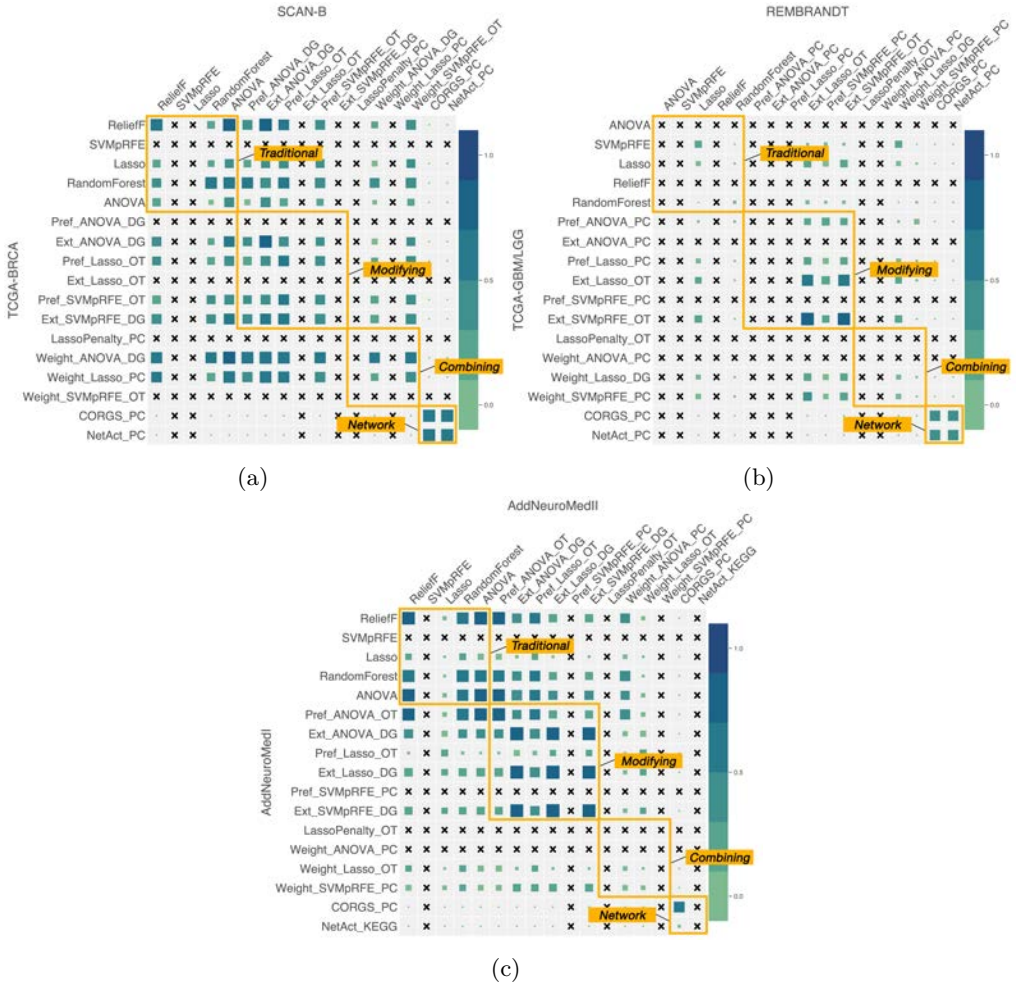


Fig. 6.12: Robustness across data sets of the same disease domain for enrichments (GO terms for Alzheimer's disease and glioma, MSigDB oncogenic signatures for breast cancer, pathways for network approaches) retrieved for feature sets selected by the tested approaches on a) TCGA-BRCA, b) TCGA-GBM/LGG, and c) AddNeuroMedI data sets. High scores mean a high semantic similarity of enrichments and as such indicate a higher robustness and biological relevance of enrichments.

both a high classification performance and robust and biologically relevant features. However, on these data sets, we also observe a high classification power of the baseline approaches *Random* and *KBonly*. Sometimes, they even outperform some of the tested approaches, though randomly selected features seldom show a biological relevance.

Prior knowledge approaches have a similar performance to traditional approaches and do not agree more on feature sets, though they retrieve robust enrichments more often

When comparing classification performances on all data sets, we cannot find large-scale performance improvements in prior knowledge approaches compared to traditional ap-

proaches. Instead, all approaches often perform similarly, with differences in MCC scores below 0.02. Furthermore, we do not observe a higher agreement between feature sets of prior knowledge, unless the same traditional approach is used or the same knowledge base is applied. Indeed, SVM-RFE (*SVMpRFE*) as a traditional approach is the only approach that shows superior classification performance across all data sets and is not outperformed by any other approach. However, while *SVMpRFE* excels in classification, its feature sets do not reveal a biological relevance, whereas prior knowledge approaches that combine *SVMpRFE* with a knowledge base are able to select features that show a biological relevance to a particular extent, at an equal or only slightly lower classification performance level. From the traditional approaches, especially *ANOVA*,*ReliefF*, and *RandomForest* often retrieve enrichments that are robust across data sets of the same disease domain. However, these enrichments are generally retrieved in small numbers, whereas prior knowledge approaches retrieve more enrichments in much higher numbers and still show robustness across data sets.

LassoPenalty Performs Worst, while Network Approaches show Robust Feature Sets and Enrichments

LassoPenalty was implemented by Zheng et al. and is a prior knowledge approach that applies an advanced integration strategy to incorporate prior knowledge. However, *LassoPenalty* is unconvincing in this case study. Across all data sets, it shows only intermediate classification performance and performs particularly badly on cross-validation data sets, where, in parts, its classification performance even falls short of randomly selected features. Furthermore, it does not select features that are robust across data sets, nor does it retrieve any enrichments in most cases. In contrast, network approaches perform much better regarding the robustness of feature sets and enrichments. Irrespective of the disease domain they are applied to, network approaches always perform at the lower end in classification. Network approaches are, however, capable of selecting similar features across two data sets of the same domain which, more importantly, yield enrichments that are typically highly robust.

6.3.4 Comparing Traditional Approaches with Adaptations Therof Using Prior Knowledge

This subsection concentrates on comparing traditional approaches with adapations therof that integrate prior knowledge, e.g. by adapting a traditional ANOVA approach by an additional prefiltering step that uses prior knowledge. This subsection, therefore, addresses the earlier posed question: *Q4: Compared to a high-complexity traditional approach, is it sufficient to use a low-complexity prior knowledge approach?* To answer this question, we compare approaches from the three main traditional feature selection categories — namely *ANOVA* (filter), *Lasso* (embedded), and *SVM-RFE* (wrapper) — with adaptations using prior knowledge that are described in Section 4.4 — namely prefiltering (*Pref*), extending (*Ext*), and weighting (*Weight*) approaches.

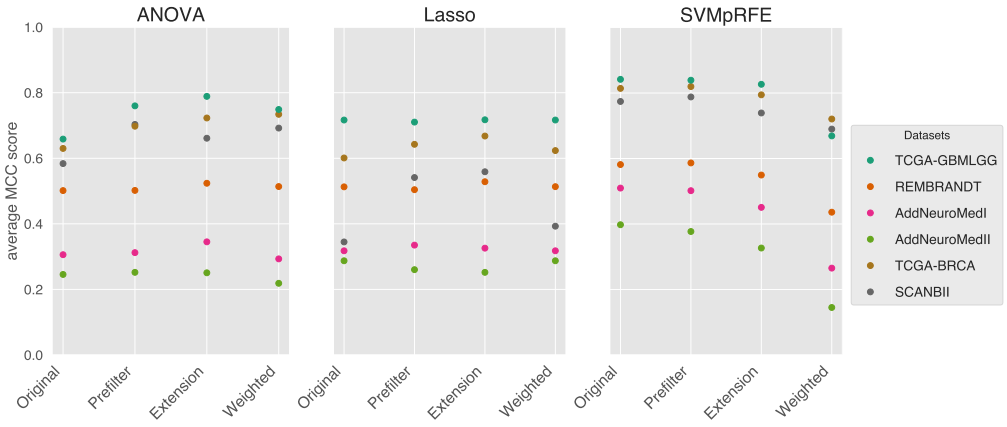


Fig. 6.13: Average classification performances (for 5 to 25 features, measured by average MCC scores) on all six data sets for ANOVA, Lasso, and SVM-RFE and approaches combining them with prior knowledge in a prefiltering (*Pref*), extending (*Ext*), and weighting (*Weight*) manner, respectively. prefiltering and extending adaptations typically lead to a slight performance improvement, whereas weighting adaptations consistently show lowest MCC scores.

Figure 6.13 depicts average classification performances of feature sets (5 to 25 features) on the six data sets for *ANOVA*, *Lasso*, and *SVM-RFE* and prior knowledge approaches that adapt their feature sets with prior knowledge by prefiltering (*Prefr*), extending (*Ext*), or weighting (*Weight*). In the following, we refer to these approaches as *adapted* approaches. While we use each adapted approach in combination with multiple knowledge bases, i.e. DisGeNET, Open Targets, KEGG, and PathwayCommons, here, we only show the average classification performances of the best-performing combination per adapted approach, e.g. *Pref_ANOVA_OT* and *Ext_ANOVA_PC*. At this point we emphasize that we show aggregated and not maximum values of classification performances here. Therefore, the average classification score can encapsulate a) a strong development of classification scores with increasing feature set sizes, i.e. the actual minimum and maximum classification scores are much lower or higher, or b) a more constant classification performance that reaches high scores with few features and a small increase.

From Figure 6.13 we observe that adapted approaches can indeed bring performance improvements, though their intensities vary depending on the data sets and traditional approach used. Adaptations of ANOVA lead to the largest improvements, particularly for prefiltering (*Pref*) and extending (*Ext*) approaches, which show an increase of MCC score up to around 0.15 compared to the original ANOVA approach. Performance improvements are especially high on TCG-BRCA, SCAN-B (breast cancer), and TCGA-GBM/LGG (glioma) data sets, whereas they diminish for AddNeuroMedI and AddNeuroMedII (Alzheimer's disease) data sets. We observe the same pattern, though less distinctively, for Lasso and its adaptations, which can increase classification performance by nearly 0.2 on the SCAN-B data set. In contrast, adaptations of SVM-RFE only show either minor improvements or stronger deteriorations in classification performances. In

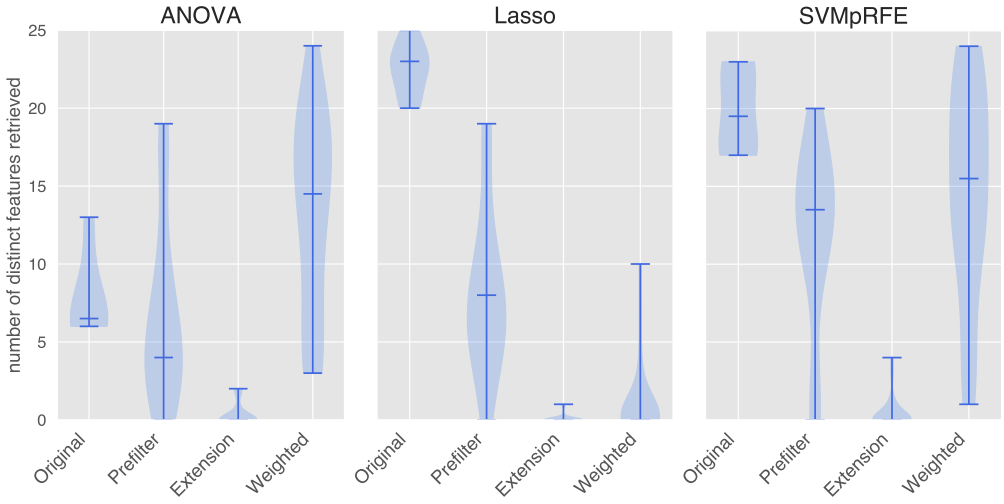


Fig. 6.14: Distributions of distinct features selected exclusively by traditional approaches and their adaptations using prior knowledge, for a) ANOVA, b) Lasso, and c) SVM-RFE. Feature sets are compared to feature sets of all approaches applied in the case study on a respective data set. For Lasso and SVM-RFE (*SVMpRFE*), adaptations using prior knowledge lead to fewer distinct features compared to the original approach.

particular, weighting (*Weight*) adaptations clearly worsen classification performances, e.g. they retrieve MCC scores lowered by at least 0.2 on the Alzheimer's disease data sets. We further observe that improvements of adaptations in classification performance are robust; i.e. we can observe these improvements — albeit at a slightly reduced level — also on the cross-validation data sets as shown in Figure 10.15 in the appendix.

Figure 6.14 depicts how many distinct feature sets traditional approaches and their adaptations typically select when comparing their feature sets with those of all approaches tested in this case study. We observe that traditional approaches generally select more distinct feature sets than their adaptations, with typically more than 7, 20, and 15 distinct features selected for *ANOVA*, *Lasso*, and *SVM-RFE*, respectively. In contrast, their adaptations most often select fewer distinct features, which is particularly pronounced for adaptations of *Lasso* and *SVM-RFE*.

Figure 6.15 depicts distributions of RBO scores of feature rankings selected by the same approach on two data sets of the same domain, i.e. high RBO scores indicate that a feature selection approach produces similar feature rankings on data sets of the same disease domain. Therefore, despite adapted approaches selecting less distinct feature sets, these feature sets are also more robust across data sets. Whereas RBO scores of traditional approaches are located around 0.4, 0.0, and 0.0 for *ANOVA*, *Lasso*, and SVM-RFE (*SVMpRFE*), their adaptations increase these scores by up to 0.4. Again, the improvements are largest for *Lasso* and SVM-RFE (*SVMpRFE*), however feature sets of *ANOVA* and its adaptations generally achieve the highest RBO scores. These effects are consistent across data sets of different disease domains.

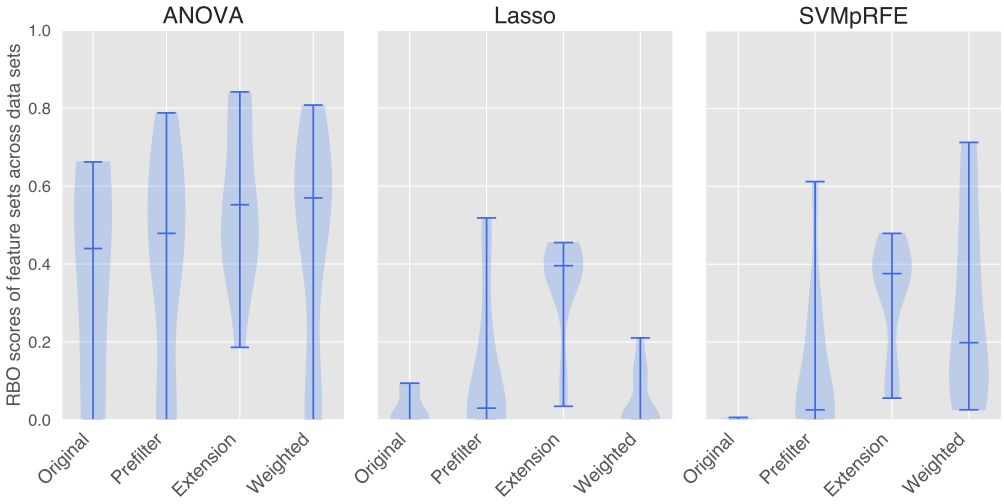


Fig. 6.15: Distributions of feature set RBO scores, i.e. robustness of feature sets across data sets, for a) ANOVA, b) Lasso, c) SVM-RFE (*SVMpRFE*) and their respective adaptations. High RBO scores indicate that a feature selection approach produces similar feature rankings on two data sets of the same disease domain, i.e. produces more robust feature sets. Adaptating a traditional approach increases feature set robustness for ANOVA, Lasso, and SVM-RFE.

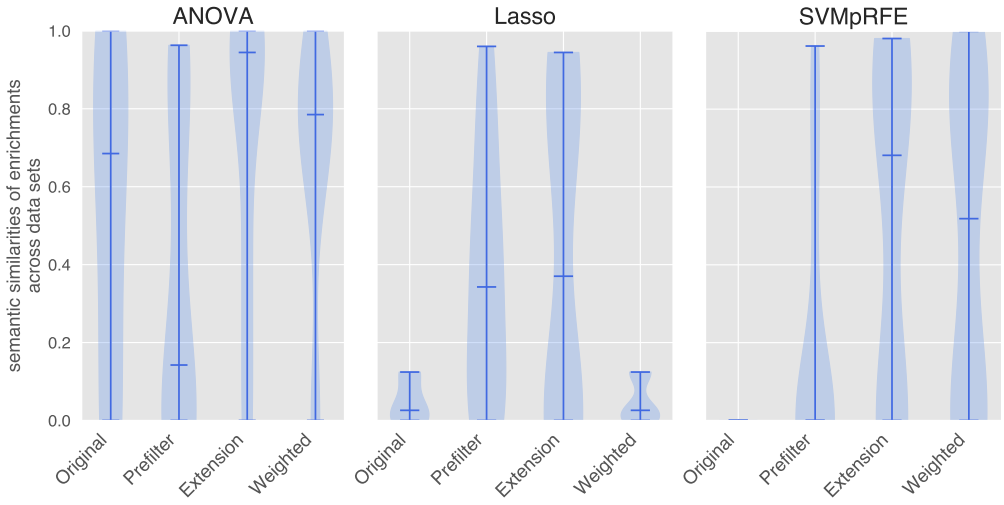


Fig. 6.16: Distributions of similarity scores for enrichments selected by an approach on two data sets from the same disease domain, i.e. enrichment robustness, for a) ANOVA, b) Lasso, and c) SVM-RFE (*SVMpRFE*) and their respective adaptations. High similarity scores indicate a higher robustness and biological relevance of the retrieved enrichments. Adaptations of Lasso and SVM-RFE show a noticeable increase in enrichment robustness.

Figure 6.16 depicts distributions of (semantic) similarities of enrichments retrieved by a tested approach across two data sets from the same disease domain. Therefore, high similarity scores mean an approach selects similar enrichments on data sets from the same domain, which indicates a true biological relevance of the selected features. The results shown in Figure 6.16 comply with what we already observed for feature set robustness: adapted approaches generally improve the robustness of enrichments, with adaptations of *Lasso* and *SVM-RFE* showing the largest improvements. While *Lasso* and *SVM-RFE* never reach a similarity score above 0.2, their adaptations reach a median similarity of up to 0.7. It is important to mention here that approaches that do not retrieve enrichments on both data sets are not included in these plots, and *Lasso* and *SVM-RFE* generally retrieve only individual enriched terms.

ANOVA Profits Most in Classification, Lasso and SVM-RFE Profit More in Feature Set Robustness and Biological Relevance

The observed improvements of adapted approaches vary across the traditional approaches used. From the original feature selection approaches, adaptations of *ANOVA* show the largest improvements in classification performance across data sets, which in parts can even outperform *Lasso* and perform on the same level as *SVM-RFE*. While the original *ANOVA* already selects feature sets that are robust and retrieve enrichments that depict a limited robustness, adaptations of *ANOVA* using prior knowledge still enhance these performances, though to a lesser extent. In contrast, adaptations of *Lasso* and *SVM-RFE*, in particular, lead to improved robustness of feature sets and enrichments. Whereas both *Lasso* and *SVM-RFE* are not capable of retrieving robust feature sets and enrichments, their adaptations, when using prior knowledge, are able to do so and improve the robustness of these by a multiple. However, these improvements at enrichment robustness come to the cost of classification performance, which deteriorates, especially for *Weighted* adaptations of *SVM-RFE*.

6.3.5 Comparing Results of Different Complexity Levels of Integration

In this subsection, we examine the performance differences between prior knowledge approaches that integrate prior knowledge at different levels, addressing the earlier posed question: *Q5: What are the benefits of applying a dense integration of prior knowledge compared to simple filtering strategies?*

In Section 4.3, we group prior knowledge approaches into modifying, combining, and network approaches based on how thoroughly prior knowledge is integrated into the feature selection process. In this subsection, we thus compare prior knowledge approaches from these three categories with each other.

Figure 6.17 depicts average MCC scores of prior knowledge approaches using 5 to 25 features for classification on an original (upper row) and cross-validation data set (lower row), grouped into categories of modifying, combining, and network approaches. From

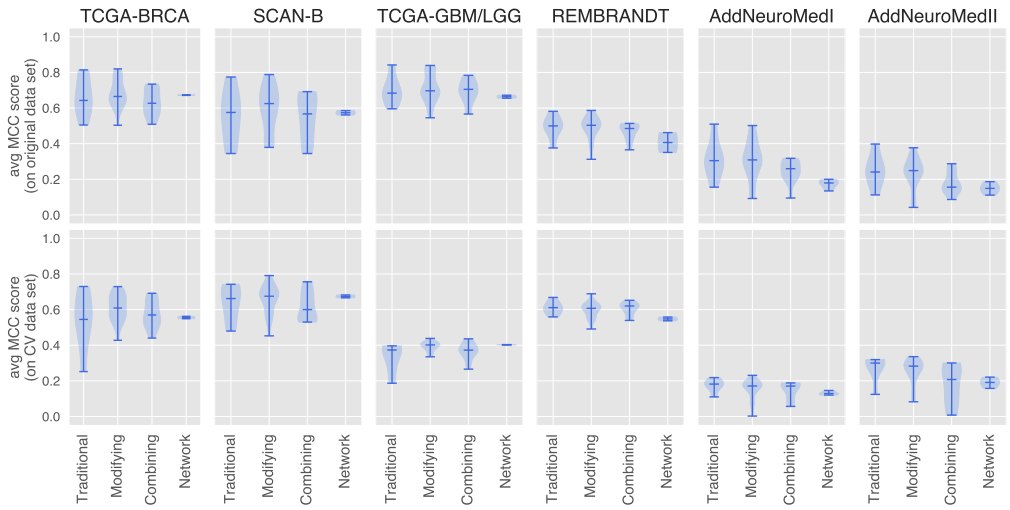


Fig. 6.17: Average classification performances (5 to 25 features, measured in MCC scores) of modifying, combining, and network approaches for all six data sets. The upper row shows the classification performances for feature sets selected on the original data set; the lower row shows the classification performances for the same feature sets on the respective cross-validation data set. While the distributions show different levels of average MCC scores for the respective data set, MCC scores tend to decrease with increasing complexity of integration strategies.

the upper row of Figure 6.17, we observe that, across all data sets, the classification performance of prior knowledge approaches decreases with an increasing integration level: while modifying approaches reach the highest classification performance, combining approaches show a slight decrease, and network approaches exhibit the lowest classification performance. These performance differences are less pronounced on cross-validation data sets (see lower row of Figure 6.17), but still visible for most data sets. Again, however, network approaches perform at the lower boundaries.

Figure 6.18 depicts robustness of feature rankings (upper row) and enrichments (lower row) when applying the same approach to two data sets from the same domain. Feature rankings robustness is measured in RBO scores, whereas enrichment robustness is measured in semantic similarity of enrichments. While network approaches showed the lowest classification performance, they outperform both modifying and combining approaches regarding the robustness of feature sets and, therefore, enrichments. Except for the glioma data sets, network approaches show the highest average RBO scores. However, even for glioma data sets the pathways, while not exactly the same, contain similar gene sets, which indicates that they are functionally similar and involved in the same biological processes. Compared to the other categories, combining approaches show the lowest robustness on both feature sets and enrichments across data sets on all data sets. Compared to network approaches, both distributions of RBO scores and semantic similarities from modifying and combining show a wide dispersion. This can, to a certain extent, be explained by the fact that there are generally fewer data points for network

approaches, as these are only combined with two knowledge bases and no traditional approach.

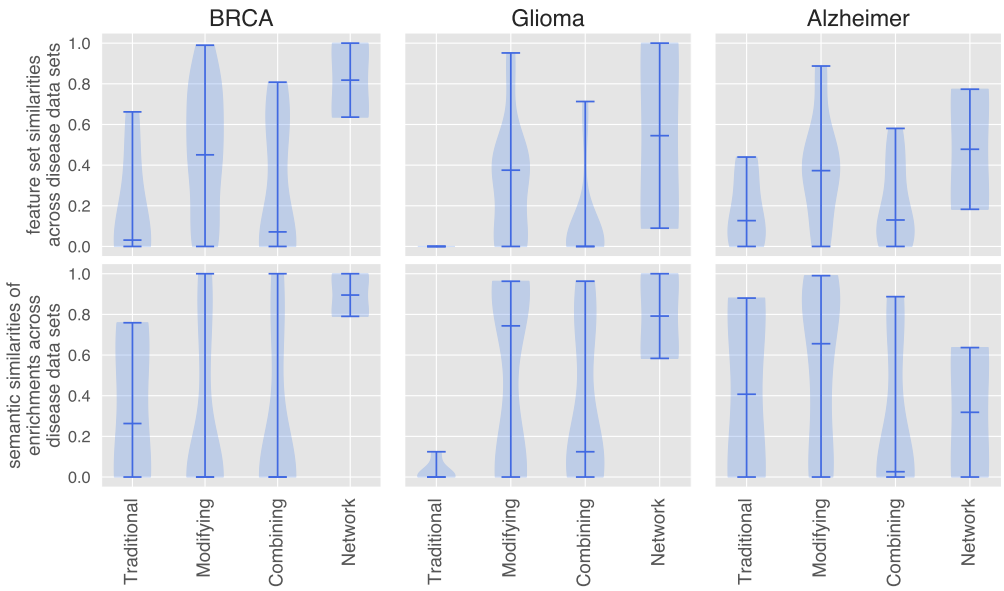


Fig. 6.18: Robustness of feature sets (upper row) and enrichments (lower row) retrieved by modifying, combining, and network approaches. Upper row: RBO scores of feature rankings selected by an approach on a particular data set and used for classification of both data sets from the same disease domain. Lower row: semantic similarities between enrichments retrieved by an approach on two data sets from the same disease domain. Network approaches maintain a high robustness of enrichments even when feature sets show decreased robustness.

Classification Performance Decreases with Increasing Integration Level, However Network Approaches Show Increased Robustness of Feature Sets and Enrichments

We observe the same pattern regarding the classification performance across all data sets: modifying approaches achieve the best classification results, while performance decreases gradually if a more sophisticated integration strategy is applied. The difference in classification performance is highest between modifying and combining approaches, whereas it narrows between combining and network approaches. Therefore, network approaches always perform at the lower end when it comes to classifying samples into their disease subtypes. On the contrary, network approaches show an increased robustness in their enrichments, which clearly outperforms approaches from the other categories, even when the selected features are not robust.

6.3.6 Comparing Results of Different Knowledge Bases

In this subsection, we examine the performance differences between prior knowledge approaches that apply different knowledge bases, addressing the earlier posed question: *Q6: How much does the choice of a knowledge base affect results?* We thus investigate whether the application of a particular knowledge base has a noticeable effect on the performance outcomes. To assess if the choice of knowledge base affects classification performance, we compare MCC scores of all prior knowledge approaches and group them by the utilized knowledge base. We further examine the robustness of feature sets and enrichments. For this, we compare similarities of feature sets and enrichments retrieved by the same prior knowledge approach on two data sets from the same disease domain and group the results by the applied knowledge base. Finally, we assess quantitative aspects of enrichments retrieved when applying a particular knowledge base, e.g. we examine if there are knowledge bases with which prior knowledge approaches generally retrieve more enrichments. We summarize our main findings at the end of this subsection.

Classification Performances

Figure 6.19 depicts win/loss plots for all tested prior knowledge approaches — grouped into modifying, combining, and network approaches — using a particular knowledge base. One plot corresponds to the classification performance for a particular data set. Blue bars indicate the combination with that knowledge base achieves the highest classification performance. Red bars indicate worse classification performance and show how much it differs from the best-performing combination. The worst classification performances are further annotated with the difference in MCC score to the best-performing combination.

As already observed in the previous sections, classification performances of prior knowledge approaches are often on a similar level. While the differences in MCC scores vary up to 0.23, most of the approaches using a particular knowledge base typically show a difference of less than 0.10. Classification performances of approaches using Open Targets, DisGeNET, and PathwayCommons vary across disease data sets. On the breast cancer data sets, combinations using DisGeNET and Open Targets show the best classification performance in 70 and 80 percent of prior knowledge approaches, respectively. In addition, their classification performance often only shows minor differences (indicated by flat red bars). On both glioma data sets, approaches using PathwayCommons consistently outperform combinations with other knowledge bases. However, the classification performances of approaches using Open Targets are on a similar level. In the Alzheimer's data sets — where all approaches achieve only low MCC scores — classification performances vary with no clear winning knowledge base combination. Besides the varying classification performances of approaches using a particular knowledge base, Figure 6.19 demonstrates that approaches using KEGG consistently have the worst classification performance across all data sets. Furthermore, while approaches using the other knowledge bases often perform similarly, the difference to approaches using KEGG is typically

larger, often lowering MCC scores by around 0.10 points. This effect is most extreme for network approaches, where the MCC scores on glioma and breast cancer data sets are lowered by 0.35 to 0.67. These differences are caused by KEGG failing to provide any relevant pathway that can be used as a new feature, and thus failing to classify the data sets at all. However, even if KEGG provides enough pathways as features for classification, e.g. on the Alzheimer's data sets, these combinations are still outperformed by network approaches using PathwayCommons.

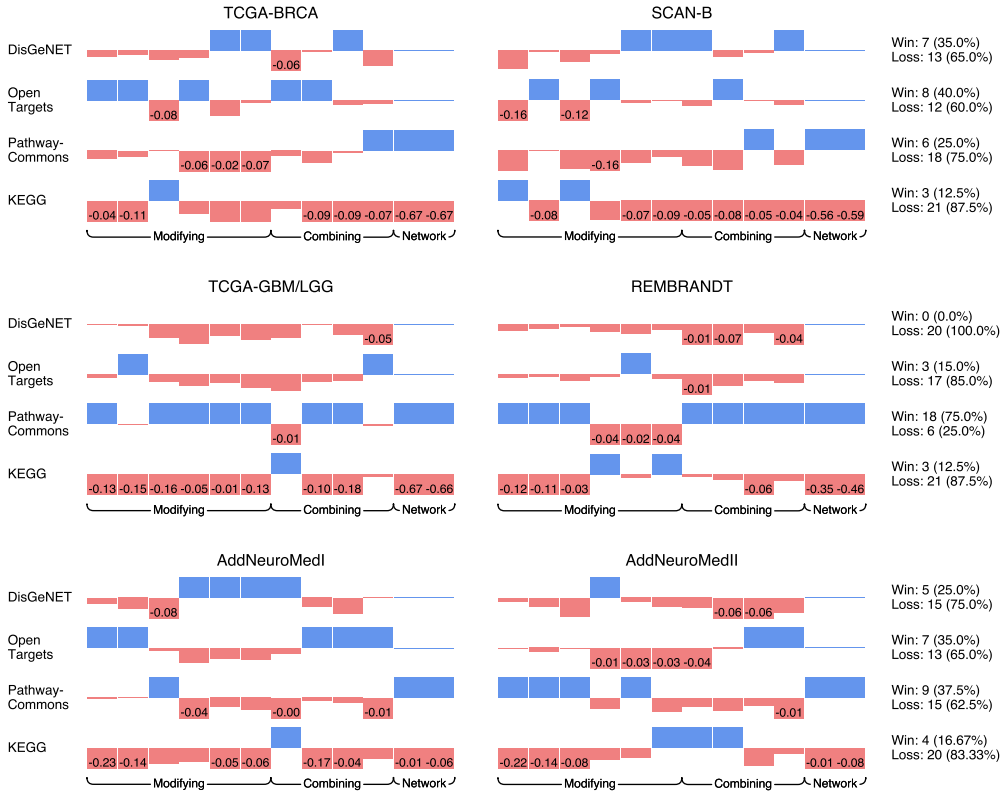


Fig. 6.19: Win/loss plots (one plot per data set) of the different prior knowledge approaches grouped by knowledge base applied. Blue bars indicate the best classification performance, height of red bars indicates distance to the best-performing approaches. Numbers on red bars depict the highest difference in MCC score to the best-performing approach. From the knowledge bases applied, KEGG consistently shows lowest classification performance across all data sets.

Robustness of Feature Sets and Enrichments

Figure 6.20 depicts the RBO scores (left side) and semantic similarities (right side) of feature sets and enrichments that are retrieved by prior knowledge approaches using a particular knowledge base on two data sets from the same disease domain. High RBO scores indicate that prior knowledge approaches select highly similar feature rankings

on both data sets, while high semantic similarities of enrichments across two data sets indicate that these are functionally related and, as such, biologically relevant. While Figure 6.20 aggregates RBO scores and semantic similarities from feature sets and enrichments retrieved from all six data sets, it still conveys the trends we recognize on the individual data sets. We refer the interested reader to Figures 10.16 to 10.18 in the appendix, which provide RBO scores and semantic similarities for the individual data sets.

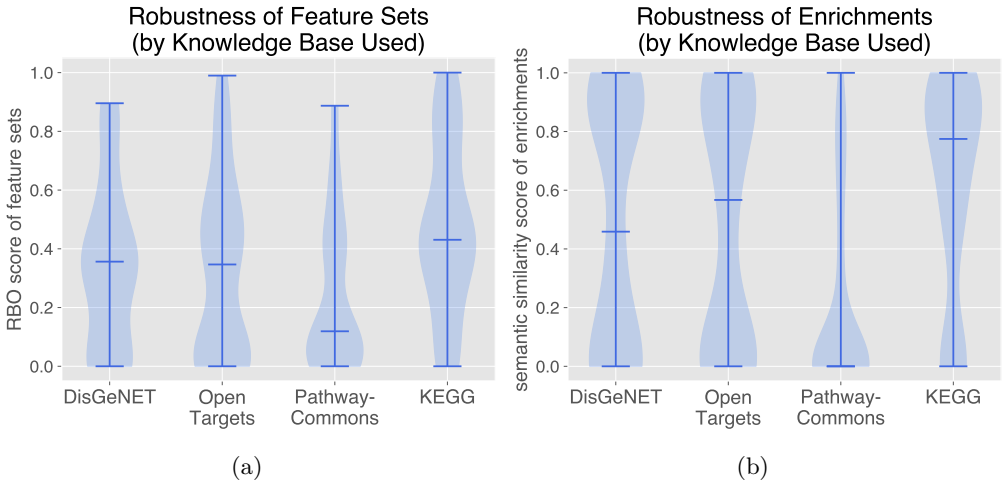


Fig. 6.20: Robustness of a) feature sets and b) enrichments across data sets, grouped by knowledge base. Feature set robustness (left side) is measured via RBO scores, enrichment robustness (right side) is measured in semantic similarity. Prior knowledge approaches using PathwayCommons generally retrieve less robust feature sets and enrichments, whereas approaches using KEGG retrieve most robust feature sets and enrichments.

From Figure 6.20 we observe that prior knowledge approaches using DisGeNET and Open Targets demonstrate a similar performance of RBO scores. Both knowledge bases enable maximum RBO scores between 0.9 and 1.0, though achieve median RBO scores of around only 0.35. In contrast, approaches applying PathwayCommons as a knowledge base perform worse, with a median RBO score of around 0.1. The maximum RBO scores of around 0.9 most likely originate from network approaches, as we already observed in Section 6.3.5 that these show high robustness. From the examined knowledge bases, KEGG shows the highest median RBO scores for nearly all data sets. The differences are more pronounced on the glioma and breast cancer data sets, while they converge for the Alzheimer's data sets (see also Figures 10.16 to 10.18 in the appendix).

The behavior of knowledge bases observed for RBO scores is even more pronounced in the semantic similarity scores of the retrieved enrichments. According to Figure 6.20, DisGeNET and Open Targets again show a similar performance with median semantic similarity scores between 0.45 and 0.55. It is worth mentioning, however, that both

DisGeNET and Open Targets show a particularly low robustness of enrichments for both breast cancer data sets, where the median similarity scores fall to 0.0 (see also Figure 10.16). Prior knowledge approaches using PathwayCommons again show the lowest similarity scores of enrichments across data sets, with a median similarity score of 0.0. The few higher semantic similarity scores most likely originate from network approaches, which typically achieve high robustness of their enrichments. As with RBO scores, prior knowledge approaches using KEGG again show outstanding performance with a median semantic similarity score close to 0.8. This effect is strongest for the glioma data sets, where median semantic similarities of enrichments even rise up to 0.9 (see also Figure 10.17).

Enrichments

Figure 6.21 depicts quantitative aspects of enrichments retrieved by prior knowledge approaches using a particular knowledge base. Figure 6.21 a) depicts overall numbers of enrichments retrieved by prior knowledge approaches using a particular knowledge base, whereas Figure 6.21 b) shows how many genes from the data sets are involved on average in an enrichment. Numbers are again aggregated from all six data sets, however, the interested reader is referred to the individual data set plots in Figures 10.19 to 10.21.

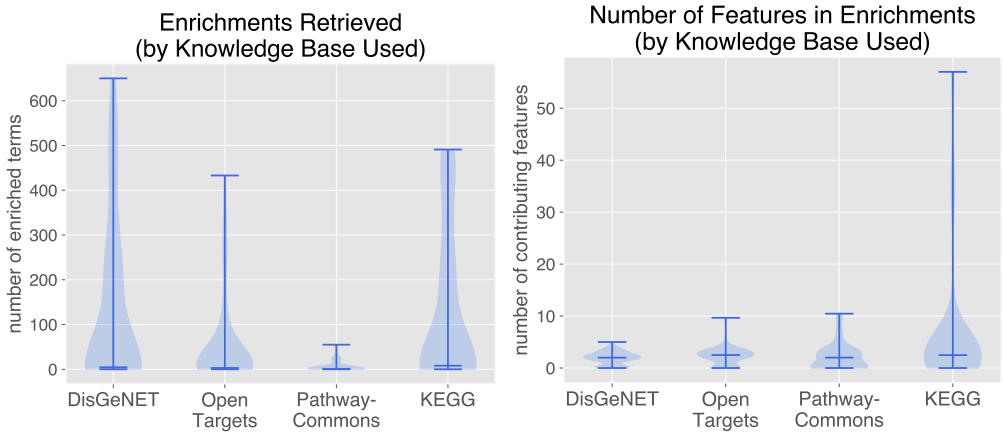


Fig. 6.21: Quantitative assessment of enrichments retrieved by prior knowledge approaches using a particular knowledge base on all six data sets, with a) showing the average number of enrichments retrieved and b) showing the average number of features being involved in an enrichment. Prior knowledge approaches using DisGeNET often retrieve many enrichments, however with only few features being actually involved, whereas approaches using KEGG still retrieve many enrichments, however with more features involved.

From Figure 6.21 we again see knowledge base specific differences in the quantitative aspects of the enrichments retrieved. The median number of enrichments has nearly the

same level for all knowledge bases, i.e. many of the approaches — independent of the knowledge base applied — retrieve no or few enrichments. If enrichments are retrieved, prior knowledge approaches using DisGeNET and KEGG generally retrieve more enrichments than the other knowledge bases, with some prior knowledge approaches yielding more than 500 enrichments. However, on closer examination of the number of enrichments retrieved for a particular data set, it becomes apparent that the median values of Figure 6.21 are caused by only few, i.e. typically less than five, enrichments retrieved for the breast cancer data sets (see also Figure 10.19). These were enriched with MSigDB oncogenic signatures, of which only 189 exist in total[5], and as such generally only few signatures are enriched. Looking at the other data sets, we see that the median number of enrichments reaches multiple hundreds on the glioma and Alzheimer's data sets, particularly for approaches using DisGeNET and KEGG (see also Figures 10.20 to 10.21 in the appendix). Out of all the knowledge bases applied, prior knowledge approaches using PathwayCommons consistently retrieve the lowest amount of enrichments, typically less than five and not more than 100 enrichments at maximum.

Besides the amount of enrichments retrieved, we also measure how many features are involved in an enrichment. From Figure 6.21 b) we see that the majority of enrichments — across all knowledge bases — generally have less than ten genes involved. All approaches show a median number of around three genes being involved in an enrichment and mainly differ by their extreme values. Here, DisGeNET, Open Targets, and PathwayCommons show the lowest upper boundaries of around 10 genes being involved in an enrichment. KEGG, on the other hand, shows extreme values of more than 50 genes being involved in an enrichment.

PathwayCommons Increases Classification Performance at the Cost of Enrichment Robustness

The preceding observations show that prior knowledge approaches using PathwayCommons achieve high classification performances, especially for the glioma data sets. At the same time, feature sets are less robust across data sets and retrieve fewer enrichments which are also robust. This indicates that the selected feature sets, while having higher distinctive ability, do not truly capture the relevant underlying biological processes. The described behavior, however, does not fully apply to network approaches. In these cases, we observe that PathwayCommons is able to retrieve multiple pathways that are significantly altered between disease subtypes and, as such, are suitable features that further exhibit a robustness across data sets of the same disease.

KEGG Retrieves more Enrichments that are Robust

While we observe increased classification performance at the cost of feature set and enrichment robustness for PathwayCommons, we notice the opposite behavior for KEGG:

[5] `https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp`, as of March 17, 2022

prior knowledge approaches using KEGG consistently show the worst classification performance. At the same time, they exhibit the highest robustness of feature sets and enrichments across data sets. Even if feature sets are not robust across data sets, they lead to enrichments that are still semantically similar. Furthermore, prior knowledge approaches using KEGG often retrieve more enriched terms than approaches using the other knowledge bases, independent from the source used for enrichments. Finally, it is important to mention that KEGG is hardly applicable to network approaches. From the tested data sets, KEGG retrieves multiple relevant pathways as features only for Alzheimer's, while it only retrieves single pathways for the other disease domains.

Open Targets and DisGeNET are a Good Compromise

Prior knowledge approaches using Open Targets and DisGeNET show intermediate performances regarding both the discriminative ability and robustness of feature sets and their enrichments. Therefore, DisGeNET and Open Targets constitute a compromise compared to KEGG and PathwayCommons, both of which excel in either classification performance or biological relevance of feature sets. However, Open Targets and DisGeNET are not immediately applicable to network approaches, as these require third-level prior knowledge, i.e. networks. Both Open Targets and DisGeNET naturally provide second-level prior knowledge, i.e. gene-disease associations, which requires considerable efforts to be transferred to third-level prior knowledge.

6.3.7 Threats to Validity of the Findings from the Case Study

The design of our conducted case study has some characteristics that might limit the validity of the conclusions drawn from our experiment results.

The choice of measures for preprocessing gene expression data sets can have a crucial impact on analysis results [51]. This is valid especially for normalization strategies, of which many approaches are available for use [1]. Our applied data sets all showed different stages of preprocessing. For example, some of our data sets were available as raw count data for which we chose the most suitable preprocessing measures based on latest insights. Others were available as normalized data to which an outdated normalization strategy had been applied, e.g. MAS5 on the REMBRANDT data set. We thus cannot rule out that the given preprocessing measures might affect our experiment results. This concern is softened, however, as we aimed to identify global trends, i.e. we drew our conclusions from observations that were visible across all data sets.

Our case study further faced the same issue as related work regarding the accessibility of prior knowledge approaches and only compares a limited number of prior knowledge approaches of other authors. Except for the approach developed and implemented by Zeng et al. (*LassoPenalty*), we had no access to the source code of any other prior knowledge approach. In order to keep the implementation effort within the scope of this thesis, we limited our case study to include prior knowledge approaches for which a) the

implementation effort is reasonable and b) that could be easily adapted to our applied classification use case. We particularly reimplemented the CORGS network approach by Lee et al. based on their descriptions in the corresponding manuscript to incorporate an external network approach in our case study [113]. While our implementation underwent critical testing, we cannot categorically preclude that it missed important aspects of Lee et al.'s method, which might have led to biased results. We further cannot entirely rule out that other sophisticated prior knowledge approaches, e.g. those described by Guo et al. or Swarnkar et al., would have delivered results that strongly differ from our insights [76, 131, 218].

Our applied knoweldge base and search terms also might have retrieved too generalized prior knowledge that cannot add much to the analysis. We incorporated knowledge bases that have a broad coverage of many diseases and genes. This was an economic decision, as we wanted to apply multiple knowledge bases to each disease domain. However, more specialized knowledge bases, e.g. like InnateDB, could have provided much more specific prior knowledge [28]. Regarding our choice of search terms for prior knowledge retrieval, we made sure to cover all disease subtypes with their various medical terms by using the metathesaurus [140]. However, we are no experts of the medical domain and might have missed aspects that are characteristic for a disease subtype, which could result in prior knowledge that is still too general.

A further critical aspect of our case study is the biological evaluation of detected biomarkers. We applied gene set enrichment strategies that used information from Gene Ontology and MSigDB to annotate our retrieved biomarkers. As both GO and MSigDB are knowledge bases, we took care that they do not include data from any of the knowledge bases used for feature selection. However, we cannot fully preclude that information from one of the knowledge bases was resembled in the other, e.g. via manual curation from the same scientific publications. As such, there remains a risk that our results, particularly the biological relevance of biomarkers, are positively biased towards prior knowledge approaches. At this moment, there is no solution to this issue and it remains subject for further investigations.

6.4 Summary

In this chapter, we described the setup and results of a case study we conducted with multiple knowledge bases and both traditional and prior knowledge approaches on data sets from three disease domains: Alzheimer's disease, glioma, and breast cancer. We first examined how much prior information is provided by the applied knowledge bases and conclude that both Open Targets and PathwayCommons provide the highest coverage for the tested disease domains. DisGeNET still returns a lot of relevant genes, though it fails to provide satisfactory evidences. KEGG, in contrast, generally provides only limited prior knowledge, which turns out to be especially challenging for the tested network approaches.

We further compared runtimes of traditional and prior knowledge approaches to assess the feasibility of prior knowledge approaches. We found that, due to the additional time needed to retrieve prior knowledge, traditional approaches require less computation time than prior knowledge approaches. They only supersede traditional approaches when applied in a prefiltering manner on data sets with large dimensions and require a substantial amount of runtime when network information is retrieved and processed.

When comparing the tested approaches on data sets from the different disease domains, they generally achieved the highest performances on breast cancer data sets. Regarding the overall classification performances, all tested approaches perform on similar levels, often with only minor differences. Prior knowledge approaches, however, often outperform traditional approaches regarding the robustness of their feature sets and both the quantity and robustness of enrichments retrieved. Network approaches, in particular, distinguished themselves by selecting and retrieving particularly robust feature sets and enrichments. However, *LassoPenalty* — as a prior knowledge approach applying sophisticated integration strategies — consistently showed the worst performance in all aspects under consideration, i.e. distinctive ability, feature set robustness, and enrichments.

When comparing traditional approaches directly to prior knowledge approaches that combine them with prior knowledge, e.g. in a prefiltering manner, we observed that *ANOVA* showed the largest improvements in classification, which increased up to the levels of the generally best-classifying *SVM-RFE*. In contrast, *Lasso* and *SVM-RFE* benefit from prior knowledge integration particularly in terms of feature set robustness and enrichments.

When comparing the effectiveness of prior knowledge approaches that apply different integration strategies for prior knowledge, we observed a general decrease in classification performance with more advanced integration strategies.

Finally, we assessed whether the choice of knowledge base had an impact on the effectiveness of prior knowledge approaches. It turns out that there seems to be a drawback between classification performance and feature set and enrichment robustness: while prior knowledge approaches using PathwayCommons showed an increased classification performance, they did not select robust feature sets and did not retrieve many robust enrichments. In contrast, KEGG consistently showed the worst classification performance, yet it retrieved many enrichments that proved to be robust across data sets. Open Targets and DisGeNET turned out to constitute a compromise between the aforementioned two, as they often showed reasonable classification performance and still retrieved robust enrichments.

7

Discussion

This chapter consolidates and discusses findings from our work presented in the previous chapters. Coming from our two research questions on the applicability and effectivity of prior knowledge approaches, we clarify how our contributions — namely our presented formal concepts, their technical realization in Comprior, and the conducted case study — address these and also elaborate on potential limitations. We conclude this chapter by highlighting the challenges for prior knowledge approaches which remain open.

7.1 Improving the Applicability of Prior Knowledge Approaches

Prior knowledge approaches are not widely applied in practice, although biological databases experience an ever increasing growth in information content and research assumes integrative approaches to address current issues of traditional biomarker detection approaches [19, 151]. Still, studies for which only single omics data is available only incorporate biological information at the very end to biologically interpret the detected biomarkers.

The observations above led us to our first research question, namely:

RQ1: *How can we improve the applicability of prior knowledge approaches in practice and subsequently enable better comparability?*

After conducting a qualitative comparison of existing prior knowledge approaches and their key characteristics, we found that their actual application in practice is negatively affected by their inflexibility and considerable efforts for prior knowledge curation. Most approaches are custom-tailored to a particular use case with particular requirements to the format of applied prior knowledge. Furthermore, prior knowledge retrieval typically involves considerable curation efforts, and applying a prior knowledge approach to a different domain requires implementing a de novo prior knowledge retrieval.

7.1.1 Generalized Approaches and Unified Definitions for Prior Knowledge

Our fundamental concept to improve the technical applicability of prior knowledge approaches is to retrieve prior knowledge from online knowledge bases. This allows for streamlining the retrieval process and agreeing on a uniform definition for prior knowledge so that resources can be used interchangeably.

In Chapter 4, we defined concepts for both prior knowledge and prior knowledge approaches. Intending to facilitate a streamlined prior knowledge retrieval, we first reviewed existing online knowledge bases regarding their available biological information and how it is accessible. We subsequently derived a formal definition of prior knowledge as it is available in online knowledge bases. Prior knowledge is thus available in three levels: a list of biological entities, a list of (scored) biological entities, and a list of networks. We further described how prior knowledge can be transformed from one level to the other and discussed potential constraints. Having a clear definition of prior knowledge and how to transform it into different levels increases flexibility in application for prior knowledge approaches. We further reviewed existing prior knowledge approaches regarding their key characteristics and identified commonalities that allowed us to classify them into the groups of modifying, combining, and network approaches. We explained what integration strategy is generally applied by approaches of each category and what level of prior knowledge is used. For each category of prior knowledge approaches, we further provided formal descriptions of own prior knowledge approaches that are generally applicable with the highest flexibility, e.g. by combining any traditional approach with prior knowledge of a particular type, which is not restricted to a particular use case. Some of our concepts can be considered to be generalizations of approaches found in related work. For example, Fang et al. and Jungjit et al. can be seen as specialized implementations of our prefiltering approach, while RelSim and SoFoCles are specific forms of our extension approach [58, 98, 132, 148]. Jungjit et al. further present a specialized implementation of our weighting approach [98]. However, our generalized concepts address the aforementioned applicability issues by allowing the highest flexibility regarding the use of prior knowledge and combination with other approaches.

7.1.2 A Technical Infrastructure for Development and Evaluation

We implemented our presented concepts for both prior knowledge and prior knowledge approaches into a research framework called *Comprior*. Comprior is an evaluation platform for prior knowledge approaches, i.e. it supports the complete analysis from feature selection to classification and enrichment, and provides access to multiple knowledge bases and multiple feature selection approaches — both traditional and prior knowledge approaches. This is our contribution to enable an effective development and benchmarking of prior knowledge approaches, which ultimately addresses our first research question. Comprior improves applicability and subsequently enables comparability of prior knowledge approaches in mainly two ways.

First, Comprior makes prior knowledge approaches accessible. In our qualitative review of existing approaches, only a few approaches actually had source code or executables available for external use. Comprior is meant to be used and further extended by the research community. Researchers can use Comprior's technical infrastructure to straightforwardly implement novel approaches and benchmark them against both traditional and prior knowledge approaches. Thus realizing novel prior knowledge approaches with Comprior enhances accessibility and overall reusability of prior knowledge approaches.

Second, Comprior streamlines the retrieval process of prior knowledge. Although we have derived a unified definition for the different levels of prior knowledge, the retrieval process for each individual knowledge base can still be cumbersome and involve considerable implementation efforts. Comprior decouples the actual prior knowledge retrieval from feature selection and instead implements our formal definitions of prior knowledge as interfaces that can be used by any feature selection approach. Still, Comprior currently supports a limited set of knowledge bases for which the retrieval process is already implemented. However, it is meant and was designed to be easily extensible by the community, and the retrieval processes only have to be implemented once and be reused by multiple approaches.

Still, Comprior is a prototypical implementation that has limited functionality with regards to the available knowledge bases, preprocessing methods, feature selection approaches, performance measurements, and visualizations. In particular, Comprior currently does not cover normalization as an important preprocessing step and, furthermore, does not support evaluation in a prediction context, e.g. for survival rates or treatment outcomes. Extending Comprior by these functionalities and thus broadening its application range is the primary objective for future releases of Comprior.

7.1.3 Transferability of our Concepts to Other Omics Domains

We defined our aforementioned concepts on prior knowledge and prior knowledge approaches in the context of their application to gene expression data, i.e. in a transcriptomics context. However, feature selection is also applied to other omics data, e.g. proteomics, methylation, single-cell, and multi-omics data [125, 254, 259, 266]. Besides that, the idea of integrating prior knowledge is not new and has also been applied to these domains in a use-case-tailored fashion [69, 126, 239]. Therefore, the question here is how applicable our concepts are to other omics domains. While we cannot provide a definite answer on this topic without in-depth domain knowledge, we suggest considering the following aspects.

The central point is to identify the connecting factors for prior knowledge in the data set. In the context of feature selection, these connecting factors are — as with gene expression data — the features and the use case domain. While the features constitute the items of interest, e.g. genes or proteins, the use case domain, e.g. a disease, limits the information space. Furthermore, our definition of pior knowledge is not restricted

to a particular entity type, e.g. genes, but instead deals with *biological entities* in general. As long as the features of a data set correspond to such a biological entity, e.g. a gene or protein, it is potentially possible to incorporate prior knowledge as by our definitions. Consequently, our definition of prior knowledge can also be applied to omics domains that deal with other biological entities. Looking at how feature selection is conducted in other omics domains, we observe that the general classification of feature selection approaches into filter, wrapper, embedded, hybrid, and ensemble approaches is consistent across the domains. Even the same basic statistical approaches are applied, e.g. SVM-RFE [125, 254, 259, 266]. At this point, we omit a consideration of necessary data preprocessing steps because suitable approaches are available. For our further considerations, we expect that the respective omics data artifacts fulfill the necessary requirements to apply feature selection approaches to them. As long as a feature selection approach produces a candidate set of features and fulfils our described requirements, it should be possible to apply our presented concepts. Modifying approaches, which filter or extend an existing feature set that has been retrieved by a feature selection approach, can be straightforwardly applied because the feature selection approach does not directly interact with the retrieved prior knowledge. Combining approaches that apply formal frameworks typically use standard statistical approaches, e.g. penalized regression strategies or Bayesian priors. These methods are already applied to the data sets of many biological domains [92]. Combining approaches that apply a process-oriented combination often apply processing strategies that are tailored towards a particular use case and data type, which renders them unlikely to be transferable across omics data types. Nevertheless, it is still possible to develop novel combining approaches for omics domains other than transcriptomics. In regards to network approaches, research has already applied biological networks, e.g. protein-protein interaction networks, to other omics domains, e.g. to proteomics and epigenomics [69, 126].

However, the corresponding prior knowledge must also be available for integration. Our approach focuses on the automatic retrieval of prior knowledge and integration thereof into the analysis. As such, it is essential that prior knowledge is provided in a publicly available database that allows for automatic prior knowledge retrieval. While we have shown that there is a plethora of knowledge bases available on genes and their relevance for particular diseases (see also Tables 2.2 to 2.4 in Section 2.4), the situation can be different for other biological entities and domains of interest. For proteomic data, *UniProtKB* may constitute a good starting point, as it enjoys great popularity for providing a comprehensive collection of functional information on proteins and rich annotation therof [228]. For epigenetic data artifacts, such as methylation data, *diseaseMeth* and *PubMeth* provide information on DNA methylation in disease contexts, e.g. cancer [144, 256]. However, not all of them provide automatic information retrieval, e.g. via a RESTful API. It is further possible that the actual retrieval methods differ from what we applied in this work. For gene expression data, most of the knowledge bases take over large parts of information retrieval, e.g. by providing gene- and disease-

centric search functionality. However, we cannot say whether this is also the case for knowledge bases of other domains.

7.1.4 Does More Flexibility Result in Improved Application in Practice?

Thus far, our contributions so far aim to improve the applicability of prior knowledge approaches by making them more flexible for different use cases and streamlining prior knowledge retrieval. While these are important first steps to encourage the research community for an application in practice, they merely cover the practical aspects of applicability. However, there are further limitations to consider and concerns to address to enable a widespread adoption of prior knowledge approaches.

Comprior already addresses many of the practicability issues for prior knowledge approaches by providing a range of default knowledge bases for prior knowledge retrieval and a comprehensive framework for setting up custom experiments. Comprior, as software tool, is tailored towards an effective evaluation of approaches and not towards including prior knowledge approaches in custom analysis workflows. Traditional approaches are still a step ahead regarding their practicability as they are typically available as R or Python packages without much technical overhead, which allows a straightforward incorporation into individual analysis workflows. The incorporation of online knowledge bases by prior knowledge approaches further poses additional requirements on the usability: to function properly, they require a stable internet connection, reachable knowledge bases, and a continuous maintenance to resolve API changes by knowledge bases.

Furthermore, besides all the advantages that are expected from incorporating prior knowledge early in the analysis, there are multiple issues that might hinder the community from relying on such approaches. First, incorporating prior knowledge can introduce bias into the analysis towards well-annotated genes [83]. The actual risk for it differs for the individual prior knowledge approaches: while prefiltering approaches, which typically remove all genes without annotation information, are particularly affected by this annotation bias, the risk is lower for modifying or network approaches — depending on their actual integration strategy. Another crucial aspect is related to the quality of prior knowledge. Automatically retrieved and integrated prior knowledge, as we have described, does not undergo quality control once retrieved from a knowledge base. While the information contained in most popular knowledge bases is of high quality, not being able to influence which prior knowledge is actually incorporated can likely lead to low confidence in the method. However, users can, to a certain extent, regulate the quality of prior knowledge by a diligent choice of high-quality knowledge bases and thoughtfully selected query terms for prior knowledge retrieval. A solution to address the confidence issue could be to introduce an optional, intermediate step for quality control that allows users to review and filter the retrieved prior knowledge before forwarding it to the actual feature selection step.

Second, it is important to assess how misinformation in knowledge bases, e.g. misannotation, has the potential to negatively affect analysis results. While most biological information collected by knowledge bases undergoes strict review processes, e.g. as described for Uniprot, there is still a chance of incorporating misinformation, e.g. due to inconsistencies or errors along the process [227]. The risk of introducing misinformation further increases when the biological information is computationally derived, which is likely to be the case for the majority of biological information in the future [97]. Detecting and correcting misinformation in knowledge bases is an ongoing field of research that has already greatly improved data quality [37, 201, 250]. For prior knowledge approaches, to date there are no studies that examine the impact of misinformation on the analysis results. However, the severity of misinformation propagating throughout the analysis mainly depends on a) how much weight is granted to prior knowledge along the feature selection process, and b) how popular that misinformation is. Modifying prior knowledge approaches are not likely to be severely affected, as the analysis still mainly focuses on the signals in the data. Still, they can facilitate the gene annotation bias by removing unannotated genes and focusing only on well-researched genes [83]. Combining approaches that apply a formal framework can assign custom weights to the biological relevance of a gene, and misinformation has a larger potential to propagate. While such information is typically quantified by evidence, e.g. by a higher score to indicate that more studies have confirmed a fact, it is still possible that it is misinformation, e.g. because of false insights from contaminated data spread [26]. Network approaches are more likely to be robust to noise because they consider the joint expression behavior of multiple genes, e.g. a network's member genes, and relate them to sample phenotypes. If a network was misannotated, e.g. with incorrect functional information, and thus wrongly considered to be relevant, its member genes will likely not show strong signals in the data. If single interactions were misannotated, e.g. with wrong interaction partners or adverse effects, it is unlikely that these will have a strong effect unless these errors occur at a high frequency in the network. In the end, it is crucial for prior knowledge approaches to carefully select resources for prior knowledge and balance its influence. Prior knowledge should never dominate the analysis but rather be considered as additional information that supports the identification of the true signals in the data.

Third, the information provided by knowledge bases is subject to change and can impede the reproducibility of analysis results. Most of the available knowledge bases update their content regularly (we provide a corresponding overview in Tables 2.2 to 2.4 in Section 2.4). Such updates can include revisions and extensions of existing information which have the potential to change outcomes of prior knowledge approaches. It has been recently shown by Tomczak et al. that the evolution of Gene Ontology can affect the interpretation and reproducibility of experiments over time [233]. While their analyses focused on enrichment analyses, their findings are also relevant for prior knowledge approaches, and their results of the same experiment setting will likely change over time.

Finally, a major challenge in the evaluation of prior knowledge approaches lies in assessing the biological validity of their results. Typically, knowledge bases are only used for validation after the analysis, e.g. via gene set enrichment [215]. Prior knowledge approaches use knowledge bases already during the analysis, thus the resources from these knowledge bases must not be used for validation. Especially when meta knowledge bases are used for feature selection, finding adequate resources for validation can become a hard task because meta knowledge bases already integrate a wide range of resources. It is thus essential to thoroughly review the knowledge bases and their sources of information, so as to avoid accidentially introducing bias into the assessment.

Some of the issues related to incorporating prior knowledge into the analysis are also relevant for traditional analyses that incorporate biological knowledge only at the end of the analysis, e.g. evolving knowledge bases and contained misinformation. However, these issues can have more severe consequences for prior knowledge approaches, as these directly influence the analysis results. It is therefore crucial that prior knowledge approaches always carefully balance the influence they grant to prior knowledge in the analysis and that users carefully assess the quality of prior knowledge in advance. Lastly, prior knowledge approaches will not be adopted until their effectiveness is proven. Our case study provides initial insights into if and how prior knowledge affects performance results.

7.2 The Impact of Integrating Prior Knowledge into Feature Selection

With Comprior as a powerful benchmarking tool at hand, we conducted a comprehensive case study to examine the performance of both traditional and prior knowledge approaches on data sets from multiple disease domains. In particular, with the findings from this case study, we address our second research question:

RQ2: *What is the impact of integrating prior biological knowledge on different analysis levels of biomarker detection regarding the*

a) delivery of interpretable and biologically meaningful results,

b) robustness across approaches and data sets, and

c) computational complexity and transparency?

In the following, we discuss the key insights we retrieved from the results of our conducted case study. We further discuss potential threats that could affect the internal validity of this case study.

7.2.1 The Choice of Knowledge Base Affects Performance Results

We examined the four knowledge bases Open Targets, DisGeNET, PathwayCommons, and KEGG with regards to their coverage and impact on performance results. Open

Targets and DisGeNET both showed a moderate coverage across all disease domains and showed similar classification performances with no clear winner. Although Open Targets delivered generally higher relevance scores than DisGeNET, prior knowledge approaches using the latter retrieved more enrichments. The question of whether these enrichments are truly relevant for the respective use case or are only of more general nature was not covered and is a subject for further studies. The effects are more visible when comparing PathwayCommons and KEGG: PathwayCommons as a meta knowledge base delivered considerably more pathways than KEGG for all disease domains. However, the high coverage in PathwayCommons did not lead to more robust feature sets or enrichments, although the classification performance remained competitive. In contrast, KEGG only showed a limited coverage. Prior knowledge approaches using KEGG almost always showed a clear loss in classification performance, which cannot be made up for by the many more enrichments retrieved for feature sets. What is more, KEGG most often failed to deliver enough pathways for network approaches to work. This renders KEGG unsuitable for our tested approaches. Based on these insights, we recommend using Open Targets for modifying and combining approaches and PathwayCommons for network approaches.

7.2.2 Prior Knowledge Approaches are Feasible, but not Real-Time

We measured the runtime performances of our tested approaches and, more specifically, also of the actual prior knowledge retrieval across different knowledge bases. Prior knowledge approaches were almost always outperformed by traditional approaches, which typically took less than a minute to compute. Most often, the time needed for prior knowledge retrieval alone exceeded the overall runtime of traditional approaches. Prior knowledge retrieval took particularly long for network information: while it often took less than a minute for Open Targets and DisGeNET, it required multiple hours for PathwayCommons. These high retrieval times are mainly caused by the additional processing step to transform the retrieved network information from its original format, e.g. JSON, into a processible data structure.

We further observed differences in overall runtime performances between the individual categories of prior knowledge approaches. Modifying prior knowledge approaches showed the smallest differences, and prefiltering approaches could actually achieve runtimes that were able to compete with traditional approaches when executed on large data sets. Runtime performances of combining approaches were quite diverse: while they increased only marginally for our generalized weighting approach compared to traditional approaches, the approach by Zeng et al. using Lasso regression needed multiple hours for computation [261]. This is, however, the curse of more complex machine learning approaches and will consequently affect other sophisticaed integrative approaches as well. While we have not tested process-oriented prior knowledge approaches, we expect that they will also show considerably higher runtimes than traditional approaches. As multiple processing steps are typically involved, e.g. clustering or network-building, this will surely take its

toll on computation runtime [131, 218, 234]. These high runtimes are only outperformed by network approaches, which, in addition to the already high retrieval times of network information, require an additional and computationally intensive processing step that maps the original feature space to a new one for further processing. In fact, feature mapping accounts for a major part of overall runtime performance and can require up to multiple hours.

With the increasing runtime performances for prior knowledge approaches, the analysis is likely shifted from delivering instant results to a more time-consuming process. However, we want to emphasize that, in current bioinformatics analyses, the focus is not on computational runtimes but on improving the quality of results. This is also reflected in research manuscripts for most feature selection approaches, which generally do not cover computational runtimes in their evaluations. Researchers are thus willing to adopt an approach if it promises improved results, whereas aspects like computational runtimes are currently seldom considered. In fact, bioinformatics starts to deal with computation-intensive tasks by moving to cloud computing and developing methods for effective profiling and resource allocation [204, 236].

7.2.3 Marginal Improvement in Classification Performance, but more Enrichments and Higher Robustness

In general, the classification performance of all tested approaches in our case study was on a very similar level and differences thereof were typically only marginal. This complies with observations made in recent research in this domain [3, 131, 250]. However, traditional approaches select different feature sets than prior knowledge approaches, and feature sets of the latter retrieve more enrichments and show a higher robustness across data sets. This suggests that prior knowledge approaches are more capable of detecting the truly underlying biological processes, whereas traditional approaches likely happen to detect signals from other, unrelated processes that may overlap by accident. This is plausible, as bulk RNAseq data aggregates the expression levels of a population of cells. As such, they always represent a mixture of signals from multiple processes taking place in many cells at the same time. As a consequence, there is high overlap and a lot of noise, e.g. signals of processes going on in particular cells and not being relevant for the use case. Evidence for this is what has been recognised as *random bias* in bulk RNAseq data sets: randomly selected genes show a high and robust predictive or distinctive power that lies above the expected [164, 241]. As shown by Shimoni, random bias particularly affects some of the TCGA gene expression data sets. For some of them, adjustment to the signature of the proliferating cell nuclear antigen (PCNA) helps to lower the effects of random bias [200]. PCNA is a proliferation promoting protein whose expression shows a correlation with a wide number of genes [63, 134]. From our case study, we can confirm that the classification power of randomly selected genes was particularly high for both TCGA and the SCAN-B data sets. This complies with the observations by Shimoni, who detected the highest random bias for the TCGA-GBM/LGG data set. However,

our own experiments and the results of Shimoni showed that random bias could not be fully addressed by PCNA adjustment either. This indicates that there might be other biological processes overlapping here that allow accidentally classifying samples into the right classes by proxy of this process. Consequently, we suggest putting more emphasis on assessing the biological relevance of the selected feature sets and, in particular, their enrichments, and also on examining their robustness across data sets.

7.2.4 Different Integration Levels Affect Biomarker Results

We compared prior knowledge approaches applying different levels of integration, i.e. modifying, combining, and network approaches, to examine whether more sophisticated integration strategies have an observable effect on the performance results. Indeed, we identified differences in both classification performance and enrichment robustness between modifying, combining, and network approaches. In particular, we observed two aspects.

First, classification performance decreases with increasing integration levels. While modifying approaches showed the highest classification performances, both combining and network approaches showed a decrease, with network approaches always performing at the lower boundaries. This behavior seems logical, as more sophisticated integration strategies give more weight to external information and rely less on the signals in the data.

Second, there seems to be a tradeoff between classification performance and the quantity and quality of enrichments. We observed that approaches that showed a higher classification performance typically were not very successful in retrieving enrichments for their feature sets. If they did, these enrichments did not show high robustness across data sets. Network approaches — which generally performed at the lower end of classification — consistently showed the highest robustness of enrichments. We noticed this behavior not only for the different categories of prior knowledge approaches but also on an individual basis. For example, prior knowledge approaches using PathwayCommons showed the highest classification performance on the glioma data sets. At the same time, their feature sets seldom retrieved enrichments, and if so, these enrichments were not robust across data sets. Vice versa, prior knowledge approaches applying KEGG consistently showed lowest the classification performance, though they excelled in the number of enrichments retrieved for their feature sets — and their robustness. This robustness across data sets indicates that the retrieved enrichments are biologically relevant for the actual use case. However, it is still possible that they also accidentally captured an unrelated biological process that happened to be present in both individual data sets.

7.2.5 Modifying or Network Approaches are the Methods of Choice

Based on the insights of our case study, we recommend applying either modifying or network approaches for biomarker detection on gene expression data. Comparing prior

knowledge approaches of the different categories, we observed that modifying approaches showed the largest improvements in classification performance. For example, applying prior knowledge in a prefiltering fashion before executing ANOVA increased classification performance to nearly the same level as the best-performing SVM-RFE. We also observed that combining approaches typically showed a classification performance lower than modifying, yet slightly higher than network approaches. However, their classification performance often decreased on cross-validation data sets, i.e. they were not robust. In particular, deriving feature-specific penalty terms (*LassoPenalty*, combining approach) often showed a low classification performance on the original data set but performed substantially worse when applied on a cross-validation data set. In addition, their feature sets seldom retrieved any enrichments. We have not tested process-oriented prior knowledge approaches and thus it remains unclear whether these types of prior knowledge approaches would exhibit a better and more robust performance than prior knowledge approaches that use formal frameworks.

While modifying approaches, in particular, showed superior classification performance most of the time, network approaches excelled regarding the robustness of retrieved enrichments. We argue that enrichments that are retrieved by the same approach on two independent data sets from the same domain are biologically relevant. Even when the feature sets, i.e. pathways, differed across data sets, their enrichments were still semantically similar and, as such, are likely to represent the same or similar biological processes. This agrees with expectations and observations of other researches, which argue that a module-based approach will better capture biological processes and show an increased robustness [41, 262]. These findings once more support the idea that analyses should focus more on retrieving robust enrichments instead of robust gene sets because processes are typically too complex to be captured by a limited set of genes. Still, network approaches need to improve their classification performance and require advanced computation strategies, as they otherwise have excessive runtime performances.

Given the benefits and drawbacks of both modifying and network approaches, the final choice for an approach depends on a researcher's individual criteria. If the final feature space should be the original one, i.e. genes, the classification performance stays high, and the analysis runs not much longer than traditional approaches, then the choice should be to use a modifying approach, e.g. prefiltering ANOVA by relevant genes. If the main focus is on retrieving robust and biologically relevant feature sets and enrichments, regardless of computational runtimes and losses in classification performance, then network approaches should be favored.

7.2.6 Do Prior Knowledge Approaches Keep Their Promises?

Research suggests that the early integration of biological context, e.g. via prior knowledge, can address current issues of traditional approaches, i.e. lead to biomarkers that are both robust and actually biologically relevant. Our case study is the first of its kind to examine whether these expectations hold for prior knowledge approaches in a

broader context. According to the results from our case study, these expectations have only been partially fulfilled. Indeed, prior knowledge approaches retrieved feature sets of higher biological relevance, i.e. enrichments were more robust across data sets. Most of the prior knowledge approaches also showed a competitive classification performance compared to traditional approaches. However, the observed improvements in classification performance were not robust, i.e. classification performance fell back to the same level as traditional approaches on a second data set. Furthermore, prior knowledge approaches — except for network approaches — did not show an increased robustness of the selected features themselves, i.e. the same approach selected different feature sets on two data sets from the same domain. However, these features then still retrieved semantically similar enrichments, so it is likely that the same underlying biological process was identified, yet probably at different stages of activity. In the end, gene expression data from microarrays and bulk RNAseq contains a mixture of different cell populations, with multiple biological processes going on at the same time, snapshotting gene activity of the same process at different stages. From the nature of gene expression data and what we have observed here, we actually question the general assumption of whether it is possible at all to find features in the original feature space, i.e. genes, that lead to a robust classification performance across data sets. This is a strong argument for the power of network approaches, as these select modules as features that can capture these similarities. Still, network approaches showed consistently lower classification performance, which as of now we can only explain with potential noise in the data. We therefore recommend that the choice for a prior knowledge approach should depend on the main focus of the analysis: in a rather diagnostic use case, i.e. a reliable classification of samples is important, modifying prior knowledge approaches are the method of choice. If the aim is to detect novel biomarkers that have a true biological relevance, then network approaches should be favored.

One result which raises questions is the unexpectedly poor performance observed for combining approaches. Especially when sophisticated strategies are applied, e.g. the feature-specific penalties by Zeng et al., neither classification performance nor biological relevance was improved [261]. Even worse, they showed a much decreased robustness in all areas when applied on a different data set, which indicates either overfitting or lacking quality of the applied prior knowledge.

At this point, the findings and conclusions drawn from our case study are generalizeable to all traditional and prior knowledge approaches, but only in the current setting of our case study. The focus of this thesis, in particular, lay on the theoretical concepts and their practical realization for feature selection. Retrieving actual biological insights with relevant implications, e.g. in a clinical context, would require further domain knowledge and experimental validation in a lab, which were not available for our case study.

7.3 Directions for Future Work

Our work provides a first step towards understanding how prior knowledge can be flexibly incorporated into the analysis of gene expression data and what positive effects it has on analysis outcomes. Our results raised important questions that can serve as starting points for future investigations.

In particular, future work should investigate the interplay between the classification performance and biological relevance of the selected features. From our results, we observed a tradeoff between both, meaning that researchers currently have to decide in favor of one of them, simultaneously neglecting the other. Future prior knowledge approaches should strive to achieve competitive results in both aspects, which is only possible with an in-depth understanding of their mutual relationship. Focus should also be put on combining approaches, as the results of our case study attested them unexpectedly poor performance. As we assume the quality of prior knowledge to be a potential cause for this, a first starting point is to compare combining approaches using prior knowledge of different quality levels, e.g. comparing highly specific and manually curated prior knowledge with that retrieved from knowledge bases with more general search terms. This likely involves case studies using simulated data, which has the advantage of providing a single truth and testing how well approaches capture the true biological processes in the data. These simulation studies should then be carried forward to a general examination of how the quality of prior knowledge affects analysis outcomes. In particular, future studies should investigate if and how misinformation propagates in the different types of prior knowledge approaches and whether the degree of specialization has a visible effect on performance results. To further assess the effectiveness of prior knowledge approaches, they should be compared to other types of integrative analyses, e.g. those applying multi-omics. It should be investigated whether prior knowledge approaches can keep up with multi-omics, being less computationally complex and requiring fewer data artifacts for the analysis. As we have previously discussed the general possibility, a next step would be to transfer our concepts to a multi-omics setting and assess whether this brings reasonable benefits for the analysis.

An upcoming field of research concerns population-specific differences in gene expression. Research has started to investigate whether disease mechanisms differ in other populations. Indeed, results of recent studies indicate that there are population-specific differences, e.g. in our applied disease domains, that might have a clinical impact [100, 107, 110]. For example, Pan et al. identified distinct immune gene expression profiles for breast cancer in Asian populations compared to European populations [147]. The majority of existing biological findings, and consequently the prior knowledge provided in knowledge bases, originates from studies on western populations [110, 147]. All the data sets we applied in our case study contain probes taken from individuals from Sweden (SCAN-B), Finland (AddNeuroMedI/II), and the USA (TCGA, REMBRANDT), the majority of which thus having European ancestry. We can therefore assume that the

above described bias towards western populations in knowledge bases has no severe effect on our analysis. It is, however, relevant to find out whether prior knowledge approaches work equally well on data sets from non-western populations that show different molecular signatures for particular diseases, and if so, what measures must be undertaken for prior knowledge approaches to address these concerns. As this is an emerging field of research, we expect further insights to follow in the coming years.

8

Conclusion

This chapter concludes the thesis by summarizing the key findings of our work. It discusses our main contributions and how they address our initially posed research questions.

At the beginning of this thesis, we identified multiple shortcomings with traditional, purely data-driven feature selection approaches on gene expression data. In particular, biomarkers detected with traditional feature selection approaches show a low robustness and questionable biological relevance. Prior knowledge approaches, which incorporate prior biological information directly into the feature selection process, are expected to resolve these issues and lead to more robust biomarkers with actual biological relevance. Throughout this thesis, we investigated how we can improve the application of prior knowledge approaches in practice and whether they can mitigate the present issues of purely data-driven feature selection approaches. Based on our observations, we formulate two research questions: 1) how to foster the application and comparability of prior knowledge approaches in practice, and 2) how the integration of prior knowledge actually affects classification performance, biological relevance of the retrieved features, and computational complexity. We addressed these research questions with the three key contributions of this thesis.

In our first contribution, we identified a general use case inflexibility and individual requirements to the incorporated prior knowledge as the two major drawbacks of existing approaches. In response to these issues, our first contribution provides a uniform definition of prior knowledge and how it can be flexibly incorporated by prior knowledge approaches. The central idea is to streamline and unify the retrieval of prior knowledge from available online knowledge bases that provide the latest research insights, e.g. on gene-disease associations. We examined what kind of prior knowledge is typically incorporated by existing approaches and reviewed what information is available in suitable knowledge bases. From these insights, we subsequently derived three levels in which prior knowledge is available for integration: as a list of *biological entities*, as a list of *scored biological entities*, and as a list of *networks*. To even further increase the flexibility of prior knowledge approaches and exploit the full range of available knowledge bases, we

described how prior knowledge can be transformed from one level to the other, e.g. how a list of networks can be transformed to a list of entities and vice versa. We further reviewed related work on prior knowledge approaches according to the applied integration strategies and the kind of prior knowledge required. We subsequently identified three general categories of prior knowledge approaches: *modifying*, *combining*, and *network approaches*. To further address the flexibility issue, we described novel approaches that incorporate prior knowledge, based on our definitions, and are therefore applicable to multiple disease domains. Our formal concepts on prior knowledge and its integration into feature selection address our first research question, by providing both novel and flexible prior knowledge approaches for application in practice and a formal framework that can be applied by future prior knowledge approaches.

In our second contribution, we applied our theoretical concepts to the benchmarking framework *Comprior*. Comprior allows for evaluating feature selection approaches by rapidly setting up benchmark experiments that cover selected preprocessing steps, the actual feature selection, and subsequent classification with extensive cross-validation. Comprior provides visualizations for multiple evaluation measures to assess the effectiveness of the tested approaches, ranging from feature ranking comparisons to standard classification measures and gene set enrichment analysis. Comprior further streamlines the prior knowledge retrieval and encapsulates it from the actual feature selection, thus allowing a prior knowledge approach to be combined with any of the available knowledge bases. The currently integrated knowledge bases and incorporated prior knowledge approaches already constitute an appropriate framework for a comprehensive benchmark. Furthermore, Comprior was intentionally designed in a modular fashion that allows for extending its functionality, e.g. to include other knowledge bases or implement novel prior knowledge approaches. Comprior also addresses our first research question by providing the technical infrastructure for rapid development and comprehensive evaluation of prior knowledge approaches, thus fostering their accessibility, flexibility, and subsequently improving overall comparability.

In our third contribution, we conducted a comprehensive case study to assess the effectiveness of integrating prior knowledge into feature selection. We used Comprior to benchmark both our own and existing prior knowledge approaches, compare their performance to traditional feature selection approaches, and examine the effect of different influence factors, e.g. the applied knowledge base or integration strategy. We carried out our case study on gene expression data sets from three disease domains, namely breast cancer, glioma, and Alzheimer's disease. We assessed the effectiveness of the tested approaches based on their classification performance, characteristics of their produced feature sets, retrieved enrichments, and their overall robustness. Our case study ultimately addressed our second research question on the effectiveness of prior knowledge approaches and further demonstrated the feasibility and applicability of our concepts on prior knowledge and prior knowledge approaches. The results of our case study showed that prior knowledge approaches positively affect the performance, particularly in terms

of more enrichments and higher robustness thereof. In our specific setting, we further identified a tradeoff between classification performance and enrichment robustness — prior knowledge approaches typically showed major improvements in either one of these properties, but not in both. From the prior knowledge approaches tested, we concluded that simple integration strategies, e.g. as applied by modifying approaches, are particularly effective in terms of classification performance without requiring too much computational runtime. In contrast, the extraction of modular features via network information required much more computational runtime, though it retrieved enrichments that proved to be more robust. Surprisingly, prior knowledge approaches that incorporated machine learning strategies, e.g. as applied by combining approaches, could neither convince in their classification performance nor in the retrieved enrichments. We further observed that the choice of knowledge base has a visible impact on the performance results. We thus suggest to use Open Targets and PathwayCommons, whereas KEGG is not suitable for network approaches due to its generally low information content. Our case study provides first insights on the effectiveness of prior knowledge approaches, and shows that already simple integration strategies, e.g. prefiltering, can have a major impact on the performance. Our results further fortify findings from other studies on network approaches and corroborate the assumption that module-based approaches lead to more robust and biologically relevant results. Based on our insights, we propose to direct future research towards a) investigations on the unexpectedly poor performance of combining prior knowledge approaches, b) examinations of the interdependency between classification performance and enrichment robustness, and c) evaluations on how the overall quality of prior knowledge affects performance results, e.g. if and how misinformation can propagate.

Starting from our initial problem statement — that prior knowledge approaches are not widely applied in practice — we identified their missing applicability and unclear effectiveness to be key obstacles. Our formal concepts on prior knowledge and its integration into feature selection, together with their implementation in Comprior, demonstrate that prior knowledge can indeed be incorporated in an efficient way and effectively address the main drawbacks that exist for traditional feature selection approaches. The case study we have conducted is the first to examine the interplay between prior biological knowledge and data-driven analyses to a larger extent, which constitutes a thorough foundation to build on and points out important directions for future research.

List of Figures

2.1 Overview of processing steps during biomarker detection 9
2.2 Process of gene expression in a cell 11
2.3 Microarray processing steps 13
2.4 Next-generation sequencing process on Illumina machines 14

3.1 Overview on prior knowledge integration strategies 30

4.1 The two possible process flows for modifying approaches 44
4.2 Process flow of combining prior knowledge approaches 46
4.3 Process overview for network approaches 47

5.1 Overview of Comprior's functionality 54
5.2 Overview of Comprior's system components 56
5.3 Class structure of the preprocessing module 59
5.4 Class structure of the FeatureSelection module 62
5.5 Class structure of the Evaluation component 63
5.6 $m:1$ mappings dealt with in Comprior 65
5.7 $1:n$ mappings dealt with in Comprior 66
5.8 Example class structure of a knowledge base 68

6.1 Summary on prior knowledge coverage across the four knowledge bases . . . 80
6.2 Average time needed to retrieve prior knowledge for a particular disease from the different knowledge bases 83
6.3 Runtime performances of traditional feature selection and prior knowledge approaches on the TCGA-BRCA data set 84
6.4 Classification performances of feature sets selected on the TCGA-BRCA data set 86
6.5 Classification performances of feature sets selected on the TCGA-GBM/LGG data set 87
6.6 Classification performances of feature sets selected on the AddNeuroMedI data set 88

6.7 Overlaps of feature sets selected by the tested approaches on TCGA-BRCA, TCGA-GBM/LGG, and AddNeuroMedI 89
6.8 Robustness across data sets of feature sets selected by the best-classifying approaches from TCGA-BRCA, TCGA-GBM/LGG, and AddNeuroMedI . 90
6.9 Semantic similarities of enriched MSigDB oncogenic signatures and pathways on the TCGA-BRCA data set 92
6.10 Semantic similarities and robustness of GO term enrichments (and pathways) on TCGA-GBM/LGG 94
6.11 Semantic similarities of enriched GO terms on AddNeuroMedI 95
6.12 Robustnes and biological relevance of enrichments retrieved for feature sets selected on TCGA-BRCA, TCGA-GBM/LGG, and AddNeuroMedI data sts. 96
6.13 Average classification performances (for 5 to 25 features) of ANOVA, Lasso, and SVM-RFE and their adaptations on all six data sets 98
6.14 Distributions of distinct features selected exclusively by an approach 99
6.15 Distributions of RBO scores of feature sets for traditional approaches and their adaptations using prior knowledge 100
6.16 Robustness of enrichments for ANOVA, Lasso, SVM-RFE, and their respective adaptations using prior knowledge 100
6.17 Average classification performances of modifying, combining, and network approaches 102
6.18 Robustness of feature sets and enrichments retrieved by modifying, combining, and network approaches across data sets. 103
6.19 Win/loss plots of the different prior knowledge approaches grouped by knowledge base applied 105
6.20 Robustness of feature sets and enrichments across data sets, grouped by knowledge base 106
6.21 Quantitative assessment of enrichments retrieved by prior knowledge approaches using a particular knowledge base 107

10.1 Detailed class structure of the KnowledgeBase component as implemented in Comprior 142
10.2 Principal component analysis (PCA) plots for the applied data sets 143
10.3 Runtime performances on the SCAN-B (breast cancer) data set of both traditional feature selection and prior knowledge approaches 144
10.4 Runtime performances on the glioma data sets of both traditional feature selection and prior knowledge approaches 146
10.5 Runtime performances of both traditional feature selection and prior knowledge approaches on the Alzheimer's disease data sets 147
10.6 Classification performances of feature sets selected on the SCAN-B data set 148

10.7 Classification performances of feature sets selected on the REMBRANDT data set 148
10.8 Classification performances of feature sets selected on the AddNeuroMedII data set 149
10.9 Overlaps of feature sets selected by the tested approaches on SCAN-B, REMBRANDT, and AddNeuroMedII 150
10.10 Robustness across data sets of feature sets selected by the best-classifying approaches from SCAN-B, REMBRANDT, and AddNeuroMedII 151
10.11 Semantic similarities of enriched MSigDB oncogenic signatures and pathways on the SCAN-B data set 152
10.12 Semantic similarities and robustness of GO term enrichments on the REMBRANDT data set 153
10.13 Semantic similarities of enriched GO terms on the AddNeuroMedII data set 154
10.14 Robustnes and biological relevance of enrichments retrieved for feature sets selected on SCAN-B, REMBRANDT, and AddNeuroMedII data sts. . 155
10.15 Average classification performances on all six data sets (when used for cross-validation) for ANOVA, Lasso, and SVM-RFE and their adaptations 156
10.16 Robustness of feature sets and enrichments on the TCGA-BRCA and SCAN-B data sets 157
10.17 Robustness of feature sets and enrichments on the REMBRANDT and TCGA-GBM/LGG data sets 158
10.18 Robustness of feature sets and enrichments on the AddNeuroMedI and AddNeuroMedII data sets 159
10.19 Quantitative comparison of enrichments retrieved for prior knowledge approaches on the TCGA-BRCA and SCAN-B data sets 160
10.20 Quantitative comparison of enrichments retrieved for prior knowledge approaches applied on the TCGA-GBM/LGG and REMBRANDT data sets 161
10.21 Quantitative comparison of enrichments retrieved for prior knowledge approaches applied on the AddNeuroMedI and AddNeuroMedII data sets . 162

List of Tables

2.1 Excerpt from an example gene expression data set 16
2.2 Online knowledge bases providing annotation information 21
2.3 Online knowledge bases providing interaction data 23
2.4 Online meta knowledge bases 25
2.5 Available data formats and endpoints for accessing online knowledge bases. 27

3.1 Review of existing approaches related to benchmarking omics data sets . . . 34

6.1 Overview of the data sets used 72
6.2 Overview of the applied preprocessing measures per data set 73
6.3 Overview of the applied traditional feature selection approaches 75
6.4 Overview of applied prior knowledge approaches 76

10.1 Overview on the expected input and returned output of relevant knowledge bases 141
10.2 Sample distributions across classes for breast cancer data sets 141
10.3 Sample distributions across classes for glioma data sets 142
10.4 Sample distributions across classes for Alzheimer's data sets 142
10.5 Search terms (and their assigned identifiers) used for retrieving prior knowledge related to Alzheimer's disease 143
10.6 Search terms (and their assigned identifiers) used for retrieving prior knowledge related to Glioma 144
10.7 Search terms (and their assigned identifiers) used for retrieving prior knowledge related to breast cancer 145
10.8 Overview on modifying approaches using prior knowledge and their respective characteristics 163
10.9 Overview on combining approaches incorporating prior knowledge via formal frameworks and their respective characteristics 164
10.10 Overview on process-oriented combining approaches using prior knowledge and their respective characteristics 165

10.11 Overview on network approaches for prior knowledge biomarker detection and their respective characteristics 166
10.12 Qualitative comparison of approaches for prior knowledge biomarker detection 167

9

Publications

9.1 Journal Articles

C. Perscheid. "Comprior: facilitating the implementation and automated benchmarking of prior knowledge-based feature selection approaches on gene expression data sets". *BMC Bioinformatics* 22.1 (2021), pp. 1–15

C. Perscheid. "Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches". *Briefings in Bioinformatics* 22.3 (2020), bbaa151

C. Perscheid. "The impact of integrating prior knowledge during biomarker detection: A case study on high-dimensional gene expression data" (2022). in preparation

C. Perscheid, B. Grasnick, and M. Uflacker. "Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches". *Journal of Integrative Bioinformatics* 16.1 (2019), p. 20180064

C. Perscheid, J. Benzler, C. Hermann, M. Janke, D. Moyer, T. Laedtke, O. Adeoye, K. Denecke, G. Kirchner, S. Beermann, et al. "Ebola outbreak containment: real-time task and resource coordination with SORMAS". *Frontiers in ICT* 5 (2018), p. 7

C. Fähnrich, K. Denecke, O. Adeoye, J. Benzler, H. Claus, G. Kirchner, S. Mall, R. Richter, M.-P. Schapranow, N. G. Schwarz, et al. "Surveillance and Outbreak Response Management System (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa". *Eurosurveillance* 20.12 (2015), p. 21071

M.-P. Schapranow, F. Häger, **C. Fähnrich**, E. Ziegler, and H. Plattner. "In-Memory Computing Enabling Real-time Genome Data Analysis". *International Journal on Advances in Life Sciences* 6.1 and 2 (2014), pp. 11–30

9.2 Conference Articles

B. Grasnick, **C. Perscheid**, and M. Uflacker. "A Framework for the Automatic Combination and Evaluation of Gene Selection Methods". In: *International Conference on*

Practical Applications of Computational Biology & Bioinformatics. Ed. by F. Fdez-Riverola, M. S. Mohamad, M. Rocha, J. F. De Paz, and P. González. Springer. Cham: Springer International Publishing, 2019, pp. 166–174

D. Tom-Aba, S. E. Toikkanen, S. Glöckner, O. Adeoye, S. Mall, **C. Fähnrich**, K. Denecke, J. Benzler, G. Kirchner, N. Schwarz, et al. "User evaluation indicates high quality of the Surveillance Outbreak Response Management and Analysis System (SORMAS) after field deployment in Nigeria in 2015 and 2018". In: *German Medical Data Sciences*. Vol. 253. 2018, pp. 233–237

C. Fähnrich, M.-P. Schapranow, and H. Plattner. "Facing the genome data deluge: efficiently identifying genetic variants with in-memory database technology". In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. 2015, pp. 18–25

M.-P. Schapranow, M. Kraus, **C. Perscheid**, C. Bock, F. Liedke, and H. Plattner. "The Medical Knowledge Cockpit: Real-time analysis of big medical data enabling precision medicine". In: *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2015, pp. 770–775

M.-P. Schapranow, **C. Perscheid**, and H. Plattner. "IT-aided business process enabling real-time analysis of candidates for clinical trials". In: *Proceedings of the 4th International Conference on Global Health Challenges*. 2015, pp. 67–73

9.3 Workshop Articles

C. Perscheid, L. Faber, M. Kraus, P. Arndt, M. Janke, S. Rehfeldt, A. Schubotz, T. Slosarek, and M. Uflacker. "A Tissue-aware Gene Selection Approach for Analyzing Multi-tissue Gene Expression Data". In: *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2018, pp. 2159–2166

C. Perscheid and M. Uflacker. "Integrating Biological Context into the Analysis of Gene Expression Data". In: *International Symposium on Distributed Computing and Artificial Intelligence*. Springer. 2018, pp. 339–343

M.-P. Schapranow, **C. Perscheid**, A. Wachsmann, M. Siegert, C. Bock, F. Horschig, F. Liedke, J. Brauer, and H. Plattner. "A federated in-memory database system for life sciences". In: *Real-Time Business Intelligence and Analytics*. Springer, 2015, pp. 19–34

C. Fähnrich, M.-P. Schapranow, and H. Plattner. "Towards integrating the detection of genetic variants into an in-memory database". In: *IEEE International Conference on Big Data*. IEEE. 2014, pp. 27–32

9.4 Technical Reports

M.-P. Schapranow and **C. Fähnrich**. "Analyze Genomes: a cloud platform enabling on-site analysis of sensitive medical data". In: *Proceedings of the HPI Future SOC Lab.*

Ed. by C. Meinel, A. Polze, G. Oswald, R. Strotmann, U. Seibold, and B. Schulzki. University of Potsdam, 2016, pp. 21–24

M.-P. Schapranow and **C. Perscheid**. "Extending Analyze Genomes to a federated in-memory database system for life sciences". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel, A. Polze, G. Oswald, R. Strotmann, U. Seibold, and B. Schulzki. University of Potsdam, 2015, pp. 99–102

M.-P. Schapranow and **C. Fähnrich**. "Provision of Analyze Genomes services in a federated in-memory database system for life sciences". In: *Proceedings of the HPI Future SOC Lab*. University of Potsdam, 2015, pp. 39–42

M.-P. Schapranow and **C. Fähnrich**. "Setting up customized genome data analysis pipelines with Analyze Genomes". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel, A. Polze, G. Oswald, R. Strotmann, U. Seibold, and B. Schulzki. University of Potsdam, 2014, pp. 11–30

K. Herbst, **C. Fähnrich**, M. L. Neves, and M.-P. Schapranow. "Applying In-Memory Technology for Automatic Template Filling in the Clinical Domain". In: *CLEF Evaluation Labs and Workshop Online Working Notes*. 2014, pp. 91–102

M.-P. Schapranow and **C. Fähnrich**. "High-Performance In-Memory Genome Project". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel, A. Polze, G. Oswald, R. Strotmann, U. Seibold, and B. Schulzki. University of Potsdam, 2014, pp. 11–30

10

Appendix

Name	Input	Output
Gene Ontology	disease or gene identifier	gene-disease associations, annotations
UniprotKB	gene identifier	functional information
Human Protein Atlas	gene identifier	functional information, expression profiles
COSMIC	disease or gene identifier	gene-disease associations (w/o score)
GWAS	disease or gene identifier	gene-disease associations (p-value = score)
IntAct	disease or gene identifier	disease-interaction associations (w/o score)
BioGRID	gene identifier	protein-protein interactions (w/o score)
CTD	gene identifier	gene interactions, gene-disease associations (w/o score)
InnateDB	gene identifier	protein-protein interactions
REACTOME	disease or gene identifier	interaction networks(pathways)
KEGG	disease or gene identifier	interaction networks(pathways)
Wikipathways	disease or gene identifier	interaction networks (pathways)
PathwayCommons	disease or gene identifier	interaction networks(pathways)
ConsensusPathDB	disease or gene identifier	scored interactions, pathways
STRING	disease or gene identifier	interaction networks
DisGeNET	disease or gene identifier	gene-disease associations
Open Targets	disease or gene identifier	gene-disease associations

Table 10.1: Overview on the expected input of relevant knowledge bases and what kind of output, i.e. related prior knowledge, they return.

Data Set	Luminal A	Luminal B	Her2-Enriched	Basal-Like	Normal-Like	Overall
TCGA-BRCA	567	207	82	194	40	1,090
SCAN-B	154	101	65	57	22	399

Table 10.2: Sample distributions across classes for breast cancer data sets.

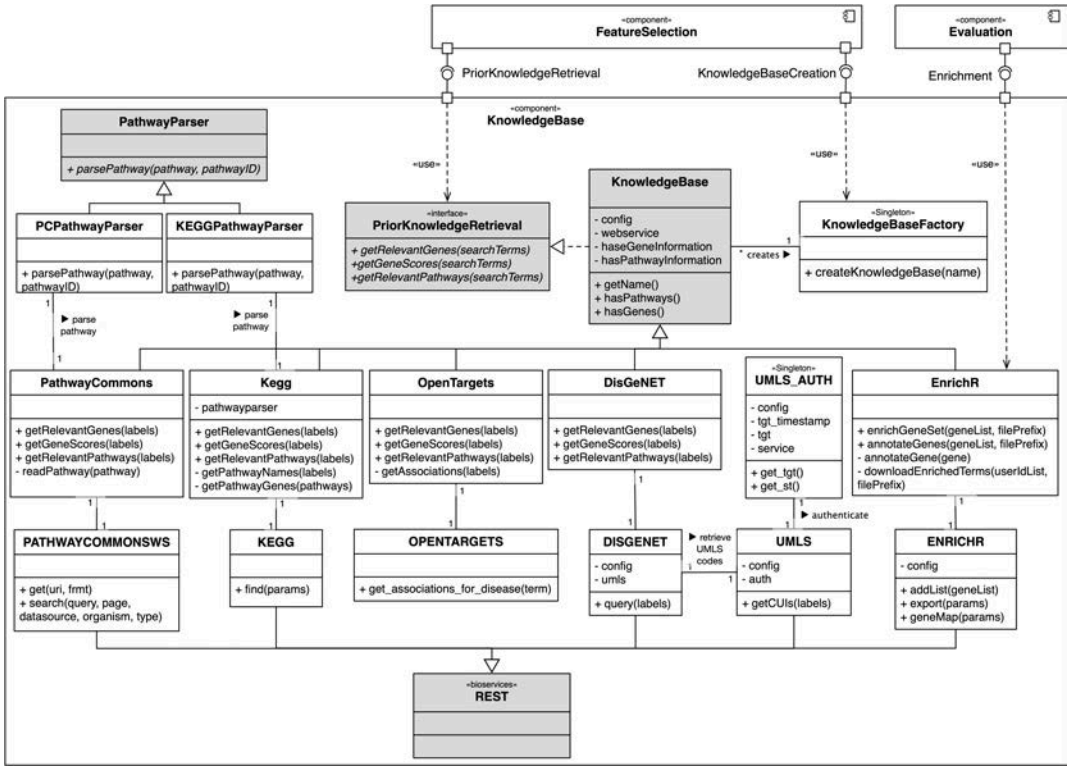


Fig. 10.1: Detailed class structure of the KnowledgeBase component as implemented in Comprior.

Data Set	Glioblastoma Multiforme	Astro-cytoma	Oligo-dendroglioma	Overall
TCGA-GBM/LGG	155	167	174	496
REMBRANDT	221	148	67	436

Table 10.3: Sample distributions across classes for glioma data sets.

Data Set	Alzheimer's Disease	Mild Cognitive Impairment	Normal	Overall
AddNeuroMedI	145	80	104	329
AddNeuroMedII	139	109	134	382

Table 10.4: Sample distributions across classes for Alzheimer's data sets.

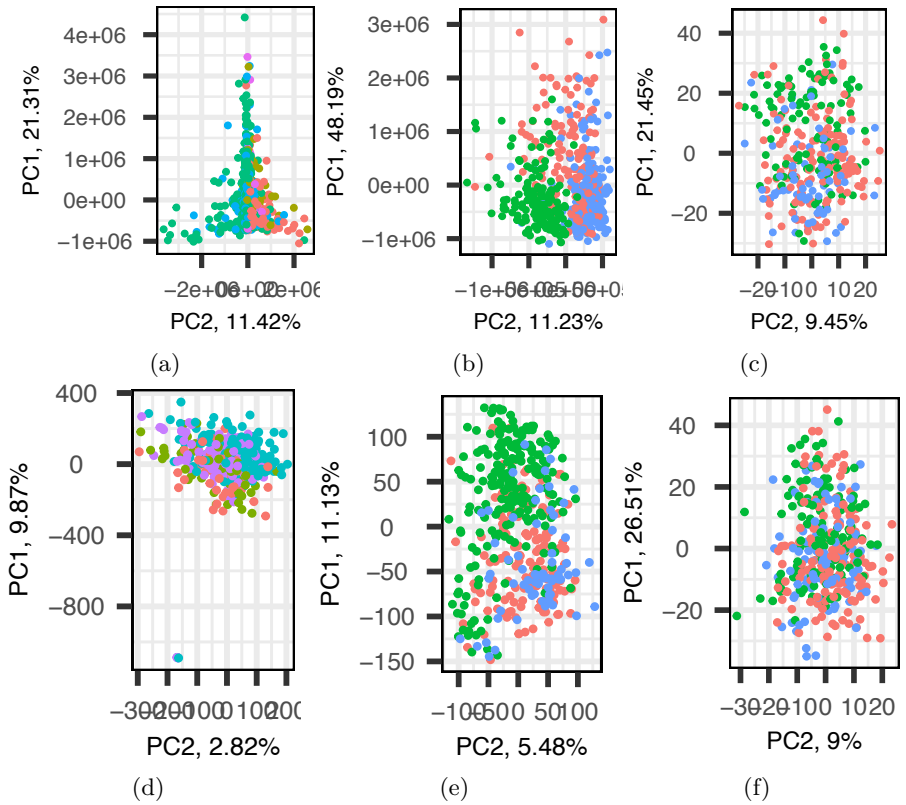


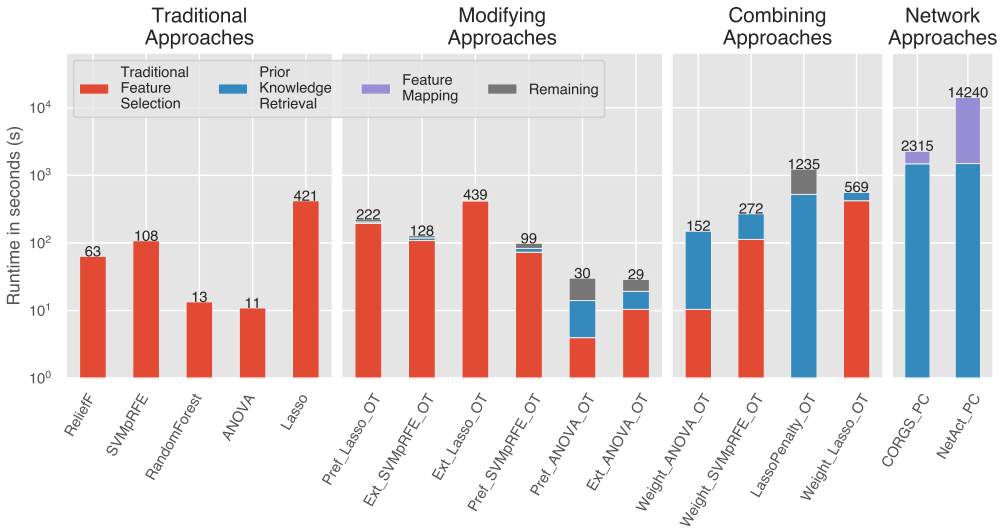
Fig. 10.2: Principal component analysis (PCA) plots for the first and second principal components for a) TCGA-BRCA, b) TCGA-GBM/LGG, c) AddNeuroMedI, d) SCAN-B, e) REMBRANDT, f) AddNeuroMedII data sets. Colors refer to disease subtypes. None of the data sets show an abnormal clustering.

ID	Search Term	ID	Search Term
1	MCI	9	Alzheimer Type Dementia
2	CTL	10	Alzheimers Dementia
3	AD	11	Senile Dementia Alzheimer Type
4	Alzheimers Disease	12	Primary Senile Degenerative Dementia
5	Alzheimer	13	Mild Cognitive Impairment
6	Alzheimer Dementia	14	Mild Neurocognitive Disorder
7	Alzheimer Sclerosis	15	Mild Cognitive Disorder
8	Alzheimer Syndrome		

Table 10.5: Search terms (and their assigned identifiers) used for retrieving prior knowledge related to Alzheimer's disease. As search terms, we used the main disease name, class labels, and their corresponding synonyms as looked up in the National Cancer Institute's metathesaurus browser (`https://ncim.nci.nih.gov/ncimbrowser/`).

ID	Search Term	ID	Search Term
1	Astrocytoma	11	Astrocytoma
2	Oligodendroglioma	12	Oligodendroglioma
3	Glioblastoma	13	Oligodendroglial Tumor
4	Glial Cell Tumor	14	Oligodendroglial Neoplasm
5	Glial Neoplasm	15	Glioblastoma
6	Glial Tumor	16	Glioblastoma Multiforme
7	Glioma	17	Astrocytic Neoplasm
8	Neoplasm of Neuroglia	18	Astrocytic Tumor
9	Neuroglial Neoplasm	19	Spongioblastoma Multiforme
10	Neuroglial Tumor		

Table 10.6: Search terms (and their assigned identifiers) used for retrieving prior knowledge related to Glioma. As search terms, we used the main disease name, class labels, and their corresponding synonyms as looked up in the National Cancer Institute's metathesaurus browser (`https://ncim.nci.nih.gov/ncimbrowser/`).



(a)

Fig. 10.3: Runtime performances on the SCAN-B (breast cancer) data set of both traditional feature selection and prior knowledge approaches. Modifying and combining approaches use Open Targets (OT) as knowledge base, whereas network approaches apply PathwayCommons (PC).

ID	Search Term	ID	Search Term
1	LumB	24	Luminal B Subtype of Breast Carcinoma
2	Basal	25	Normal Breast-Like Subtype of Breast Cancer
3	LumA	26	Normal Breast-Like Subtype of Breast Carcinoma
4	Her2	27	Invasive Breast Cancer
5	Normal	28	Invasive Breast Carcinoma
6	Breast Cancer	30	Infiltrating Breast Carcinoma
7	ERBB2 Overexpressing Subtype of Breast Carcinoma	30	Infiltrating Breast Carcinoma
8	HER2 Overexpressing Breast Carcinoma	31	Invasive Mammary Carcinoma
9	HER2 Positive Breast Cancer	32	Mammary Carcinoma
10	HER2 Positive Breast Carcinoma	33	Breast Carcinoma
11	HER2+ Breast Cancer	34	Infiltrating Carcinoma of Breast
12	Human Epidermal Growth Factor 2 Positive Carcinoma Of Breast	35	Breast Ductal Carcinoma
13	Basal-Like Breast Cancer	36	Mammary Ductal Carcinoma
14	Basal-Like Breast Carcinoma	37	Duct Adenocarcinoma
15	Basal-Like Subtype of Breast Carcinoma	38	Duct Carcinoma
16	Luminal A Breast Cancer	39	Ductal Adenocarcinoma
17	Luminal A Breast Carcinoma	40	Ductal Carcinoma of Breast
18	Luminal A Estrogen Receptor Positive Subtype of Breast Carcinoma	41	Ductal Breast Carcinoma
19	Luminal A	42	Ductal Carcinoma
20	Luminal B	43	Lobular Carcinoma
21	Luminal B Breast Cancer	44	Infiltrating Lobular Carcinoma of Breast
22	Luminal B Breast Carcinoma	45	Lobular Adenocarcinoma
23	Luminal B Estrogen Receptor Positive Subtype of Breast Carcinoma	46	Lobular Breast Carcinoma

Table 10.7: Search terms (and their assigned identifiers) used for retrieving prior knowledge related to breast cancer. As search terms, we used the main disease name, class labels, and their corresponding synonyms as looked up in the National Cancer Institute's metathesaurus browser (`https://ncim.nci.nih.gov/ncimbrowser/`).

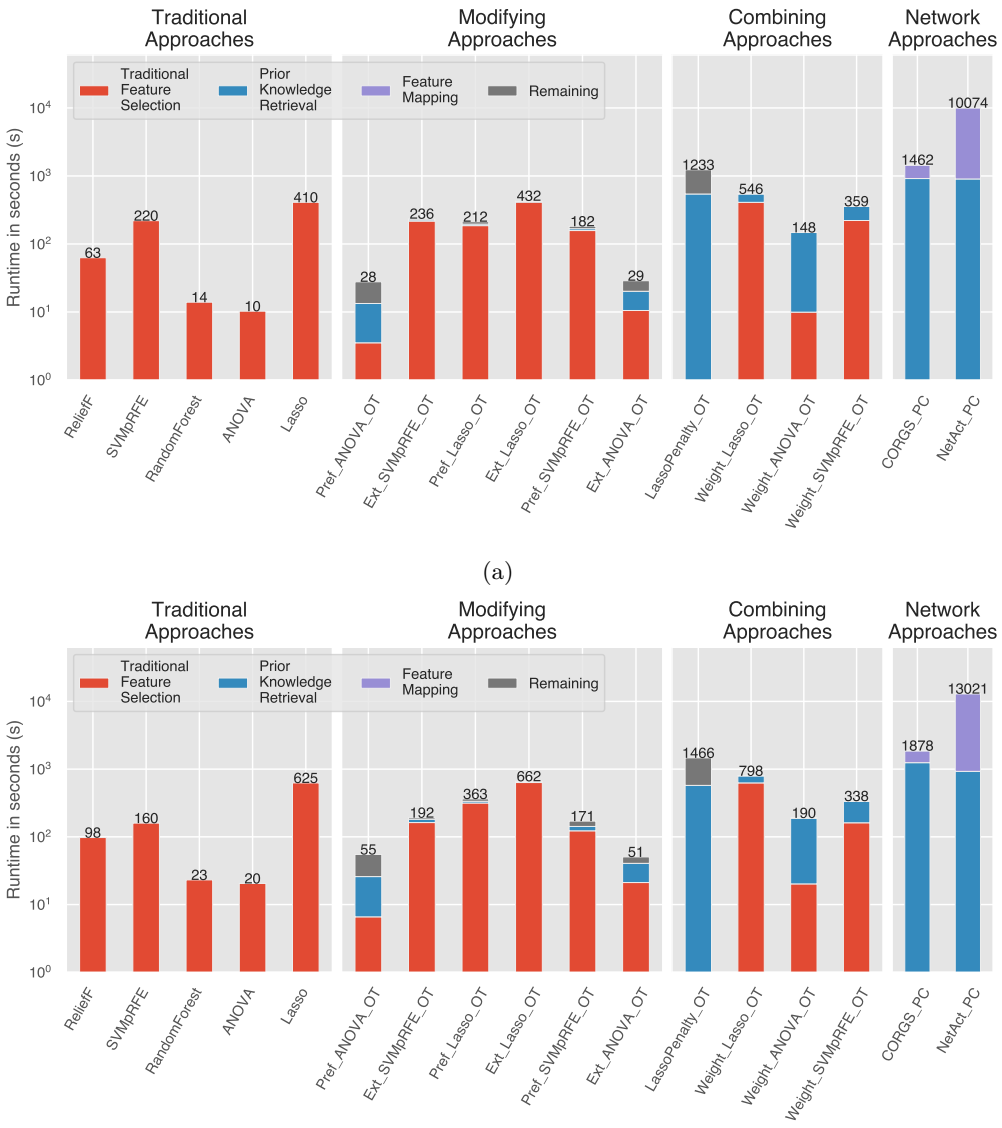
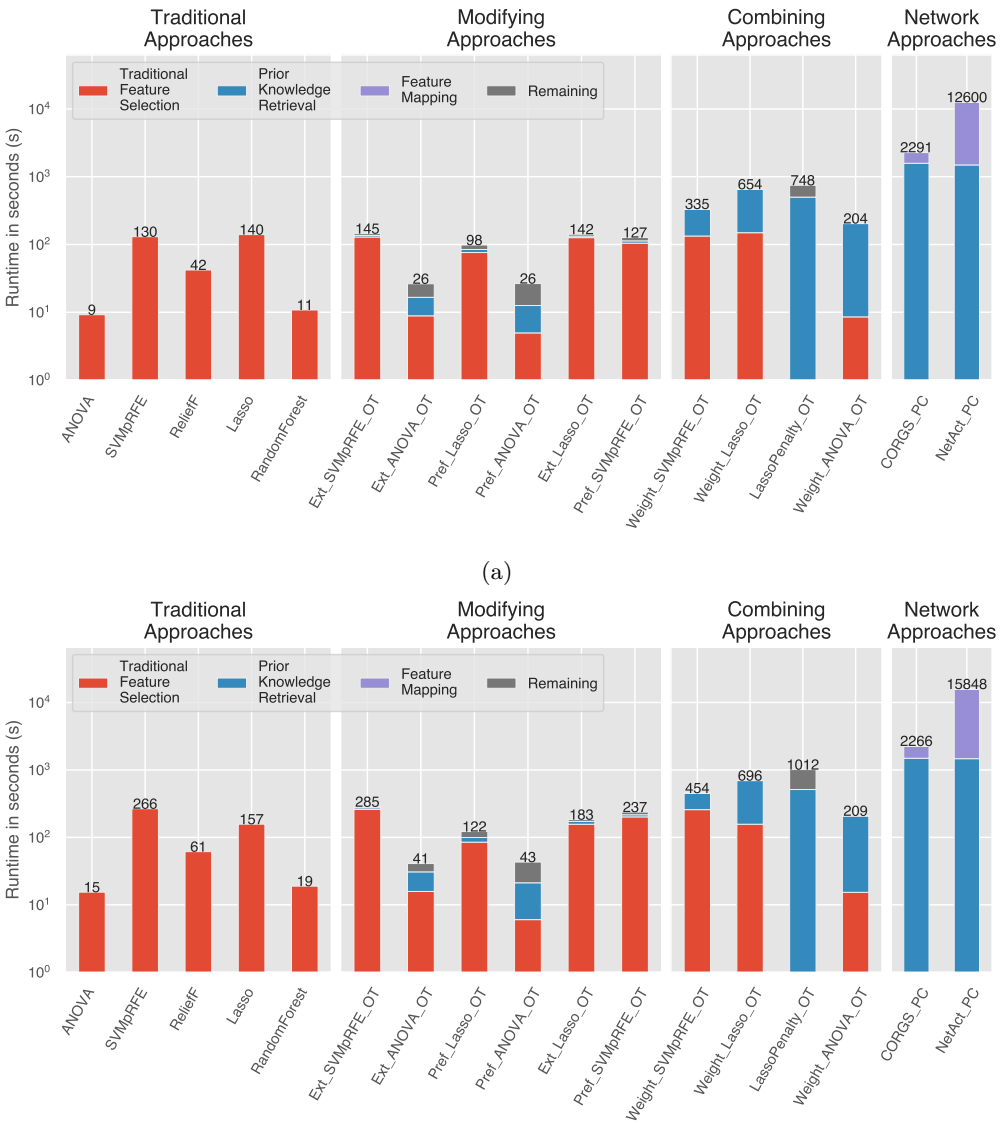


Fig. 10.4: Runtime performances on the glioma data sets of both traditional feature selection and prior knowledge approaches. a) shows performances on the REMBRANDT data set, b) shows performances on the TCGA-GBM/LGG data set. Modifying and combining approaches use Open Targets (OT) as knowledge base, whereas network approaches apply PathwayCommons (PC).



(a)

(b)

Fig. 10.5: Runtime performances of both traditional feature selection and prior knowledge approaches on the Alzheimer's disease data sets. a) shows performances on the AddNeuroMedI data set, b) shows performances on the AddNeuroMedII data set. Modifying and combining approaches use Open Targets (OT) as knowledge base, whereas network approaches apply PathwayCommons (PC).

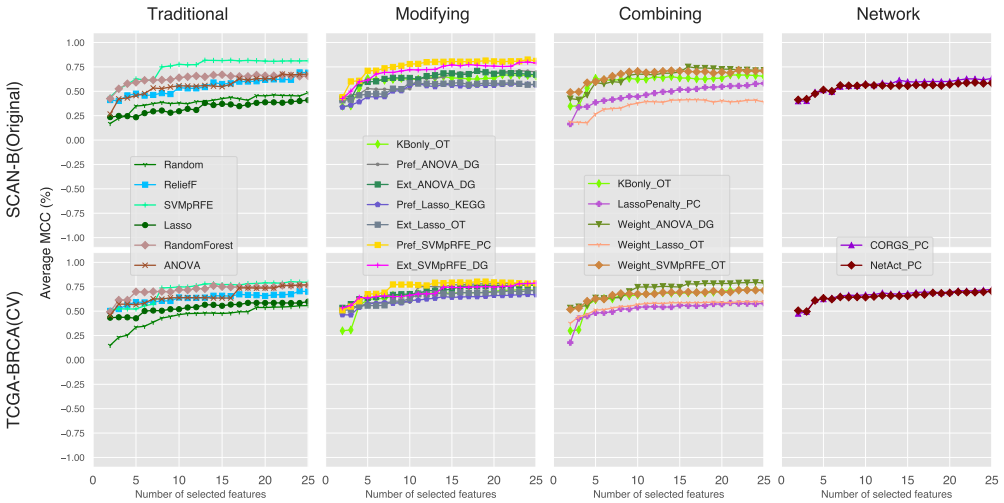


Fig. 10.6: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by the tested approaches on the SCAN-B data set, grouped into traditional, modifying, combining, and network approaches (from left to right). Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the TCGA-BRCA data set for cross-validation (CV). SVM-RFE (*SVMpRFE*) and its adaptations show highest MCC scores, network approaches perform at the lower end.

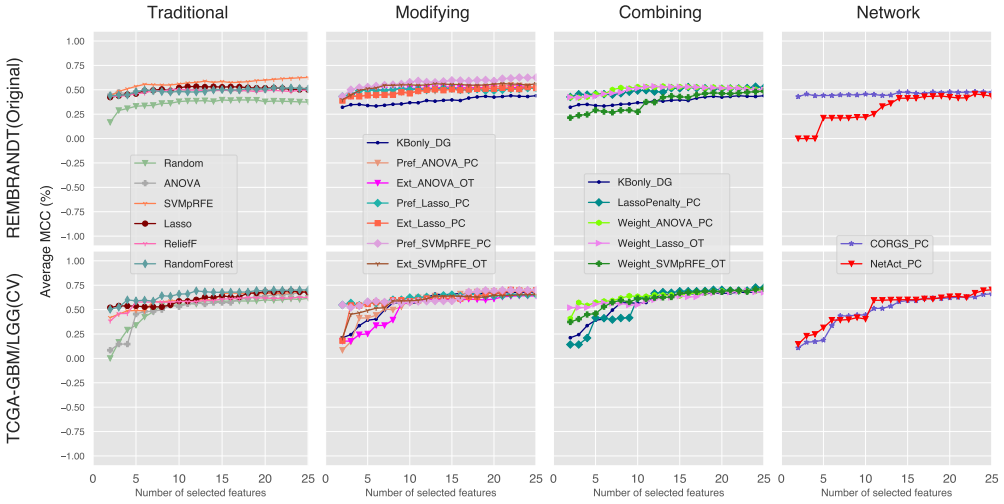


Fig. 10.7: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by the tested approaches on the REMBRANDT data set, grouped into traditional, modifying, combining, and network approaches (from left to right). Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the TCGA-GBM/LGG data set for cross-validation (CV). The tested approaches show similar MCC scores that only decrease slightly on the cross-validation data set.

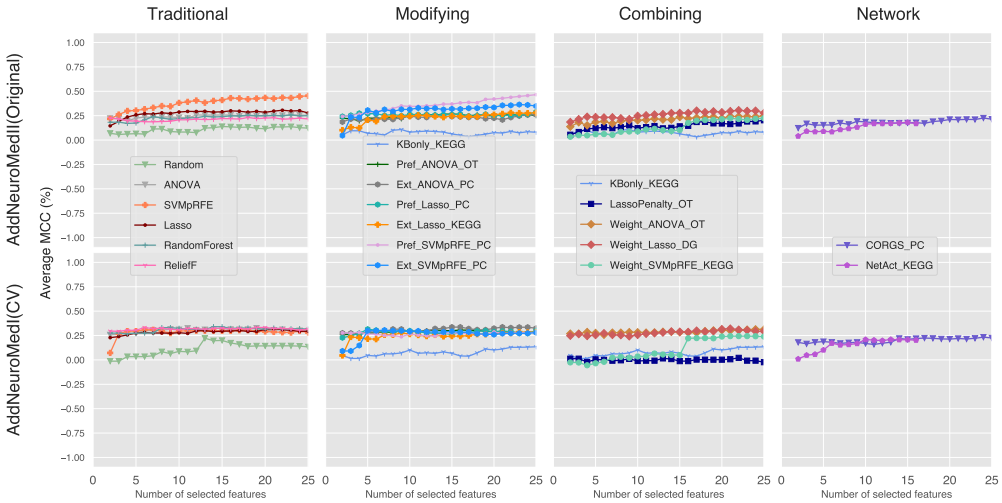
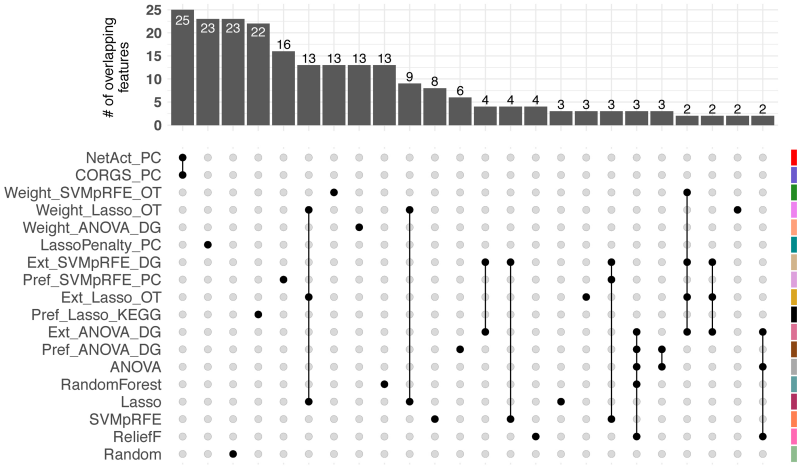
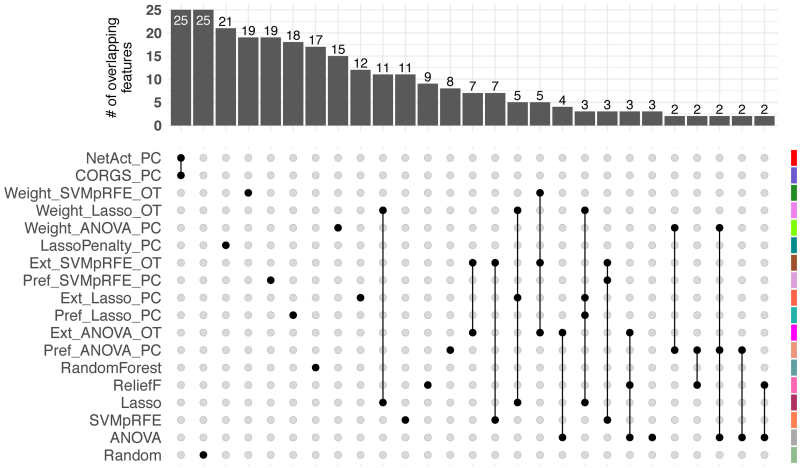


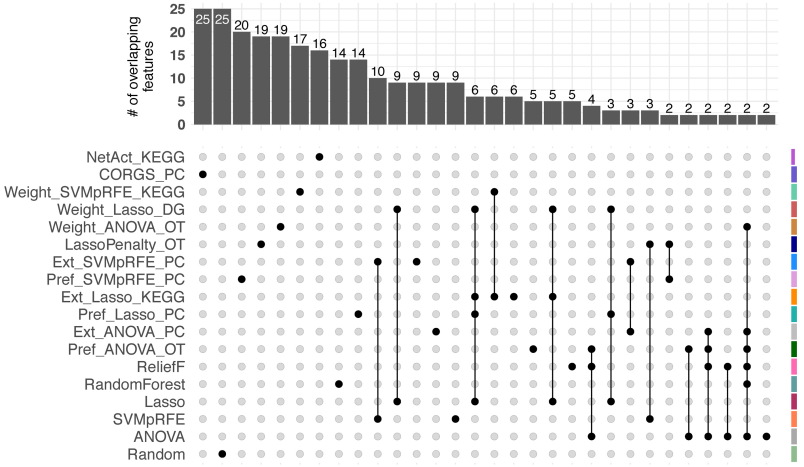
Fig. 10.8: Classification performances measured in Matthew's Correlation Coefficient (MCC) of feature sets selected by the tested approaches on the AddNeuroMedII data set, grouped into traditional, modifying, combining, and network approaches (from left to right). Upper row shows MCC scores of the feature sets used for classification on the original data set, lower rows show MCC scores of the same feature sets applied for classification on the AddNeuroMedI data set for cross-validation (CV). Again, SVM-RFE (*SVMpRFE*) and its prefiltering (*Pref*) and extending (*Ext*) adaptations perform best but cannot maintain these MCC scores on the cross-validation data set.



(a)



(b)



(c)

Fig. 10.9: Overlaps of feature sets selected with the same approaches as used for classification, for data sets a) SCAN-B, b) REMBRANDT, and c) AddNeuroMedII. We omit overlaps of single features. Often up to half of the features selected by the tested approaches are distinct, with fewest overlaps observed for SCAN-B.

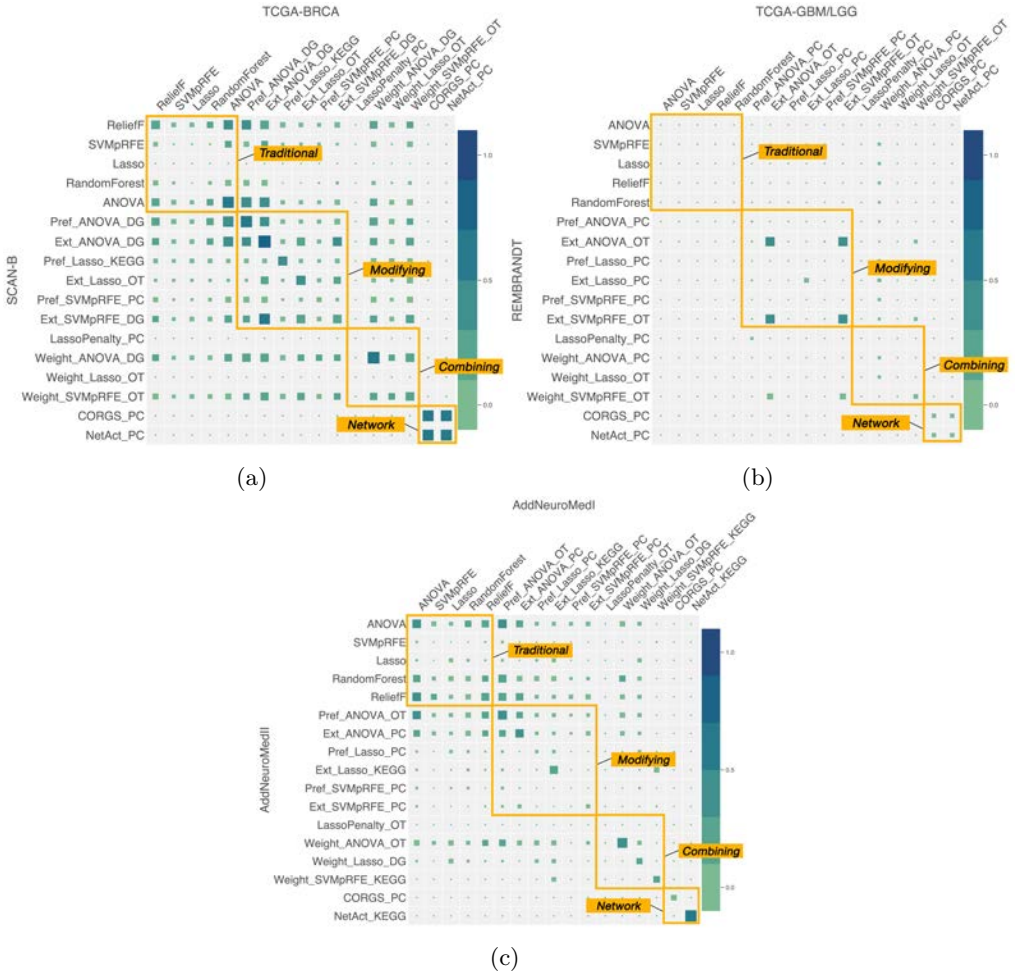


Fig. 10.10: Robustness across data sets of the same disease domain for feature sets selected by the best-classifying approaches from a) TCGA-BRCA, b) TCGA-GBM/LGG, and c) AddNeuroMedI. Robustness is measured by comparing feature rankings via ranked-biased (RBO) overlap. A high RBO score means high agreement of feature rankings across both data sets, and as such indicates a high robustness.

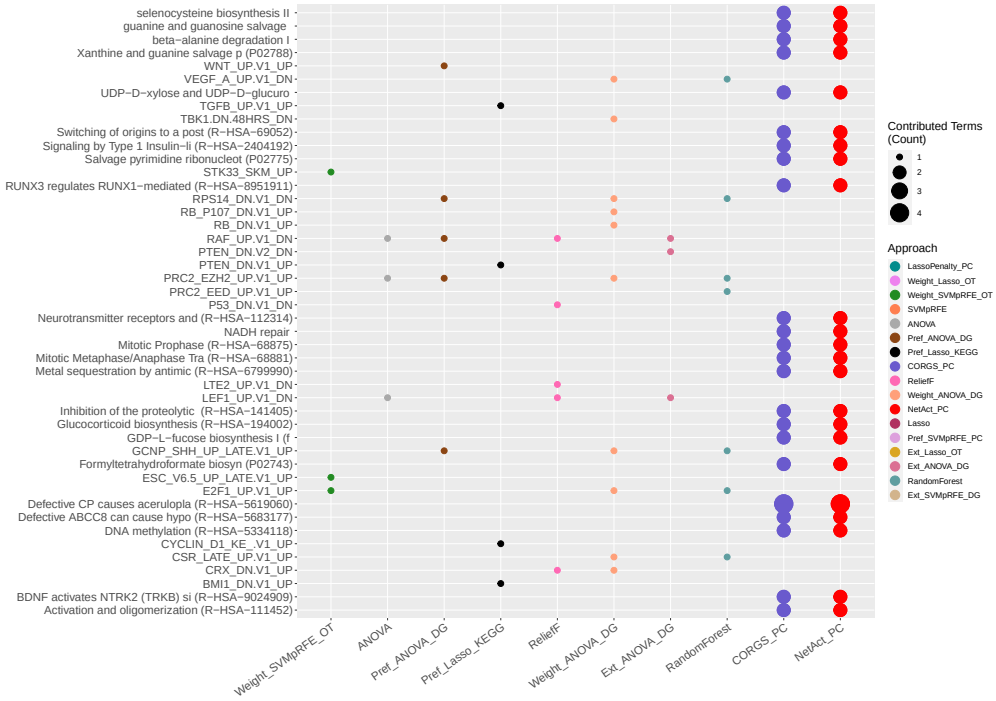


Fig. 10.11: Semantic similarities of enriched MSigDB oncogenic signatures (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the SCAN-B data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. Clusters are nearly identical for network approaches, whereas there is only few similarity between enrichments of the other approaches.

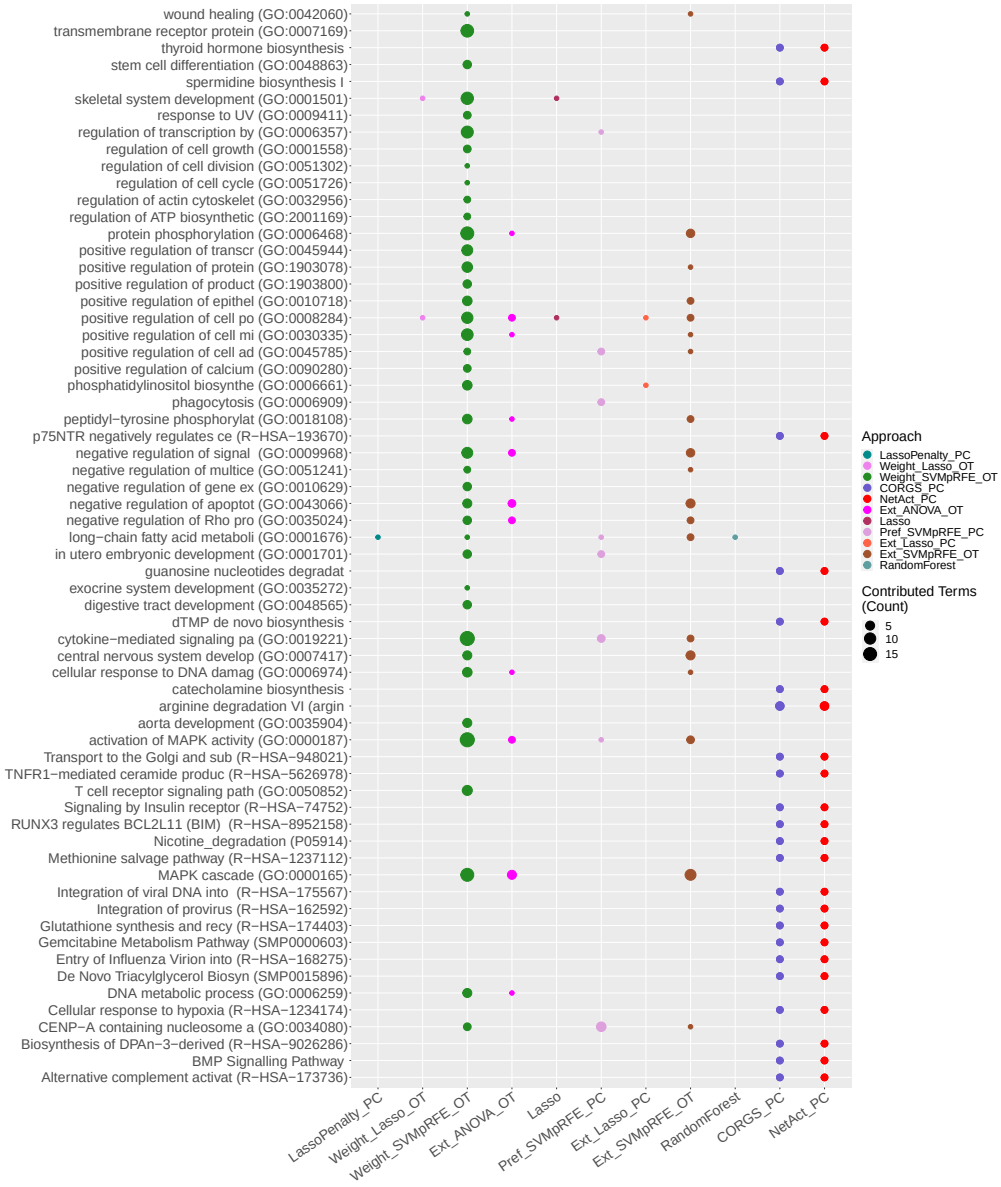


Fig. 10.12: Semantic similarities of enriched MSigDB oncogenic signatures (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the REMBRANDT data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. Clusters are nearly identical for network approaches, whereas there is only few similarity between enrichments of the other approaches. *Weight_SVMpRFE_OT* retrieves highest number of enrichments.

Fig. 10.13: Semantic similarities of enriched MSigDB oncogenic signatures (or pathways, for network approaches) of gene sets selected with the same approaches as used for classification on the AddNeuroMedII data set. We group enrichments into into semantic clusters, the name of the representative term is given on the left side. Sizes of the bubbles indicate with how many enriched terms (or pathways) an approach contributes to a cluster. There is a high overlap between *Ext_ANOVA_PC*, *Pref_ANOVA_OT*, *ANOVA*, *ReliefF*, and *RandomForest*.

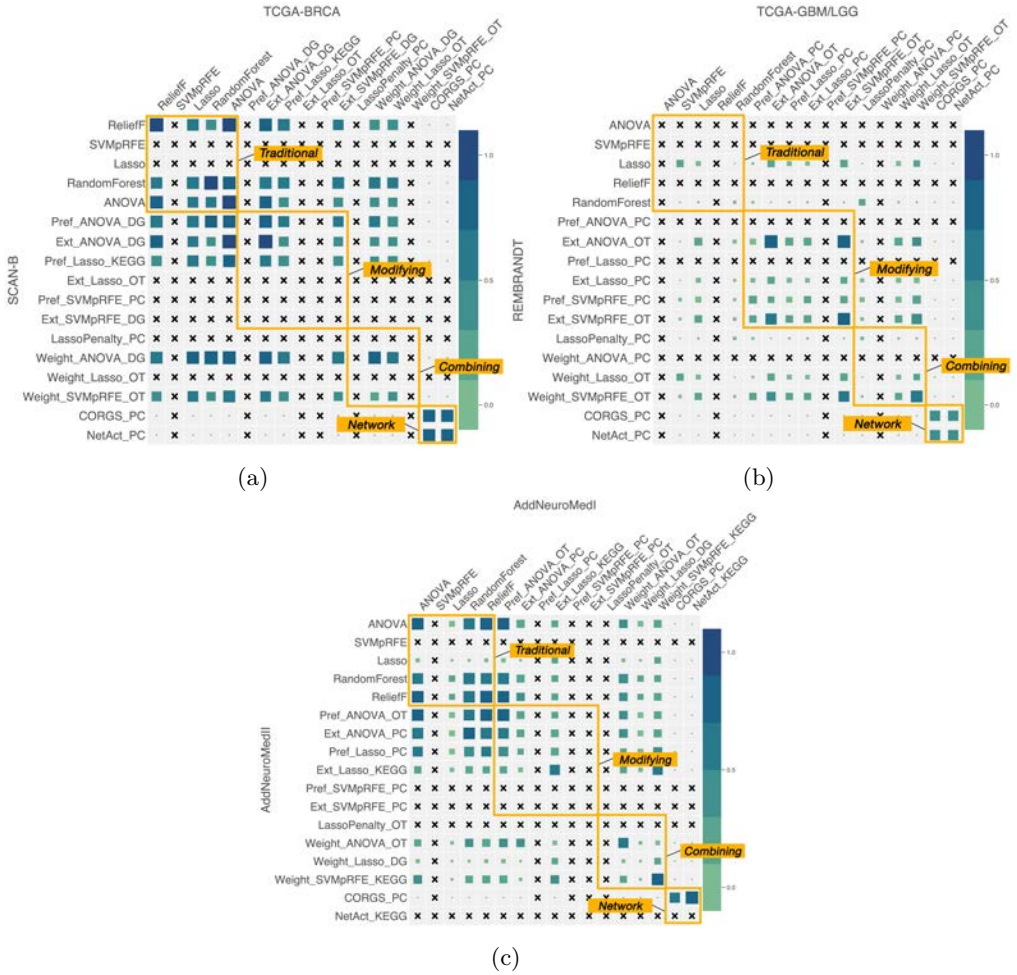


Fig. 10.14: Robustness across data sets of the same disease domain for enrichments (GO terms for Alzheimer's disease and glioma, MSigDB oncogenic signatures for breast cancer, pathways for network approaches) retrieved for feature sets selected by the tested approaches on a) SCAN-B, b) REMBRANDT, and c) AddNeuroMedII data sets. High scores mean a high semantic similarity of enrichments and as such indicate a higher robustness and biological relevance of enrichments.

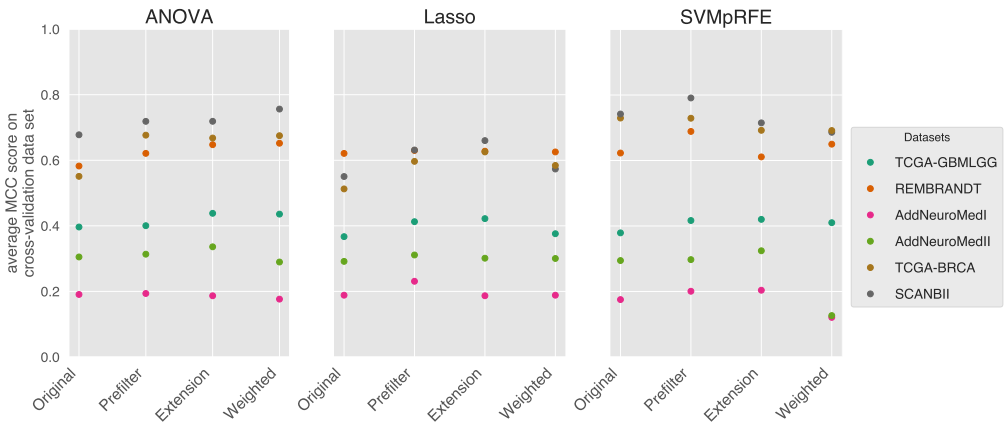


Fig. 10.15: Average classification performances (for 5 to 25 features, measured by average MCC scores) on all six data sets when used for cross-validation for ANOVA, Lasso, SVM-RFE and approaches combining them with prior knowledge in a prefiltering (*Pref*), extending (*Ext*), and weighting (*Weight*) manner, respectively. Especially prefiltering and extending approaches typically show increased average classification performance.

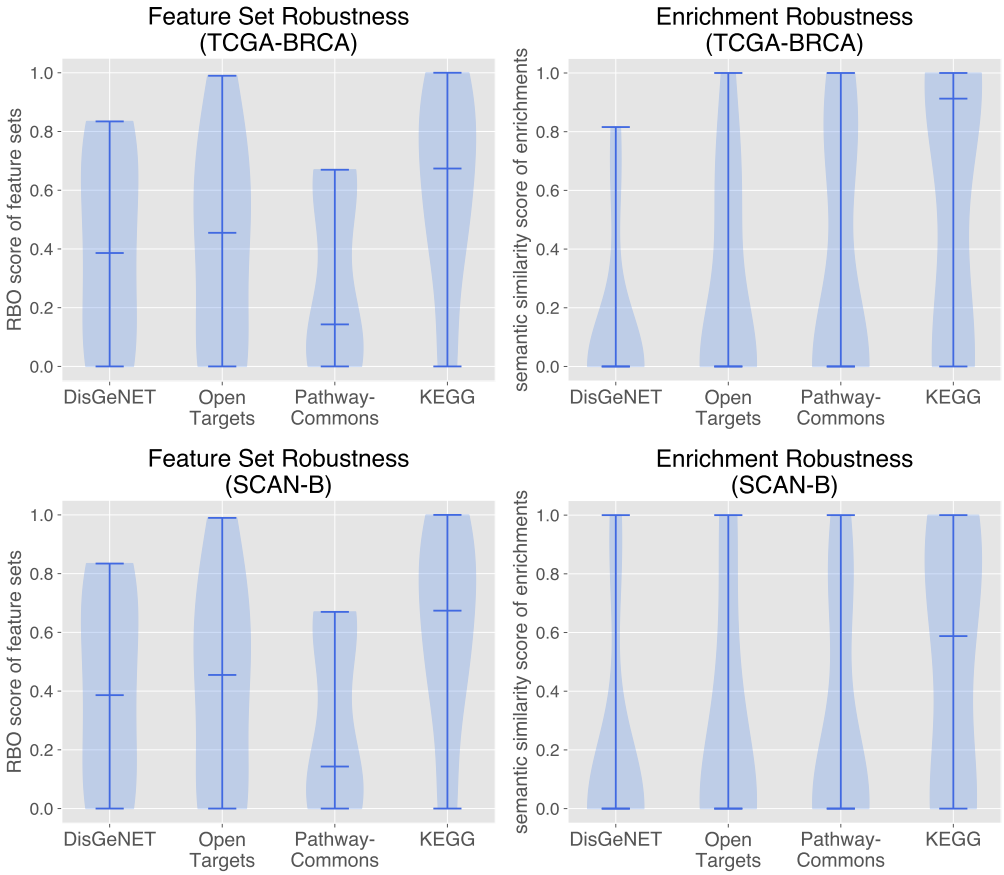


Fig. 10.16: Robustness of feature sets and enrichments on the TCGA-BRCA (upper row) and SCAN-B (lower row) data sets, grouped by knowledge base used. Feature set robustness (left side) is measured via RBO scores, enrichment robustness (right side) is measured in semantic similarity scores. Approaches using PathwayCommons retrieve less robust feature sets, whereas approaches using KEGG retrieve most robust feature sets and enrichments.

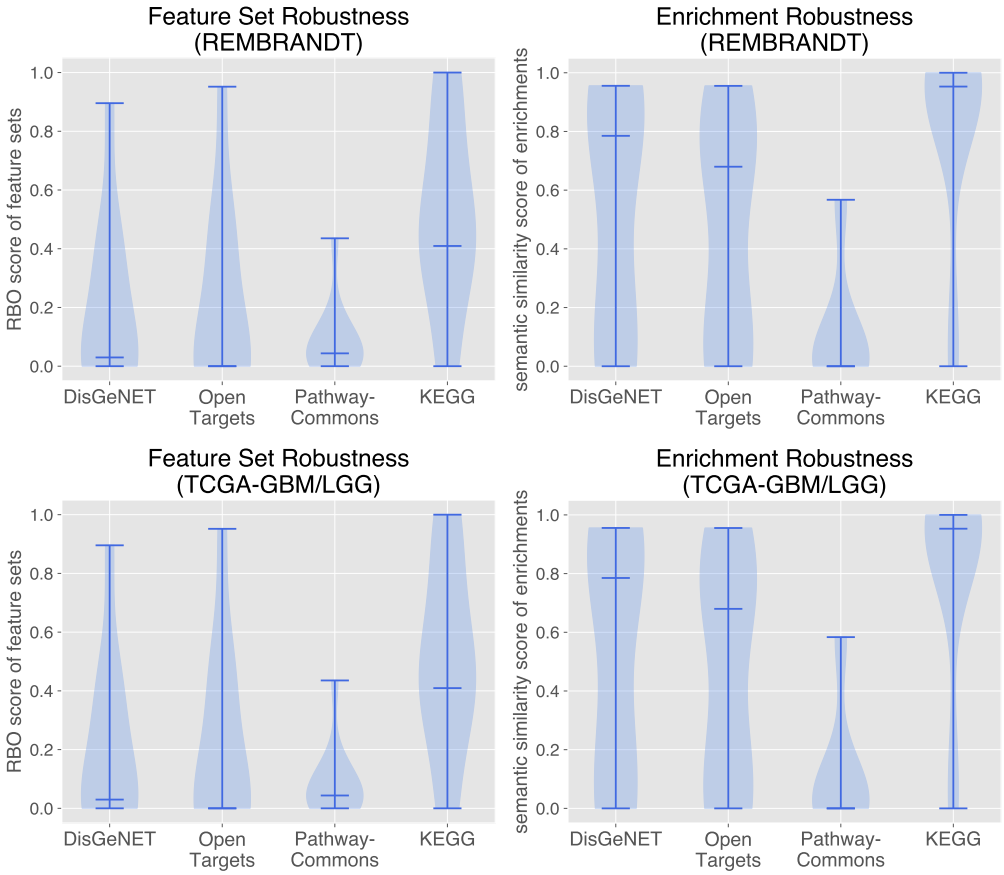


Fig. 10.17: Robustness of feature sets and enrichments on the REMBRANDT (upper row) and TCGA-GBM/LGG (lower row) data sets, grouped by knowledge base used. Feature set robustness (left side) is measured via Rank-biased overlap (RBO) scores, enrichment robustness (right side) is measured in semantic similarity scores. Except for PathwayCommons, which shows equally low robustness of features and enrichments, all other knowledge bases show an increased robustness of enrichments, even when feature set robustness is low.

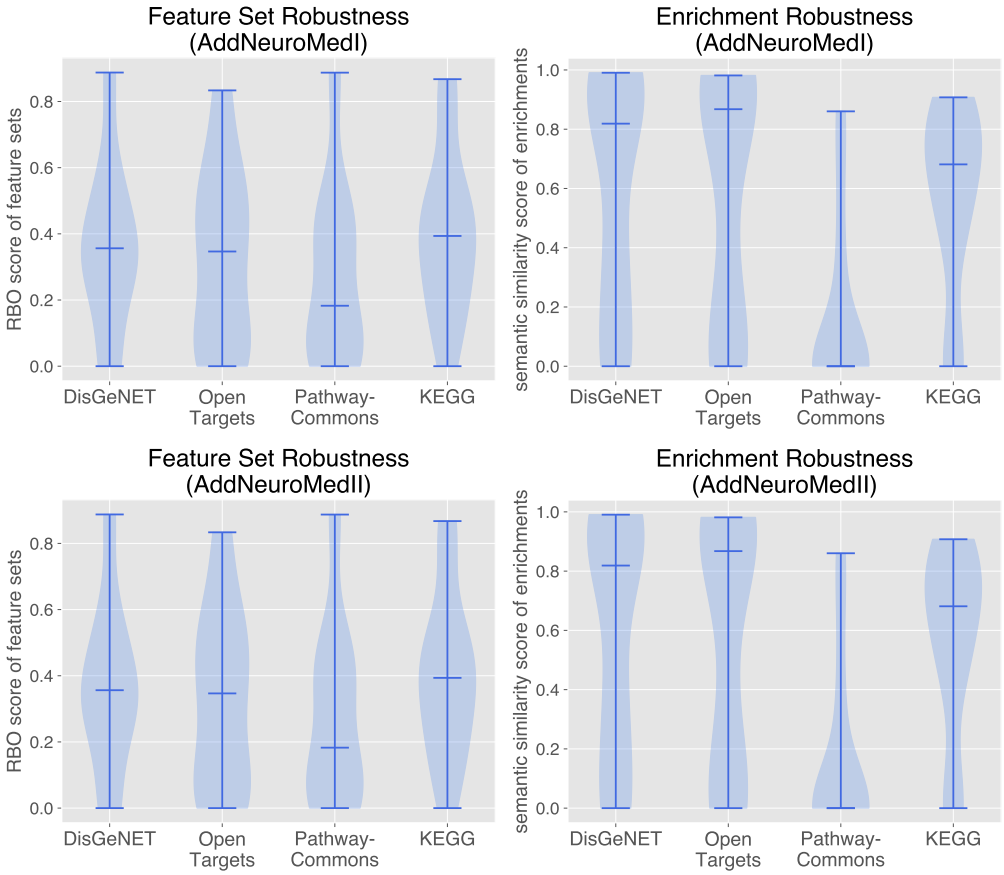


Fig. 10.18: Robustness of feature sets and enrichments on the AddNeuroMedI (upper row) and AddNeuroMedII (lower row) data sets, grouped by knowledge base used. Feature set robustness (left side) is measured via Rank-biased overlap (RBO) scores, enrichment robustness (right side) is measured in semantic similarity scores. Approaches using PathwayCommons retrieve less robust feature sets.

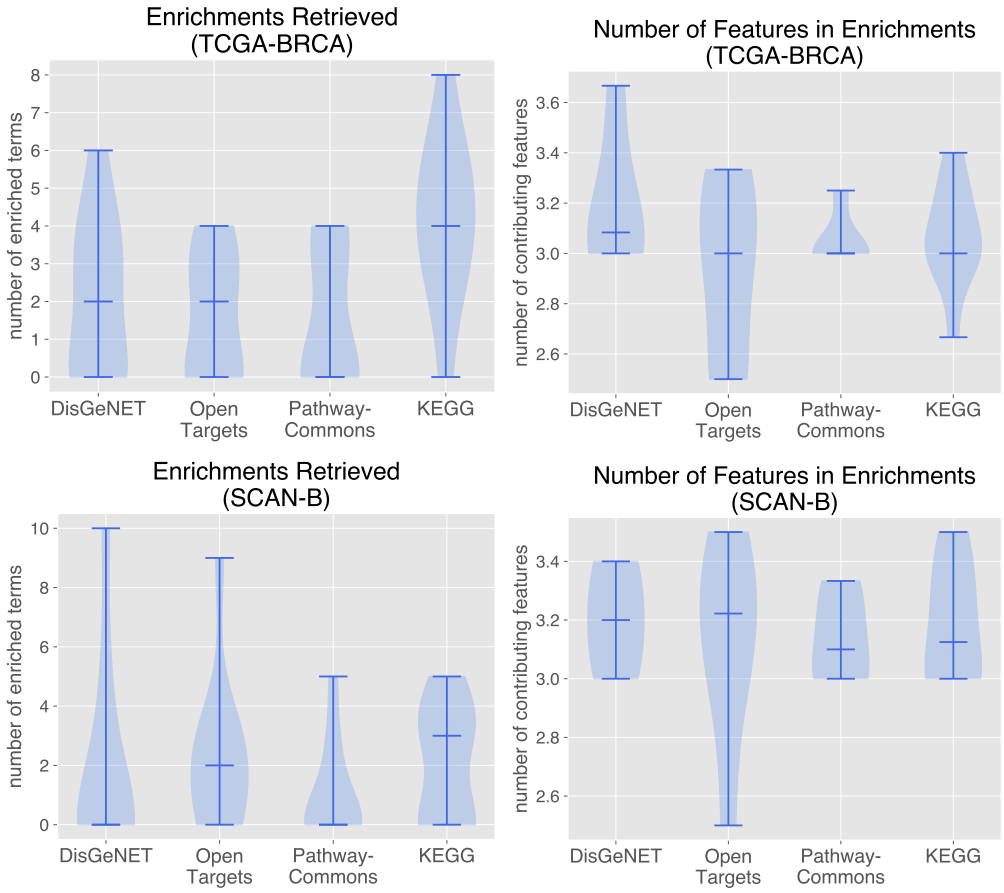


Fig. 10.19: Quantitative comparison of enrichments (MSigDB oncogenic signatures) retrieved for prior knowledge approaches using a particular knowledge base when applied to the TCGA-BRCA (upper row) and SCAN-B (lower row) data sets. Left side depicts overall number of enrichments retrieved (with median), right side depicts number of features involved in the enrichments.

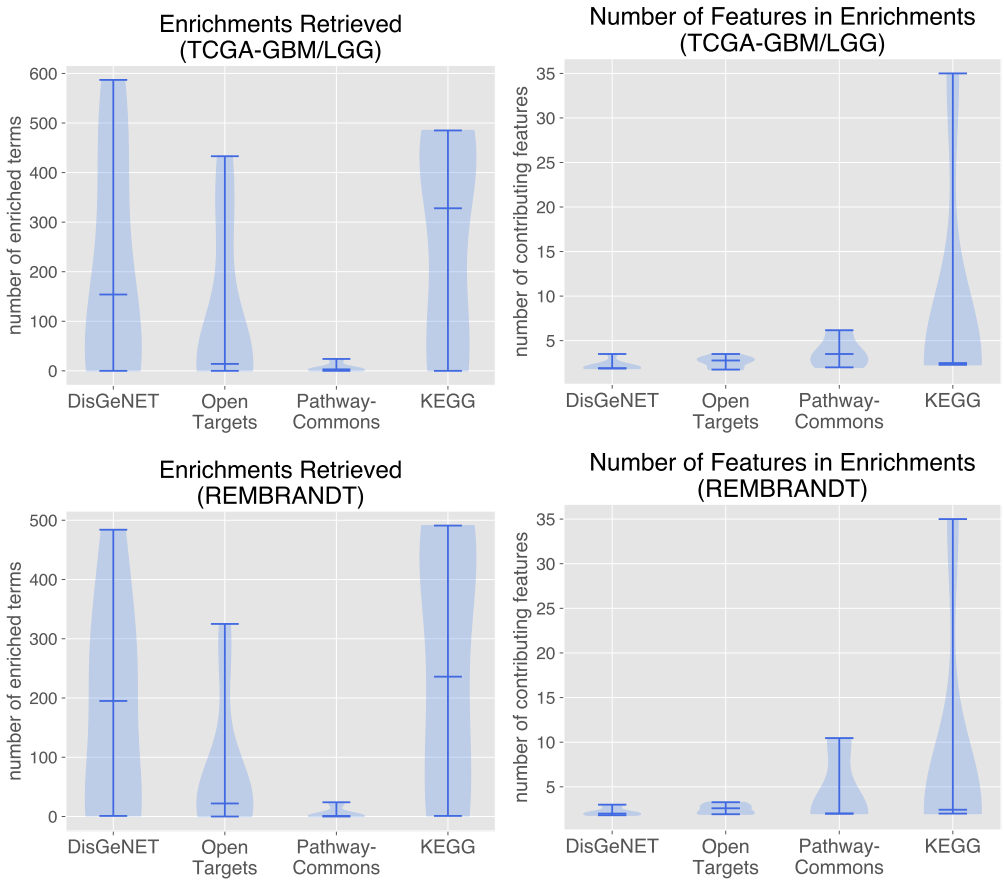


Fig. 10.20: Quantitative comparison of GO term enrichments retrieved for prior knowledge approaches using a particular knowledge base when applied to the TCGA-GBM/LGG (upper row) and REMBRANDT (lower row) data sets. Left side depicts overall number of enrichments retrieved (with median), right side depicts number of features involved in the enrichments.

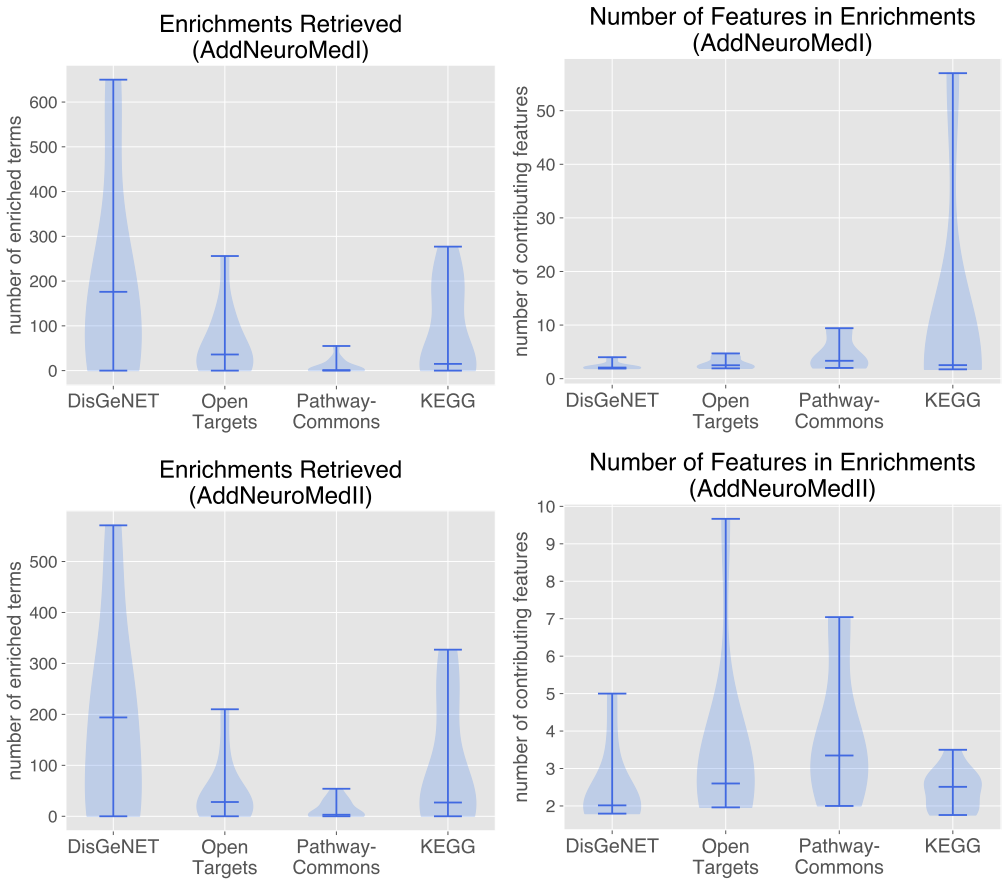


Fig. 10.21: Quantitative comparison of GO term enrichments retrieved for prior knowledge approaches using a particular knowledge base and applied to the AddNeuroMedI (upper row) and AddNeuroMedII (lower row) data sets. Left side depicts overall number of enrichments retrieved (with median), right side depicts number of features involved in the enrichments.

Subclass	Approach	Year	Data Source	Output		Use Case		
				ranked pred. genes	others	clust./ pred. class.	gene-dis. assoc.	biom. det.
filtering	Fang et al. [58]	2014	GO, KEGG	•		•		
	Jungjit et al. [98]	2014	KEGG	•		•		
	Perscheid et al. [159]	2018	DisGeNET, KEGG	•		•		
extending	BPFS [15]	2009	KEGG	•		•		
	SoFoCles [148]	2010	GO	•		•		
	RelSim [132]	2017	STRING	•			•	

Table 10.8: Overview on modifying approaches using prior knowledge and their respective characteristics. All approaches return ranked gene sets and most of them are used in a clustering or classification context. Abbreviations: predictions (pred.), clustering (clust.), classification (class.), gene-disease association (gene-dis. assoc.), biomarker detection (biom. det.).

Subclass	Approach	Year	Data Source	Output		Use Case		
				ranked pred. genes	others	clust./ pred. class.	gene-dis. assoc.	biom. det.
formal framework	Vert & Kanehisa [244]	2003	KEGG		linear features	•		
	Xu & Zhang [257]	2005	GO	•		•		
	Li & Li [117]	2008	KEGG		•		•	
	Srivastava et al. [207]	2008	GO	•		•		
	Zhu, Shen, Pan [265]	2009	KEGG		•		•	
	Zhao et al. [264]	2010	COSMIC, GO, KEGG	•		•		
	Stingo & Vannucci [210]	2010	KEGG	•		•		
	Stingo et al. [209]	2011	KEGG		genes and pathways		•	
	Gillies et al. [68]	2012	GO		•			•
	iBVS [153]	2013	KEGG		genes and pathways	•		•
	Jungjit et al. [98]	2014	KEGG	•		•		
	Raghu et al. [171]	2017	DisGeNET, KEGG	•		•		

Table 10.9: Overview on combining approaches incorporating prior knowledge via formal frameworks and their respective characteristics. GO and KEGG are predominantly used as external knowledge bases, and many approaches are used in a clustering or classification context. Abbreviations: predictions (pred.), clustering (clust.), classification (class.), gene-disease association (gene-dis. assoc.), biomarker detection (biom. det.).

Subclass	Approach	Year	Data Source	Output		Use Case		
				ranked pred. genes	others	clust./ pred. class.	gene-dis. assoc.	biom. det.
process-oriented	GeneRank [135]	2005	GO	•			•	
	CGI [127]	2006	MIPS [146]		gene-disease association scores			•
	Qi and Tang [169]	2007	GO	•		•	•	
	Aragues [8]	2008	BioGRID, GO, IntAct	•		•		
	Tseng & Yu [234]	2011	GO	•			•	
	Mitra et al. [133]	2012	GO	•		•		
	Swarnkar [217]	2014	HPRD [104], IntAct, MINT [119]		ranked gene sets			•
	Park et al. [149]	2014	I2D [29]		functionally related gene pairs		•	
	Swarnkar [218]	2015	HPRD, IntAct, MINT		hub genes		•	
	Acharya et al. [3]	2017	GO	•		•		
	Mahapatra et al. [131]	2018	HPRD, IntAct, MINT		hub genes	•		
	Acharya et al. [2]	2020	GO, HitPredict [123]	•		•		•
	PrognoSIT [18]	2021	MSigDB, PID [186]	•		•		
	Family Rank [184]	2021	STRING	•		•		

Table 10.10: Overview on process-oriented combining approaches using prior knowledge and their respective characteristics. GO and interaction knowledge bases are predominantly used, and the approaches are used in different contexts. Abbreviations: predictions (pred.), clustering (clust.), classification (class.), gene-disease association (gene-dis. assoc.), biomarker detection (biom. det.).

Approach	**Year**	**Data Source**	**Output**		**Use Case**		
			ranked pred. genes	others	clust./ pred. class.	gene-dis. assoc.	biom. det.
Guo et al. [76]	2005	GO		functional expression profiles	•		
Chuang et al. [41]	2007	BIND [13], HPRD, Reactome		subnetwork feature values		•	
Quanz et al. [170]	2008	KEGG		pathway feature values	•		
Lee et al. [113]	2008	MSigDB [118]		pathway activity score	•		
Chen & Wang [38]	2009	GO		"supergenes"	•		
Zhang [262]	2013	BIND, HPRD, IntAct, Reactome, and others		functional modules	•		
Gu et al. [74]	2014	KEGG		pathway activity scores	•		•
Alcaraz et al. [5]	2016	HPRD, HTRIdb [24], HumanNet [114], I2D		pathway enrichment scores	•		•

Table 10.11: Overview on network approaches for prior knowledge biomarker detection and their respective characteristics. All of them return some kind of extracted features, e.g. network modules. Most of them are applied in a clustering or classification context. Abbreviations: predictions (pred.), clustering (clust.), classification (class.), gene-disease association (gene-dis. assoc.), biomarker detection (biom. det.).

Table 10.12: Qualitative comparison of approaches for prior knowledge biomarker detection. (•) means that code/software might be available upon author request. Abbreviations used for performance measures: Predictive mean squared error (PMSE), mean squared error (MSE), Matthew's correlation coefficient (MCC), area under curve (AUC), receiver operating characteristic (ROC), false negative rate (FNR), positive predictive value (PPR), posterior inclusion probability (PIP), standard error (SE), standard squared errors (SSE), normalized root mean squared error (NRMSE).

Approach	**Applicability**		**Characteristics**				**Validation**			**Comparison**	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Vert & Kanehisa [244]		+ general purpose approach	− not addressed	+ outperforms traditional	− removes unmapped genes	+ via variability of neighbor genes	− none	+ ROC	+ holdout	•	
Guo et al. [76]		− specific for GO usage	+ functional modules as features	+ selects less features + robust to noise	− removes unmapped genes	+ via functional annotations	+ literature review	+ accuracy + sensitivity + precision	+ 5-fold	•	
GeneRank [135]	•	+ incorporates any network	− tends to include genes of same pathway	+ more robust to noise compared to random	− removes unmapped genes	+ higher weight for hub genes irresp. of expression value	+ expression change analysis	+ AUC	− none	•	
Xu & Zhang [257]		+ add-on for any traditional approach	− not addressed	+ selects less random genes − performs similar to trad. approaches	− removes unmapped genes	− not addressed	− none	+ sensitivity	− none	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
CGI [127]	(•)	− requires labeled data	− not addressed	+ better performance ± depends on knowledge base reliability	− removes unmapped genes	+ positive/negative gene-gene associations	+ known disease genes ± annotation (GO, input knowledge)	+ area under cumulative distribution function	− none	•	•
Chuang et al. [41]		− only works with labeled data	+ subnetworks as features	+ more robust across datasets + better performance	+ combines multiple networks − removes unmapped genes	+ via gene-gene interaction networks	+ genes from GO and MSigDB + cancer hallmark enrichment	+ AUC	+ 5-fold + across datasets	•	
Qi & Tang [169]		+ works with any trad. approach	+ selects representatives	+ better performance + selects less genes + robust to noise	− removes non-annotated genes	− not addressed	− none	+ accuracy	+ LOOCV	•	
Aragues et al. [8]		− specific for cancer genes	− not addressed	+ better results	+ combines multiple networks − removes unmapped genes	+ via gene-gene interaction networks	+ literature review + annotation (Reactome, UniProt, GO)	+ PPR + sensitivity	− none		

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Lee et al. [113]		− requires labeled data − requires predefined pathways	+ pathways as features	+ slightly more robust across datasets + slightly better accuracy	− removes unmapped genes	+ via pathway activity scores	+ literature review	+ accuracy + AUC	+ 5-fold + across datasets		•
Li & Li [117]		+ takes any network as input	− not addressed	+ better performance + selects more connected genes	− removes unmapped genes	− via network constraints	+ literature review	+ sensitivity + specificity + PMSE	+ 10-fold + across datasets	•	
Quanz et al. [170]		− requires predefined pathway features	+ pathways as features	+ large improvement for real-world data − low improvement for simulated data	− removes unmapped genes	+ via gene-gene interaction networks	− none	+ accuracy + sensitivity + specificity + MCC	+ 10-fold	•	
Srivastava et al. [207]		+ configurable weight for prior knowledge	+ groups similar genes	+ more biologically relevant genes	− not addressed	+ via functional similarities	+ literature review	− none	− none		

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
BPFS [15]	•	+ takes any network as input	+ selects genes far apart in pathway	− not very decisive	+ keeps unmapped genes via randomization	+ via pathway information	+ literature review	+ accuracy	+ holdout + across datasets	•	
Chen & Wang [38]		+ references to used R packages	+ selects *supergenes* from pathways	± only slightly better performance	+ keeps unmapped genes via co-expression	+ via pathways	+ literature review	+ AUC + ROC	− none	•	
Path-Boost [21]		+ can use any network information	+ boosts interaction partners of differentially expressed genes	± only slightly better than other approaches	+ keeps unmapped genes	+ via network information	− none	+ MSE + SE	+ holdout	•	•
Zhu et al. [265]		+ applicable to different use cases	+ groups similar genes	+ better performance + more clinically relevant genes	± possible via default weights	+ via network-based SVM	+ known disease genes	+ FNR + SE	+ holdout	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
RRFE [95]	•	+ can use any network as input		+ selected genes more consistent + more robust to noise + selects more biologically relevant genes	+ default ranks for unmapped genes	+ higher weight for hub genes irrespective of expression value	+ pathway enrichment (KEGG)	+ AUC + gene set comparisons	+ 10-fold	•	•
SoFoCles [148]	(•)	+ user interface + flexibly combines similarity measures / trad. approaches	± adds semantically similar genes	+ more biologically relevant genes	+ keeps non-annotated genes	− none	+ literature review	+ accuracy + sensitivity + FPR + F_1	+ LOOCV	•	
Stingo & Vannucci [210]		+ takes any network as input	− selects similar genes within network	+ detects discriminative genes	+ keeps unmapped genes	+ via MRF	+ literature review	+ PIP	+ holdout	•	
Zhao et al. [264]		+ classifies prior knowledge types ± formal guide to include prior knowledge	± depends on integrated knowledge bases	+ more biologically relevant genes − no performance improvement	+ every gene receives a rank	± depends on integrated knowledge base	+ literature review + known disease genes + similarities (GO)	+ accuracy	− none	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Stingo et al. [209]		+ takes any network as input	+ also selects pathways	+ selects less false positives + more biologically relevant pathways	− removes unmapped genes ± can be kept via default priors	+ via MRF + also inter-pathway dependencies	+ literature review	+ PIP	+ holdout		
Tseng & Yu [234]		+ combines multiple scoring functions + integrates multiple knowledge bases	− not addressed	+ better performance + selects less genes	− removes non-annotated genes	± depends on the applied scoring functions	+ decision tree analysis	+ accuracy	+ LOOCV	•	•
Gillies et al. [68]		+ pseudo-code provided	± adds genes involved in similar processes	+ better performance	+ keeps non-annotated genes via trad. approach	+ via semantic similarity	+ literature review	+ sensitivity + FPR + SSE	− none	•	
Mitra et al. [133]		− only works with GO + independent of disease domain	+ clusters genes	− not addressed	+ allows to identify new relationships	− not addressed	+ literature review	+ accuracy	+ 10-fold	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
iBVS [153]		− requires prefiltered gene sets − high runtime	− not addressed	+ better performance	− removes non-annotated genes	+ via MRF	+ literature review	+ PIP + ROC	+ holdout	•	
Zhang et al. [262]		− requires labeled data	+ detects network motifs	+ more robust to noise in dataset + more robust across datasets	+ combines multiple networks − removes unmapped genes	+ via gene-gene interaction networks	+ annotation (OMIM)	+ AUC	+ 5-fold + across datasets	•	
AdaL-net [216]		+ can use any connectivity information	± favors connected genes	+ more stable gene sets ± only slightly better results	− removes unmapped genes	+ via network information	+ literature review	+ MSE + sensitivity + specificity + MCC + gene set sizes + FP number	− none	•	•
Fang et al. [58]		− requires predefined pathways	+ selects pathway representative genes	+ slightly better performance + selects less genes	− removes non-annotated genes	+ via pathway information	+ literature review	+ accuracy	+ LOOCV	•	•

Table 10.12: continued

Approach	Applicability	Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Gu et al. [74]		+ pseudo code provided + references to R applied packages − requires predefined pathway features	+ addresses inner-pathway redundancy	+ slightly better performance + more biologically relevant genes	+ keeps unmapped genes via *unknown* pathway	+ via gene-gene interaction networks	+ literature review	+ accuracy	+ 5-fold	•	•
Jungjit et al. [98]		+ general purpose approach − prefiltered input	− not addressed	− slightly worse performance − no biological relevance	− removes non-annotated genes	− not addressed	− none	+ hemming loss + Wilcoxon signed rank	+ LOOCV	•	
Jungjit et al. [98]		+ general purpose approach − requires manually filtered pathways	+ selects less genes	+ more biologically relevant genes − better performance	− removes unmapped genes	− not addressed	+ literature review − annotation (KEGG = input knowledge)	+ hamming loss + Wilcoxon signed rank	+ LOOCV	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Park et al. [149]		+ general purpose approach − requires labeled data	− not addressed	+ robust for imbalanced classes + selects biologically relevant genes	− removes unmapped genes	− via gene correlations	+ literature review + annotation (GO)	+ accuracy + AUC	+ 10-fold	•	
Swarnkar et al. [217]		+ incorporates any network	+ via network approach	+ more biologically relevant genes	+ combines multiple networks − removes unmapped genes	+ via gene-gene interaction networks	+ known disease genes + annotation (GO)	+ accuracy	− none	•	
Swarnkar et al. [218]		+ incorporates any network	+ selects hub genes	+ selects less gene + slightly better performance	+ combines multiple networks	+ via gene-gene interaction network	+ literature review + annotation (GO, KEGG)	+ accuracy + sensitivity + specificity + precision + F_1 + MCC	− none		•
Alcaraz et al. [5]	•	+ incorporates any network + web service with user interface + low runtime	+ subnetworks as features	+ detects biologically relevant pathways	− removes unmapped genes	+ via gene-gene interaction networks	+ literature review + annotation (KEGG)	− none	− none	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Shao & Conrad [198]		+ can use any kind of network	+ via network motifs as features	+ better accuracy + more robust with network motifs as features	− removes unmapped genes	+ via interaction network	− none	+ AUC	+ 10-fold	•	
Acharya et al. [3]		− complex workflow of multiple tools	+ clusters similar genes	− selects more genes − worse performance	− removes unmapped genes	− not addressed	− none	+ accuracy + silhouette + Dunn index	+ 40-fold	•	
RelSim [132]	(•)	+ low computational complexity − requires labeled data	± adds semantically similar genes	+ more biologically relevant genes	+ keeps non-annotated genes via trad. approach	− none	+ literature review + annotation (GO, KEGG)	+ accuracy	+ LOOCV	•	
Raghu et al. [171]		+ pseudo-code provided − no details on DisGeNET usage	− not addressed	+ more biologically relevant genes − no performance improvement	± depends on definition of importance score	− not addressed	+ known gene sets + annotation (KEGG)	+ misclass. rate + F_1	+ 5-fold	•	

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Know-GRFF [75]		− prior knowledge = score	− not addressed	+ selects fewer and more stable features + reduced predictive error	− not addressed	− cannot incorporate interactions	+ literature review	+ Jaccard Index + TPR + FPR + MSE	− none	•	
Mahapatra et al. [131]		+ incorporates any network	+ extracts hub genes from subnetworks	+ more biologically relevant genes	+ combines multiple networks − removes unmapped genes	+ gene-gene interactions via network	+ gene set enrichment analysis + annotation (DAVID)	+ accuracy + sensitivity + specificity + precision + F_1 + MCC	+ 10-fold		•
Perscheid et al. [159]	•	+ combines knowledge bases / trad. approaches	− not addressed	+ better performance + selects less genes	− removes non-annotated genes	± depends on traditional approach	− none	+ accuracy	− none	•	
Acharya et al. [2]		+ requires unlabeled data	+ addressed via clustering	− not addressed	− remove non-annotated genes	− not addressed	± GO enrichment + literature review	+ F_1 + accuracy + specificity + sensitivity	− none	•	•

Table 10.12: continued

Approach	Applicability		Characteristics				Validation			Comparison	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
RLasso-Cox [247]	•	+ can use any network information	− not addressed	+ higher prognostic accuracy + fewer genes selected	− removes unmapped genes	− lower penalty for topologically important genes	− none	+ C-index + AUC	+ holdout + across datasets + AUC	•	•
xtune [261]	•	+ can use any and multiple kinds of prior knowledge	− not addressed	+ better accuracy results + fewer genes selected − can become unstable for > 50.000 features	+ default penalty terms for features w/o prior knowledge	+ can be encoded in penalty terms	− none	+ R^2 + AUC + number of selected features	+ 5-fold	•	
Prog-noSIT [18]	•	+ takes any set of pathways	− not addressed	+ more robust	− removes unmapped genes ± highly depends on input pathways	+ uses interaction data	+ literature review	+ NRMSE	+ 4-fold	•	

Table 10.12: continued

Approach	**Applicability**		**Characteristics**				**Validation**			**Comparison**	
	Availability	Flexibility	Redundancy	Robustness	Incompleteness	Interactions	Biological	Performance	Cross-Validation	Traditional	Prior Knowledge
Family Rank [185]	•	+ takes any network as input	+ addressed by grouping similar genes	− not addressed	− remove unmapped genes	+ uses interaction data	− none	+ AUC + differences in group medians/means	+ 10-fold	•	•

Bibliography

[1] F. Abbas-Aghababazadeh, Q. Li, and B. L. Fridley. "Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing". *PloS One* 13.10 (2018), e0206312.

[2] S. Acharya, L. Cui, and Y. Pan. "Multi-view feature selection for identifying gene markers: a diversified biological data driven approach". *BMC Bioinformatics* 21.18 (2020), pp. 1–31.

[3] S. Acharya, S. Saha, and N. Nikhil. "Unsupervised gene selection using biological knowledge: application in sample clustering". *BMC Bioinformatics* 18.1 (2017), p. 513.

[4] B. Afsari and E. J. Fertig. "Gene Set BenchMark". *R Package Version 1.14.0* 3 (2021).

[5] N. Alcaraz et al. "Robust de novo pathway enrichment with KeyPathwayMiner 5". *F1000Research* 5 (2016), p. 1531.

[6] S. Anders and W. Huber. "Differential expression analysis for sequence count data". *Nature Precedings* 11 (2010), R106.

[7] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13.5 (2015), pp. 971–989.

[8] R. Aragues, C. Sander, and B. Oliva. "Predicting cancer involvement of genes from heterogeneous data". *BMC Bioinformatics* 9 (2008), p. 172.

[9] B. Aranda et al. "PSICQUIC and PSISCORE: accessing and scoring molecular interactions". *Nature Methods* 8.7 (2011), pp. 528–529.

[10] M. Ashburner et al. "Gene Ontology: tool for the unification of biology". *Nature Genetics* 25.1 (2000), pp. 25–29. URL: `http://www.geneontology.org/`.

[11] J. Audoux et al. "SimBA: A methodology and tools for evaluating the performance of RNA-Seq bioinformatic pipelines". *BMC Bioinformatics* 18.1 (2017), pp. 1–14.

[12] F. Azuaje and O. Bodenreider. "Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study". In: *Proceedings of the*

Fourth IEEE Symposium on Bioinformatics and Bioengineering. IEEE. 2004, pp. 317–324.

[13] G. D. Bader, D. Betel, and C. W. Hogue. "BIND: the biomolecular interaction network database". *Nucleic Acids Research* 31.1 (2003), pp. 248–250.

[14] B. Baik, S. Yoon, and D. Nam. "Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data". *PloS One* 15.4 (2020), e0232271.

[15] N. Bandyopadhyay et al. "Pathway-based feature selection algorithm for cancer microarray data". *Advances in Bioinformatics* 2009 (2009).

[16] D. A. Barbie et al. "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1". *Nature* 462 (2009), pp. 108–112.

[17] T. Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". *Nucleic Acids Research* 41.D1 (2012), pp. D991–D995.

[18] A. B. Bektaş and M. Gönen. "PrognosiT: pathway/gene set-based tumour volume prediction using multiple kernel learning". *BMC Bioinformatics* 22.1 (2021), pp. 1–15.

[19] R. Bellazzi and B. Zupan. "Towards Knowledge-Based Gene Expression Data Mining". *Journal of Biomedical Informatics* 40.6 (2007), pp. 787–802.

[20] P. Bellot et al. "NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference". *BMC Bioinformatics* 16.1 (2015), p. 312.

[21] H. Binder and M. Schumacher. "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models". *BMC Bioinformatics* 10.1 (2009), pp. 1–11.

[22] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. "A review of feature selection methods on synthetic data". *Knowledge and Information Systems* 34 (2013), pp. 483–519.

[23] V. Bolón-Canedo et al. "A review of microarray datasets and applied feature selection methods". *Information Sciences* 282 (2014), pp. 111–135.

[24] L. A. Bovolenta, M. L. Acencio, and N. Lemke. "HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions". *BMC Genomics* 13.1 (2012), p. 405.

[25] L. Breiman. "Random forests". *Machine Learning* 45.1 (2001), pp. 5–32.

[26] F. P. Breitwieser, M. Pertea, A. V. Zimin, and S. L. Salzberg. "Human contamination in bacterial genomes has created thousands of spurious proteins". *Genome Research* 29.6 (2019), pp. 954–960.

[27] C. W. Brennan et al. "The somatic genomic landscape of glioblastoma". *Cell* 155.2 (2013), pp. 462–477.

[28] K. Breuer et al. "InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation". *Nucleic Acids Research* 41.D1 (2012), pp. D1228–D1233. URL: `https://www.innatedb.com`.

[29] K. R. Brown and I. Jurisica. "Unequal evolutionary conservation of human protein interactions in interologous networks". *Genome Biology* 8.5 (2007), R95.

[30] C. Brueffer et al. "Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network—breast initiative". *JCO Precision Oncology* 2 (2018), pp. 1–18.

[31] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". *BMC Bioinformatics* 11.1 (2010), pp. 1–13.

[32] M. Carlson. *UniProt.ws: R interface to UniProt web services*. R package version 2.26.0. 2018. URL: `http://www.bioconductor.org/packages/release/bioc/html/UniProt.ws.html`.

[33] E. G. Cerami et al. "Pathway Commons, a web resource for biological pathway data". *Nucleic Acids Research* 39.suppl_1 (2010), pp. D685–D690. URL: `http://www.pathwaycommons.org`.

[34] A. Chatr-Aryamontri et al. "The BioGRID interaction database: 2017 update". *Nucleic Acids Research* 45.D1 (2017), pp. D369–D379.

[35] E. Y. Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". *BMC Bioinformatics* 14.1 (2013), p. 128.

[36] J. Chen and B. Yuan. "Detecting functional modules in the Yeast protein–protein interaction network". *Bioinformatics* 22.18 (2006), pp. 2283–2290.

[37] J. Chen, N. Geard, J. Zobel, and K. Verspoor. "Automatic consistency assurance for literature-based Gene Ontology annotation". *BMC Bioinformatics* 22 (2021), p. 565.

[38] X. Chen and L. Wang. "Integrating biological knowledge with gene expression profiles for survival prediction of cancer". *Journal of Computational Biology* 16.2 (2009), pp. 265–278.

[39] M. Chiesa, G. I. Colombo, and L. Piacentini. "DaMiRseq—an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification". *Bioinformatics* 34.8 (2018), pp. 1416–1418.

[40] S. Chowdhury and R. R. Sarkar. "Comparison of human cell signaling pathway databases - evolution, drawbacks and challenges". *Database* 2015 (2015). bau126.

[41] H.-Y. Chuang et al. "Network-based classification of breast cancer metastasis". *Molecular Systems Biology* 3 (2007), p. 140.

[42] T. Cokelaer et al. "BioServices: a common Python package to access biological web services programmatically". *Bioinformatics* 29.24 (2013), pp. 3241–3242. URL: `http://bioservices.readthedocs.io/en/master/`.

[43] D. Croft et al. "The Reactome pathway knowledgebase". *Nucleic Acids Research* 42.D1 (2013), pp. D472–D477. URL: `https://reactome.org`.

[44] S. van Dam, T. Craig, and J. P. de Magalhaes. "GeneFriends: a human RNA-seq-based gene and transcript co-expression database". *Nucleic Acids Research* 43.D1 (2014), pp. D1124–D1132. URL: `http://genefriends.org`.

[45] M. Dash and H. Liu. "Feature selection for classification". *Intelligent Data Analysis* 1.3 (1997), pp. 131–156.

[46] A. P. Davis et al. "The comparative toxicogenomics database: update 2017". *Nucleic Acids Research* 45.D1 (2016), pp. D972–D978. URL: `https://ctdbase.org`.

[47] E. Demir et al. "The BioPAX community standard for pathway data sharing". *Nature Biotechnology* 28.9 (2010), pp. 935–942.

[48] D. Dernoncourt, B. Hanczar, and J.-D. Zucker. "Analysis of feature selection stability on high dimension and small sample data". *Computational Statistics & Data Analysis* 71 (2014), pp. 681–693.

[49] C. E. Determan Jr. "OmicsMarkeR". *R Package Version 1.19.0* (2017).

[50] R. Díaz-Uriarte and S. A. De Andres. "Gene selection and classification of microarray data using random forest". *BMC Bioinformatics* 7.1 (2006), p. 3.

[51] M.-A. Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". *Briefings in Bioinformatics* 14.6 (2013), pp. 671–683.

[52] L. Ein-Dor, O. Zuk, and E. Domany. "Thousands of Samples are Needed to Generate a Robust Gene List for Predicting Outcome in Cancer". *Proceedings of the National Academy of Sciences* 103.15 (2006), pp. 5923–5928.

[53] L. Ein-Dor et al. "Outcome Signature Genes in Breast Cancer: Is there a Unique Set?" *Bioinformatics* 21.2 (2004), pp. 171–178.

[54] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. "Cluster Analysis and Display of Genome-Wide Expression Patterns". *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868.

[55] C. Fähnrich, M.-P. Schapranow, and H. Plattner. "Towards integrating the detection of genetic variants into an in-memory database". In: *IEEE International Conference on Big Data*. IEEE. 2014, pp. 27–32.

[56] C. Fähnrich, M.-P. Schapranow, and H. Plattner. "Facing the genome data deluge: efficiently identifying genetic variants with in-memory database technology". In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. 2015, pp. 18–25.

[57] C. Fähnrich et al. "Surveillance and Outbreak Response Management System (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa". *Eurosurveillance* 20.12 (2015), p. 21071.

[58] O. H. Fang, N. Mustapha, and M. N. Sulaiman. "An integrative gene selection with association analysis for microarray data classification". *Intelligent Data Analysis* 18.4 (2014), pp. 739–758.

[59] A. Finney and M. Hucka. "Systems biology markup language: level 2 and beyond". *Biochemical Society Transactions* 31 (2003), pp. 1472–1473.

[60] S. A. Forbes et al. "COSMIC: somatic cancer genetics at high-resolution". *Nucleic Acids Research* 45.D1 (2016), pp. D777–D783. URL: `https://cancer.sanger.ac.uk/cosmic`.

[61] P. A. Futreal et al. “BRCA1 mutations in primary breast and ovarian carcinomas”. *Science* 266.5182 (1994), pp. 120–122.

[62] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design patterns: elements of reusable object-oriented software*. Vol. 99. Wokingham, UK: Addison-Wesley Reading, 1995.

[63] X. Ge et al. “Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues”. *Genomics* 86.2 (2005), pp. 127–141.

[64] L. Geistlinger et al. “Toward a gold standard for benchmarking gene set enrichment analysis”. *Briefings in Bioinformatics* 22 (1 2021), pp. 545–556.

[65] P.-L. Germain, A. Sonrel, and M. D. Robinson. “pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single-cell RNA-seq preprocessing tools”. *Genome Biology* 21 (2020), p. 227.

[66] P.-L. Germain et al. “RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods”. *Nucleic Acids Research* 44.11 (2016), pp. 5054–5067.

[67] J. Giles. “Internet encyclopaedias go head to head”. *Nature* 438 (2005), pp. 900–901.

[68] C. E. Gillies et al. “Improved feature selection by incorporating gene similarity into the LASSO”. *International Journal of Knowledge Discovery in Bioinformatics* 3.1 (2012), pp. 1–22.

[69] W. W. B. Goh and L. Wong. “Integrating networks and proteomics: moving forward”. *Trends in Biotechnology* 34.12 (2016), pp. 951–959.

[70] B. Grasnick, C. Perscheid, and M. Uflacker. “A Framework for the Automatic Combination and Evaluation of Gene Selection Methods”. In: *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Ed. by F. Fdez-Riverola et al. Springer. Cham: Springer International Publishing, 2019, pp. 166–174.

[71] R. L. Grossman et al. “Toward a shared vision for cancer genomic data”. *New England Journal of Medicine* 375.12 (2016), pp. 1109–1112.

[72] F.-N. B. W. Group. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Silver Spring, CO, and Bethesda, MD, USA: National Institute of Health, Food and Drug Administration (US), 2016. URL: `www.ncbi.nlm.nih.gov/books/NBK326791/` (visited on 05/17/2021).

[73] M. Gruenpeter et al. “M2.15 Assessment report on ’FAIRness of software’”. Version 1.1. *Zenodo* (2020). URL: `https://doi.org/10.5281/zenodo.4095092`.

[74] J.-L. Gu, Y. Lu, C. Liu, and H. Lu. “Multiclass classification of sarcomas using pathway based feature selection method”. *Journal of Theoretical Biology* 362 (2014), pp. 3–8.

[75] X. Guan and L. Liu. “Know-GRRF: domain-knowledge informed biomarker discovery with random forests”. In: *International Conference on Bioinformatics and Biomedical Engineering*. Springer. 2018, pp. 3–14.

[76] Z. Guo et al. "Towards precise classification of cancers based on robust gene functional expression profiles". *BMC Bioinformatics* 6 (2005), p. 58.

[77] Y. Gusev et al. "The REMBRANDT study, a large collection of genomic data from brain cancer patients". *Scientific Data* 5.1 (2018), pp. 1–9.

[78] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines". *Machine Learning* 46.1-3 (2002), pp. 389–422.

[79] M. Hall et al. "The WEKA data mining software: an update". *ACM SIGKDD Explorations Newsletter* 11.1 (2009), pp. 10–18.

[80] J. Han et al. "Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma". *Scientific Reports* 8.1 (2018), pp. 1–11.

[81] T. Hart et al. "Finding the active genes in deep RNA-seq gene expression studies". *BMC Genomics* 14.1 (2013), pp. 1–7.

[82] A.-C. Haury, P. Gestraud, and J.-P. Vert. "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures". *PloS One* 6.12 (2011), e28210.

[83] W. A. Haynes, A. Tomczak, and P. Khatri. "Gene annotation bias impedes biomedical research". *Scientific Reports* 8.1 (2018), pp. 1–7.

[84] K. Herbst, C. Fähnrich, M. L. Neves, and M.-P. Schapranow. "Applying In-Memory Technology for Automatic Template Filling in the Clinical Domain". In: *CLEF Evaluation Labs and Workshop Online Working Notes*. 2014, pp. 91–102.

[85] H. Hermjakob et al. "IntAct: an open source molecular interaction database". *Nucleic Acids Research* 32.suppl_1 (2004), pp. D452–D455. URL: `https://www.ebi.ac.uk/intact/`.

[86] Z. M. Hira and D. F. Gillies. "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data". *Advances in Bioinformatics* 2015 (2015).

[87] K. L. Howe et al. "Ensembl 2021". *Nucleic Acids Research* 49.D1 (2020), pp. D884–D891.

[88] J. Hua, W. D. Tembe, and E. R. Dougherty. "Performance of Feature-Selection Methods in the Classification of High-Dimension Data". *Pattern Recognition* 42.3 (2009), pp. 409–424.

[89] D. W. Huang, B. T. Sherman, and R. A. a. Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources". *Nature Protocols* 4.1 (2009), p. 44.

[90] F. W. Huang et al. "Highly recurrent TERT promoter mutations in human melanoma". *Science* 339.6122 (2013), pp. 957–959.

[91] E. Hubbell, W.-M. Liu, and R. Mei. "Robust estimators for expression analysis". *Bioinformatics* 18.12 (2002), pp. 1585–1592.

[92] K. Ickstadt, M. Schäfer, and M. Zucknick. "Toward integrative Bayesian analysis in molecular biology". *Annual Review of Statistics and Its Application* 5 (2018), pp. 141–167.

[93] M. P. van Iersel et al. “Presenting and exploring biological pathways with PathVisio”. *BMC Bioinformatics* 9.1 (2008), pp. 1–9.

[94] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza. “Filter versus wrapper gene selection approaches in DNA microarray domains”. *Artificial Intelligence in Medicine* 31.2 (2004), pp. 91–103.

[95] M. Johannes et al. “Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients”. *Bioinformatics* 26.17 (2010), pp. 2136–2144.

[96] W. E. Johnson, C. Li, and A. Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. *Biostatistics* 8.1 (2007), pp. 118–127.

[97] C. E. Jones, A. L. Brown, and U. Baumann. “Estimating the annotation error rate of curated GO database sequence annotations”. *BMC Bioinformatics* 8.1 (2007), pp. 1–9.

[98] S. Jungjit, A. A. Freitas, M. Michaelis, and J. Cinatl. “Extending multi-label feature selection with KEGG pathway information for microarray data analysis”. In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE. 2014, pp. 1–8.

[99] A. Kamburov et al. “ConsensusPathDB: toward a more complete picture of cell biology”. *Nucleic Acids Research* 39.suppl_1 (2010), pp. D712–D717. URL: `https://www.consensuspathdb.org`.

[100] Z. Kan et al. “Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures”. *Nature Communications* 9.1 (2018), pp. 1–13.

[101] M. Kanehisa and S. Goto. “KEGG: Kyoto encyclopedia of genes and genomes”. *Nucleic Acids Research* 28.1 (2000), pp. 27–30. URL: `http://kegg.jp/`.

[102] Kanehisa Laboratories. *KEGG Markup Language*. 2016. URL: `https://www.genome.jp/kegg/xml/docs/` (visited on 02/25/2022).

[103] S. Kerrien et al. “Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions”. *BMC Biology* 5.1 (2007), pp. 1–11.

[104] T. Keshava Prasad et al. “Human protein reference database – 2009 update”. *Nucleic Acids Research* 37.suppl_1 (2008), pp. D767–D772.

[105] P. K. Kimes and A. Reyes. “Reproducible and replicable comparisons using SummarizedBenchmark”. *Bioinformatics* 35.1 (2019), pp. 137–139.

[106] I. Kononenko. “Estimating attributes: analysis and extensions of RELIEF”. In: *European Conference on Machine Learning*. Ed. by F. Bergadano and L. De Raedt. Springer. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 171–182.

[107] H. Koo et al. “Ethnic delineation of primary glioblastoma genome”. *Cancer Medicine* 9.19 (2020), pp. 7352–7359.

[108] G. Koscielny et al. “Open Targets: a platform for therapeutic target identification and validation”. *Nucleic Acids Research* 45.D1 (2016), pp. D985–D994. URL: `http://targetvalidation.org`.

[109] B. Kramarz et al. “Improving the Gene Ontology resource to facilitate more informative analysis and interpretation of Alzheimer’s disease data”. *Genes* 9.12 (2018), p. 593.

[110] B. W. Kunkle et al. “Novel Alzheimer disease risk loci and pathways in African American individuals using the African genome resources panel: a meta-analysis”. *JAMA Neurology* 78.1 (2021), pp. 102–113.

[111] J. M. Lancaster et al. “BRCA2 mutations in primary breast and ovarian cancers”. *Nature Genetics* 13.2 (1996), pp. 238–240.

[112] C. Lazar et al. “A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.4 (2012), pp. 1106–1119.

[113] E. Lee et al. “Inferring Pathway Activity Toward Precise Disease Classification”. *PLoS Computational Biology* 4.11 (2008), e1000217.

[114] I. Lee et al. “Prioritizing candidate disease genes by network-based boosting of genome-wide association data”. *Genome Research* 21.7 (2011), pp. 1109–1121.

[115] J. T. Leek and J. D. Storey. “Capturing heterogeneity in gene expression studies by surrogate variable analysis”. *PLoS Genetics* 3.9 (2007), e161.

[116] Y. Leung and Y. Hung. “A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7.1 (2010), pp. 108–117.

[117] C. Li and H. Li. “Network-constrained regularization and variable selection for analysis of genomic data”. *Bioinformatics* 24.9 (2008), pp. 1175–1182.

[118] A. Liberzon et al. “The Molecular Signatures Database Hallmark Gene Set Collection”. *Cell Systems* 1.6 (2015), pp. 417–425.

[119] L. Licata et al. “MINT, the molecular interaction database: 2012 update”. *Nucleic Acids Research* 40.D1 (2012), pp. D857–D861.

[120] D. Lin. “An information-theoretic definition of similarity”. In: *Proceedings of the 15th International Conference on Machine Learning*. Vol. 98. San Francisco, CA, USA: Morgan Kaufmann, 1998, pp. 296–304.

[121] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe. “Model-based variance-stabilizing transformation for Illumina microarray data”. *Nucleic Acids Research* 36.2 (2008), e11–e11.

[122] H. Liu, L. Liu, and H. Zhang. “Ensemble gene selection by grouping for microarray data classification”. *Journal of Biomedical Informatics* 43.1 (2010), pp. 81–87.

[123] Y. López, K. Nakai, and A. Patil. “HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species”. *Database* 2015 (2015).

[124] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology* 15.12 (2014), p. 550.

[125] M. Lualdi and M. Fasano. "Statistical analysis of proteomics data: a review on feature selection". *Journal of Proteomics* 198 (2019), pp. 18–26.

[126] X. Ma, P. Sun, and Z.-Y. Zhang. "An integrative framework for protein interaction network and methylation data to discover epigenetic modules". *IEEE/ACM transactions on Computational Biology and Bioinformatics* 16.6 (2018), pp. 1855–1866.

[127] X. Ma, H. Lee, L. Wang, and F. Sun. "CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data". *Bioinformatics* 23.2 (2006), pp. 215–221.

[128] J. MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". *Nucleic Acids Research* 45.D1 (2016), pp. D896–D901. URL: `https://www.ebi.ac.uk/gwas/`.

[129] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. "Entrez Gene: gene-centered information at NCBI". *Nucleic Acids Research* 33.suppl_1 (2005), pp. D54–D58. URL: `https://www.ncbi.nlm.nih.gov/gene`.

[130] S. Mahajan and S. Singh. "Review on Feature Selection Approaches Using Gene Expression Data". *Imperial Journal of Interdisciplinary Research* 2.3 (2016), pp. 356–364.

[131] S. Mahapatra, B. N. Mandal, and T. Swarnkar. "Biological networks integration based on dense module identification for gene prioritization from microarray data". *Gene Reports* 12 (2018), pp. 276–288.

[132] P. Maji, E. Shah, and S. Paul. "RelSim: an integrated method to identify disease genes using gene expression profiles and PPIN based similarity measure". *Information Sciences* 384 (2017), pp. 110–125.

[133] S. Mitra and S. Ghosh. "Feature selection and clustering of gene expression profiles using biological knowledge". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6 (2012), pp. 1590–1599.

[134] G.-L. Moldovan, B. Pfander, and S. Jentsch. "PCNA, the maestro of the replication fork". *Cell* 129.4 (2007), pp. 665–679.

[135] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. "GeneRank: using search engine technology for the analysis of microarray experiments". *BMC Bioinformatics* 6.1 (2005), p. 233.

[136] A. Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* 5.7 (2008), pp. 621–628.

[137] P. A. Mundra and J. C. Rajapakse. "SVM-RFE with MRMR filter for gene selection". *IEEE Transactions on NanoBioscience* 9.1 (2010), pp. 31–37.

[138] Z. Mungloo-Dilmohamud, Y. Jaufeerally-Fakim, and C. Peña-Reyes. "Exploring the stability of feature selection methods across a palette of gene expression datasets". In: *Proceedings of the 6th International Conference on Biomedical and Bioinformatics Engineering*. ICBBE '19. Shanghai, China: Association for Computing Machinery, 2019, pp. 7–12.

[139] P. S. Nair and M. Vihinen. "VariBench: a benchmark database for variations". *Human Mutation* 34.1 (2013), pp. 42–49.

[140] National Cancer Institute (NCI). *NCI Metathesaurus*. 2021. URL: `https://ncim.nci.nih.gov/ncimbrowser/` (visited on 08/21/2021).

[141] National Research Council. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. Washington, DC, USA: National Academies Press, 2011. URL: `https://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research`.

[142] D. Ochoa et al. "Open Targets Platform: supporting systematic drug–target identification and prioritisation". *Nucleic Acids Research* 49.D1 (2021), pp. D1302–D1310.

[143] Y. Okamura et al. "COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems". *Nucleic Acids Research* 43.D1 (2014), pp. D82–D86. URL: `http://coxpresdb.jp`.

[144] M. Ongenaert et al. "PubMeth: a cancer methylation database combining text-mining and expert annotation". *Nucleic Acids Research* 36.suppl_1 (2007), pp. D842–D846.

[145] C. Ooi and P. Tan. "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data". *Bioinformatics* 19.1 (2003), pp. 37–44.

[146] P. Pagel et al. "The MIPS mammalian protein–protein interaction database". *Bioinformatics* 21.6 (2004), pp. 832–834.

[147] J.-W. Pan et al. "The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences". *Nature Communications* 11.1 (2020), pp. 1–12.

[148] G. Papachristoudis, S. Diplaris, and P. A. Mitkas. "SoFoCles: feature filtering for microarray classification based on Gene Ontology". *Journal of Biomedical Informatics* 43.1 (2010), pp. 1–14.

[149] C. Park, J. Ahn, H. Kim, and S. Park. "Integrative Gene Network Construction to Analyze Cancer Recurrence Using Semi-Supervised Learning". *PloS One* 9.1 (2014), e86309.

[150] J. S. Parker et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes". *Journal of Clinical Oncology* 27.8 (2009), pp. 1160–1167.

[151] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard. "Mining Gene Expression Data Using Domain Knowledge". *International Journal of Software and Informatics* 2.2 (2008), pp. 215–231.

[152] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[153] B. Peng et al. "An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways". *PloS One* 8.7 (2013), e67672.

[154] C. Perscheid. "Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches". *Briefings in Bioinformatics* 22.3 (2020), bbaa151.

[155] C. Perscheid. *Comprior GitHub Repository*. 2021. URL: `https://github.com/CPerscheid/Comprior`.

[156] C. Perscheid. "Comprior: facilitating the implementation and automated benchmarking of prior knowledge-based feature selection approaches on gene expression data sets". *BMC Bioinformatics* 22.1 (2021), pp. 1–15.

[157] C. Perscheid. *GitHub Repository for Supplementary Material*. 2022. URL: `https://github.com/CPerscheid/Dissertation_Supplementary`.

[158] C. Perscheid. "The impact of integrating prior knowledge during biomarker detection: A case study on high-dimensional gene expression data" (2022). in preparation.

[159] C. Perscheid, B. Grasnick, and M. Uflacker. "Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches". *Journal of Integrative Bioinformatics* 16.1 (2019), p. 20180064.

[160] C. Perscheid and M. Uflacker. "Integrating Biological Context into the Analysis of Gene Expression Data". In: *International Symposium on Distributed Computing and Artificial Intelligence*. Springer. 2018, pp. 339–343.

[161] C. Perscheid et al. "A Tissue-aware Gene Selection Approach for Analyzing Multi-tissue Gene Expression Data". In: *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2018, pp. 2159–2166.

[162] C. Perscheid et al. "Ebola outbreak containment: real-time task and resource coordination with SORMAS". *Frontiers in ICT* 5 (2018), p. 7.

[163] B. Pes, N. Dessì, and M. Angioni. "Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data". *Information Fusion* 35.C (2017), pp. 132–147.

[164] L. E. Peterson and T. Kovyrshina. "DNA repair gene expression adjusted by the PCNA metagene predicts survival in multiple cancers". *Cancers* 11.4 (2019), p. 501.

[165] J. Piñero et al. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes". *Database* 2015 (2015). URL: `http://disgenet.org/`.

[166] J. Piñero et al. "The DisGeNET knowledge platform for disease genomics: 2019 update". *Nucleic Acids Research* 48.D1 (2020), pp. D845–D855.

[167] A. Pratapa et al. "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". *Nature Methods* 17.2 (2020), pp. 147–154.

[168] A. L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". *Nature Genetics* 38.8 (2006), pp. 904–909.

[169] J. Qi and J. Tang. "Integrating Gene Ontology into discriminative powers of genes for feature selection in microarray data". In: *Proceedings of the ACM Symposium*

on Applied Computing. ACM. New York, NY, USA: Association for Computing Machinery, 2007, pp. 430–434.

[170] B. Quanz, M. Park, and J. Huan. "Biological pathways as features for microarray data classification". In: *International Workshop on Data and Text Mining in Bioinformatics*. 2008, pp. 5–12.

[171] V. K. Raghu, X. Ge, P. K. Chrysanthis, and P. V. Benos. "Integrated theory- and data-driven feature selection in gene expression data analysis". In: *IEEE 33rd International Conference on Data Engineering*. IEEE. 2017, pp. 1525–1532.

[172] P. Raina et al. "GeneFriends 2021: Updated co-expression databases and tools for human and mouse genes and transcripts". *bioRxiv* (2021). URL: `http://genefriends.org`.

[173] F. Rapaport et al. "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data". *Genome Biology* 14.9 (2013), pp. 1–13.

[174] U. Raudvere et al. "g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". *Nucleic Acids Research* 47.W1 (2019), W191–W198.

[175] J. Reimand et al. "Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap". *Nature Protocols* 14.2 (2019), pp. 482–517.

[176] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. "Normalization of RNA-seq data using factor analysis of control genes or samples". *Nature Biotechnology* 32.9 (2014), pp. 896–902.

[177] M. E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* 43.7 (2015), e47–e47.

[178] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26.1 (2010), pp. 139–140.

[179] M. D. Robinson and A. Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". *Genome Biology* 11.3 (2010), pp. 1–9.

[180] I. Rodchenkov et al. "Pathway Commons 2019 update: integration, analysis and exploration of pathway data". *Nucleic Acids Research* 48.D1 (2019), pp. D489–D497.

[181] Y. Saeys, I. Inza, and P. Larrañaga. "A Review of Feature Selection Techniques in Bioinformatics". *Bioinformatics* 23.19 (2007), pp. 2507–2517.

[182] S. P. P. Salifu et al. "RNA-seq analyses: benchmarking differential expression analyses tools reveals the effect of higher number of replicates on performance". *bioRxiv* (2020).

[183] A. Sarkar, Y. Yang, and M. Vihinen. "Variation benchmark datasets: update, criteria, quality and applications". *Database* 2020 (2020).

[184] M. Saul and V. Dinu. "Family Rank: A graphical domain knowledge informed feature ranking algorithm". *Bioinformatics* (2021).

[185] M. Saul and V. Dinu. "Family Rank: a graphical domain knowledge informed feature ranking algorithm". *Bioinformatics* 37.20 (2021), pp. 3626–3631.

[186] C. F. Schaefer et al. "PID: the pathway interaction database". *Nucleic Acids Research* 37.suppl_1 (2009), pp. D674–D679.

[187] T. Schaffter, D. Marbach, and D. Floreano. "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods". *Bioinformatics* 27.16 (2011), pp. 2263–2270.

[188] M.-P. Schapranow et al. "A federated in-memory database system for life sciences". In: *Real-Time Business Intelligence and Analytics*. Springer, 2015, pp. 19–34.

[189] M.-P. Schapranow and C. Fähnrich. "Provision of Analyze Genomes services in a federated in-memory database system for life sciences". In: *Proceedings of the HPI Future SOC Lab*. University of Potsdam, 2015, pp. 39–42.

[190] M.-P. Schapranow et al. "In-Memory Computing Enabling Real-time Genome Data Analysis". *International Journal on Advances in Life Sciences* 6.1 and 2 (2014), pp. 11–30.

[191] M.-P. Schapranow and C. Fähnrich. "High-Performance In-Memory Genome Project". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel et al. University of Potsdam, 2014, pp. 11–30.

[192] M.-P. Schapranow and C. Fähnrich. "Setting up customized genome data analysis pipelines with Analyze Genomes". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel et al. University of Potsdam, 2014, pp. 11–30.

[193] M.-P. Schapranow and C. Fähnrich. "Analyze Genomes: a cloud platform enabling on-site analysis of sensitive medical data". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel et al. University of Potsdam, 2016, pp. 21–24.

[194] M.-P. Schapranow and C. Perscheid. "Extending Analyze Genomes to a federated in-memory database system for life sciences". In: *Proceedings of the HPI Future SOC Lab*. Ed. by C. Meinel et al. University of Potsdam, 2015, pp. 99–102.

[195] M.-P. Schapranow, C. Perscheid, and H. Plattner. "IT-aided business process enabling real-time analysis of candidates for clinical trials". In: *Proceedings of the 4th International Conference on Global Health Challenges*. 2015, pp. 67–73.

[196] M.-P. Schapranow et al. "The Medical Knowledge Cockpit: Real-time analysis of big medical data enabling precision medicine". In: *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2015, pp. 770–775.

[197] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. "A new measure for functional similarity of gene products based on Gene Ontology". *BMC Bioinformatics* 7.1 (2006), pp. 1–16.

[198] B. Shao and T. Conrad. "Epithelial-Mesenchymal Transition Regulatory Network-Based Feature Selection in Lung Cancer Prognosis Prediction". In: *International Conference on Bioinformatics and Biomedical Engineering*. Springer. 2016, pp. 135–146.

[199] S. T. Sherry et al. “dbSNP: the NCBI database of genetic variation”. *Nucleic Acids Research* 29.1 (2001), pp. 308–311.

[200] Y. Shimoni. “Association between expression of random gene sets and survival is evident in multiple cancer types and may be explained by sub-classification”. *PLoS Computational Biology* 14.2 (2018), e1006026.

[201] N. Škunca, R. J. Roberts, and M. Steffen. “Evaluating computational Gene Ontology annotations”. In: *The Gene Ontology Handbook*. Humana Press, New York, NY, 2017, pp. 97–109.

[202] D. N. Slenter et al. “WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research”. *Nucleic Acids Research* 46.D1 (2017), pp. D661–D667. URL: `http://www.wikipathways.org`.

[203] D. Smedley et al. “The BioMart community portal: an innovative alternative to large, centralized data repositories”. *Nucleic Acids Research* 43.W1 (2015), W589–W598. URL: `http://www.biomart.org`.

[204] M. Sobrinho, M. Rosa, W. Silva, and A. Araújo. “Resource Prediction Service for Efficient Execution of Bioinformatics Workflows in Federated Cloud with Machine Learning”. In: *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2021, pp. 1975–1983.

[205] C. Soneson and M. D. Robinson. “iCOBRA: open, reproducible, standardized and live method benchmarking”. *Nature Methods* 13.4 (2016), pp. 283–283.

[206] S. Sood et al. “A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status”. *Genome Biology* 16.1 (2015), pp. 1–17.

[207] S. Srivastava, L. Zhang, R. Jin, and C. Chan. “A novel method incorporating Gene Ontology information for unsupervised clustering and feature selection”. *PloS One* 3.12 (2008), e3860.

[208] C. Stark et al. “BioGRID: a general repository for interaction datasets”. *Nucleic Acids Research* 34.suppl_1 (2006), pp. D535–D539. URL: `https://thebiogrid.org/`.

[209] F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci. “Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes”. *The Annals of Applied Statistics* 5.3 (2011), pp. 1978–2002.

[210] F. C. Stingo and M. Vannucci. “Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data”. *Bioinformatics* 27.4 (2010), pp. 495–501.

[211] D. Strbenac, G. J. Mann, J. T. Ormerod, and J. Y. Yang. “ClassifyR: an R package for performance assessment of classification with applications to transcriptomics”. *Bioinformatics* 31.11 (2015), pp. 1851–1853.

[212] D. Strbenac et al. “Quantitative performance evaluator for proteomics (QPEP): Web-based application for reproducible evaluation of proteomics preprocessing methods”. *Journal of Proteome Research* 16.7 (2017), pp. 2359–2369.

[213] L. Strömbäck and P. Lambrix. "Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX". *Bioinformatics* 21.24 (2005), pp. 4401–4407.

[214] S. Su et al. "CellBench: R/Bioconductor software for comparing single-cell RNA-seq analysis methods". *Bioinformatics* 36.7 (2020), pp. 2288–2290.

[215] A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[216] H. Sun, W. Lin, R. Feng, and H. Li. "Network-regularized high-dimensional Cox regression for analysis of genomic data". *Statistica Sinica* 24.3 (2014), pp. 1433–1459.

[217] T. Swarnkar et al. "Multiview clustering on PPI network for gene selection and enrichment from microarray data". In: *IEEE International Conference on Bioinformatics and Bioengineering*. IEEE. 2014, pp. 15–22.

[218] T. Swarnkar et al. "Identifying dense subgraphs in protein–protein interaction network for gene selection from microarray data". *Network Modeling Analysis in Health Informatics and Bioinformatics* 4 (2015), p. 33.

[219] D. Szklarczyk et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". *Nucleic Acids Research* 43.D1 (2014), pp. D447–D452. URL: `http://string-db.org/`.

[220] R. Tabares-Soto et al. "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data". *PeerJ Computer Science* 6 (2020), e270.

[221] D. Tenenbaum. *KEGGREST: client-side REST access to KEGG*. R package version 1.26.1. 2018. URL: `http://www.bioconductor.org/packages/release/bioc/html/KEGGREST.html`.

[222] The Cancer Genome Atlas Research Network. "Comprehensive molecular portraits of human breast tumours". *Nature* 490 (2012), pp. 61–70.

[223] The Cancer Genome Atlas Research Network. "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas". *New England Journal of Medicine* 372.26 (2015), pp. 2481–2498.

[224] The Cytoscape Consortium. *The Simple Interaction Format*. 2020. URL: `http://manual.cytoscape.org/en/stable/Supported_Network_File_Formats.html` (visited on 02/25/2022).

[225] The Gene Ontology Consortium. *GO Annotation File (GAF) files*. 2020. URL: `http://geneontology.org/docs/go-annotation-file-gaf-format-2.1/` (visited on 02/25/2022).

[226] The Gene Ontology Consortium. "The Gene Ontology resource: enriching a GOld mine". *Nucleic Acids Research* 49.D1 (2021), pp. D325–D334.

[227] The UniProt Consortium. "Activities at the universal protein resource (UniProt)". *Nucleic Acids Research* 42.D1 (2014), pp. D191–D198.

[228] The UniProt Consortium. "UniProt: the universal protein knowledgebase in 2021". *Nucleic Acids Research* 49.D1 (2020), pp. D480–D489. URL: http://www.uniprot.org/uniprot/.

[229] L. Tian et al. "Discovering statistically significant pathways in expression profiling studies". *Proceedings of the National Academy of Sciences* 102.38 (2005), pp. 13544–13549.

[230] R. Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282.

[231] A. Tjärnberg et al. "GeneSPIDER–gene regulatory network inference benchmarking with controlled network and data properties". *Molecular BioSystems* 13.7 (2017), pp. 1304–1312.

[232] D. Tom-Aba et al. "User evaluation indicates high quality of the Surveillance Outbreak Response Management and Analysis System (SORMAS) after field deployment in Nigeria in 2015 and 2018". In: *German Medical Data Sciences*. Vol. 253. 2018, pp. 233–237.

[233] A. Tomczak et al. "Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations". *Scientific Reports* 8.1 (2018), pp. 1–10.

[234] V. S. Tseng and H.-H. Yu. "Microarray data classification by multi-information based gene scoring integrated with Gene Ontology". *International Journal of Data Mining and Bioinformatics* 5.4 (2011), pp. 402–416.

[235] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources". *Nature Methods* 13.12 (2016), pp. 966–967. URL: https://www.omnipathdb.org.

[236] A. Tyryshkina, N. Coraor, and A. Nekrutenko. "Predicting runtimes of bioinformatics tools based on historical data: five years of Galaxy usage". *Bioinformatics* 35.18 (2019), pp. 3453–3460.

[237] M. Uhlén et al. "Tissue-based map of the human proteome". *Science* 347.6220 (2015), p. 1260419. URL: https://proteinatlas.org.

[238] G. Valabrega, F. Montemurro, and M. Aglietta. "Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer". *Annals of Oncology* 18.6 (2007), pp. 977–984. URL: https://www.sciencedirect.com/science/article/pii/S0923753419379013.

[239] I. Valavanis et al. "Cancer biomarkers from genome-scale DNA Methylation: comparison of evolutionary and semantic analysis methods". *Microarrays* 4.4 (2015), pp. 647–670.

[240] L. J. Van't Veer et al. "Gene Expression profiling predicts clinical outcome of breast cancer". *Nature* 415 (2002), pp. 530–536.

[241] D. Venet, J. E. Dumont, and V. Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". *PLoS Computational Biology* 7.10 (2011), e1002240.

[242] J. C. Venter et al. "The sequence of the human genome". *Science* 291.5507 (2001), pp. 1304–1351.

[243] P. Verschaffelt et al. "MegaGO: a fast yet powerful approach to assess functional Gene Ontology similarity across meta-omics data sets". *Journal of Proteome Research* 20.4 (2021), pp. 2083–2088.

[244] J.-P. Vert and M. Kanehisa. "Graph-driven feature extraction from microarray data using diffusion lernels and kernel CCA". In: *Advances in Neural Information Processing Systems*. 2003, pp. 1449–1405.

[245] J. Vinagre et al. "Frequency of TERT promoter mutations in human cancers". *Nature Communications* 4.1 (2013), pp. 1–6.

[246] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo. "Gene expression correlation and Gene Ontology-based similarity: an assessment of quantitative relationships". In: *Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE. 2004, pp. 25–31.

[247] W. Wang and W. Liu. "Integration of gene interaction information into a reweighted Lasso-Cox model for accurate survival prediction". *Bioinformatics* 36 (22–23 2020), pp. 5405–5414.

[248] W. Webber, A. Moffat, and J. Zobel. "A similarity measure for indefinite rankings". *ACM Transactions on Information Systems* 28.4 (2010), pp. 1–38.

[249] P. Wei and W. Pan. "Incorporating Gene Networks into Statistical Tests for Genomic Data via a Spatially Correlated Mixture Model". *Bioinformatics* 24.3 (2007), pp. 404–411.

[250] X. Wei, C. Zhang, P. L. Freddolino, and Y. Zhang. "Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons". *Bioinformatics* 36.16 (2020), pp. 4383–4388.

[251] H. G. Welch and P. C. Albertsen. "Prostate Cancer Diagnosis and Treatment After the Introduction of Prostate-Specific Antigen Screening: 1986–2005". *Journal of the National Cancer Institute* 101.19 (2009), pp. 1325–1329.

[252] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3.1 (2016), pp. 1–9.

[253] J. Willforss, A. Chawade, and F. Levander. "NormalyzerDE: online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis". *Journal of Proteome Research* 18.2 (2018), pp. 732–740.

[254] C. Wu et al. "A selective review of multi-level omics data integration using variable selection". *High-Throughput* 8.1 (2019), p. 4.

[255] Z. Xie et al. "Gene set knowledge discovery with Enrichr". *Current Protocols* 1.3 (2021), e90.

[256] J. Xing et al. "DiseaseMeth version 3.0: a major expansion and update of the human disease methylation database". *Nucleic Acids Research* 50.D1 (2021), pp. D1208–D1215.

[257] X. Xu and A. Zhang. “Selecting informative genes from microarray dataset by incorporating Gene Ontology”. In: *IEEE Symposium on Bioinformatics and Bioengineering*. IEEE. 2005, pp. 241–245.

[258] F. Yang and K. Mao. “Robust feature selection for microarray data based on multicriterion fusion”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.4 (2011), pp. 1080–1092.

[259] P. Yang, H. Huang, and C. Liu. “Feature selection revisited in the single-cell era”. *Genome Biology* 22.1 (2021), pp. 1–17.

[260] K. Yoshida and Y. Miki. “Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage”. *Cancer Science* 95.11 (2004), pp. 866–871.

[261] C. Zeng, D. C. Thomas, and J. P. Lewinger. “Incorporating prior knowledge into regularized regression”. *Bioinformatics* 37.4 (2021), pp. 514–521.

[262] Y. Zhang, J. Xuan, R. Clarke, and H. W. Ressom. “Module-based breast cancer classification”. *International Journal of Data Mining and Bioinformatics* 7.3 (2013), pp. 284–302.

[263] Y. Zhang, G. Parmigiani, and W. E. Johnson. “ComBat-seq: batch effect adjustment for RNA-seq count data”. *NAR Genomics and Bioinformatics* 2.3 (2020). lqaa078.

[264] Z. Zhao et al. “An integrative approach to identifying biologically relevant genes”. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM. 2010, pp. 838–849.

[265] Y. Zhu, X. Shen, and W. Pan. “Network-Based Support Vector Machine for Classification of Microarray Samples”. *BMC Bioinformatics* 10.1 (2009), S21.

[266] J. Zhuang, M. Widschwendter, and A. E. Teschendorff. “A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform”. *BMC Bioinformatics* 13.1 (2012), pp. 1–14.

[267] J. M. Zook et al. “Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls”. *Nature Biotechnology* 32.3 (2014), pp. 246–251.

[268] J. M. Zook et al. “Extensive sequencing of seven human genomes to characterize benchmark reference materials”. *Scientific Data* 3.1 (2016), pp. 1–26.

[269] J. M. Zook et al. “A robust benchmark for detection of germline large deletions and insertions”. *Nature Biotechnology* (2020), pp. 1–9.