

Disease Module Discovery in Systems Medicine

Olga Lazareva

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Mathias Wilhelm

Prüfer*innen der Dissertation:

1. Prof. Dr. Bernard Küster
2. Prof. Dr. Jan Baumbach

Die Dissertation wurde am 21.01.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 01.06.2022 angenommen.

Acknowledgments

I just can not believe that this is happening. These last 3 years were indeed an unbelievable journey: great travels (well, for 1 year), 5 COVID waves, amazing projects, and friendships. I can not stress enough the last part, because I think that the atmosphere of support and fun that we had in exbio was truly the best part of my experience. I can not thank Dr. Markus List enough, who was never my supervisor on paper but was a great supervisor, in fact. Thank you so much for the support, for your ability to stay positive and friendly, and always be there when needed. I would like to thank Prof. Jan Baumbach for taking me in and supporting me in obtaining my bidt fellowship. The fellowship gave me a huge motivational boost that powered me through the last 2 years. I also would like to thank Prof. Bernhard Küster for agreeing to support me in the last steps of my journey. And of course, I want to thank all members of the exbio chair for the cool discussions, fun lunch breaks and help with the actual writing of my thesis. In particular, I would like to thank Dr. Olga Tsoy, Amit Fenn, Nikolai Köhler, Laura Hernandez Lorenzo, Julian Späth, Markus Hoffmann, Melissa Klug, Tim Rose, and Dr. Olga Zolotareva for reading my thesis and helping me to improve it.

I would like to thank my family that allowed me to be where I am right now, doing what I love the most. My parents that supported all my decisions, and my sister that already knows everything about my colleagues, supervisors, and collaborators. I also would like to thank my boyfriend Tim, who supported me enormously through my dissertation writing and had to deal with all my setbacks and frustrations. I could not have dreamt of a more supportive partner.

I am very happy that this is almost over, but I am also sad to let go of everyone who was with me on this journey. Thank you again. Memories of these times will stay with me forever (just like my doctoral degree).

Abstract

Systems Medicine is an interdisciplinary field of study that brings together clinicians, biologists, statisticians, data- and computer scientists. By bringing expertise from different fields, Systems Medicine employs complex algorithmic approaches to integrate various biological data and provide a holistic understanding of the human body. Ultimately, Systems Medicine aims to redefine disease phenotypes based on molecular mechanisms rather than symptoms. Substantial attention has been directed towards the extraction of disease modules based on the aggregation of protein-protein interaction (PPI) networks and transcriptomics data.

While a large number of methods have been developed for disease module extraction, two open questions remain which I will address in this thesis. First, existing methods extract disease modules in a supervised fashion and are thus limited in their capacity to detect new endophenotypes together with the mechanisms that drive them. Second, it remains unclear to what extent these methods leverage prior knowledge about functional interactions in PPI networks as remarkably few studies have sought to examine the exact informational gain that PPI networks provide.

This cumulative thesis comprises three original publications addressing this unmet need and examining Systems Medicine approaches to bridge data from different molecular layers. The first publication describes the only unsupervised disease module extraction method, BiCoN (Biclustering Constrained by Networks). BiCoN performs simultaneous patient clustering and protein module extraction to obtain a potential mechanistic explanation of a studied condition. The obtained results conclusively showed that PPI integration made clustering results more resistant to noise and batch effects in transcriptomics data compared to results of regular clustering and biclustering methods. In light of these findings, the second publication attempts to numerically evaluate the influence of different properties of PPI networks (such as node degree distribution, individual node's degree, hub node presence) on the results of disease module extraction methods. The analysis demonstrated that disease module extraction methods could often not exploit the actual PPIs but instead relied on a degree of a particular protein. This conclusion is particularly alarming due to previous research showing that protein degree in PPI networks correlates with the number of studies conducted on a protein and does not necessarily reflect the actual number of interactions.

While the field of module discovery is currently very much focused on transcriptomics data, the third publication gives a different perspective where epigenomic regulation patterns are used as potential disease mechanisms. In this review paper, we have explored the potential for integrating epigenomics information using machine learning-based approaches for DNA methylation deconvolution, gene expression prediction, and multi-omics integration on a single-cell level. In the future, such efforts can help integrate knowledge about gene-regulatory relations in the disease module discovery process in a true systems medicine fashion.

Taken together, the findings suggest various directions that can improve disease module mining quality. Improvements are suggested on the algorithmic side, data acquisition, and modeling of the problem. These findings can shape the future of the disease mechanisms extraction and eventually contribute to safer treatments, better diagnostics, and efficient drug design.

Kurzfassung

Systemmedizin ist ein interdisziplinäres Forschungsfeld, das WissenschaftlerInnen aus der Klinik, Biologie, Statistik und Informatik zusammenbringt. Mit der Expertise aus verschiedenen Disziplinen nutzt die Systemmedizin komplexe Algorithmen, um biologische Daten zu integrieren und ein gesamtheitliches Verständnis des menschlichen Körpers zu ermöglichen. Die Systemmedizin zielt darauf ab, Krankheits-Phänotypen neu zu definieren, basierend auf molekularen Mechanismen anstatt durch Symptome. Ein wichtiger Teil, der in der Forschung viel Aufmerksamkeit erfahren hat, ist die Extraktion von Krankheitsmodulen, die Protein-Protein Interaktionsnetzwerke (PPI) mit Transkriptomik Daten integrieren.

Viele Methoden für die Extraktion von Krankheitsmodulen wurden bereits entwickelt. Zwei wichtige Fragestellungen blieben dabei aber bisher unbeantwortet, welche in dieser Dissertation adressiert werden. Erstens, existierende Methoden arbeiten ausschließlich mit überwachtem Lernen, nicht aber mit unüberwachten. Das schränkt sie in ihrer Möglichkeit ein, neue Endophänotypen mit entsprechenden Mechanismen zu finden. Zweitens, bisher ist nicht klar, welchen Einfluss die Struktur von PPI Netzwerken auf die Algorithmen haben. Nur sehr wenige Studien haben versucht diesen Informationsgehalt zu untersuchen.

Diese kumulative Dissertation beinhaltet drei Originalpublikationen, welche diese Fragen untersuchen und systemmedizinische Methoden zur Verbindung von verschiedenen molekularen Ebenen diskutieren. Die erste Publikation beschreibt die einzige veröffentlichte Methode zur unüberwachten Krankheitsmodul Extraktion, BiCoN (Biclustering Constrained by Networks). BiCoN führt ein simultanes Gruppieren von Patienten und Extrahieren von Proteinmodulen durch, um eine potentiell mechanistische Erklärung einer untersuchten Krankheit zu finden. Die Resultate dieser Studie zeigten, dass PPI Integration das Gruppieren der Patienten robuster gegenüber Rauschen und Batch Effekten in Transkriptomik Daten macht. Ausgehend von diesen Ergebnissen, untersuchte ich in der zweiten Publikation den Einfluss von verschiedenen Charakteristika der Netzwerke auf Methoden zur Extraktion von Krankheitsmodulen. Die Analyse zeigte, dass die Methoden oft nicht von den ganzen Netzwerken profitieren, sondern sich nur auf den Knotengrad einzelner Proteine fokussieren. Dieses Fazit ist alarmierend, weil in vorherigen Studien bereits gezeigt wurde, dass der Knotengrad mit der Anzahl der Studien, in denen das jeweilige Protein untersucht wurde, korreliert und nicht die wirkliche Anzahl der Interaktionen widerspiegelt.

Da sich das Feld der Modulextraktion in Krankheiten sehr auf Transkriptomik Daten fokussiert, gibt die dritte Publikation einen Blick auf epigenetische Regulationsmuster, welche für die Erklärung von Krankheitsmechanismen genutzt werden können. In diesem Review Artikel, untersuchte ich das Potential der Integration von epigenetischen Information mit Methoden des maschinellen Lernens, zur Dekonvolution von DNA methylierung, Vorhersage von Transkription, und Integration von Multi-Omik Daten auf der Einzelzell Ebene. In der

Zukunft können diese Methoden dabei helfen, Wissen über genetische Regulation in den Prozess der Extraktion von Krankheitsmodulen zu integrieren.

Die hier vorgestellten Ergebnisse zeigen verschiedene Wege, mit welchen die Erkennung von Krankheitsmodulen verbessert werden kann. Diese umfassen Algorithmen, Daten Aqoise, und Modellierung des Problems. Damit können Methoden zur Modulextraktion in der Zukunft verbessert werden und zu besseren Diagnosen, Behandlungen und Medikamenten Design beitragen.

Contents

Acknowledgments	ii
Abstract	iii
Kurzfassung	v
1. General Introduction	1
1.1. Motivation	1
1.2. Outline	3
2. Background	4
2.1. Molecular Biology and Human Disease	4
2.1.1. Molecular data and technologies	5
2.1.2. Single cell technologies	11
2.2. Molecular data integration and interpretation	14
2.2.1. Goals	14
2.2.2. Data repositories	14
2.2.3. Tools	15
2.3. Artificial intelligence in biology	17
2.3.1. Classic heuristic approaches	17
2.3.2. Machine learning	18
2.3.3. Artificial neural networks in biomedicine	22
2.3.4. Common challenges in machine learning	23
2.4. Network and Systems Medicine	25
2.4.1. Molecular networks	25
2.4.2. De novo endophenotyping	27
2.4.3. Active Module Identification	28
2.5. Objective	30
3. General Methods	32
3.1. BiCoN algorithmic framework description	32
3.1.1. Metaheuristic approach	32
3.2. Active module identification methods evaluation	34
3.2.1. PPI networks and randomizations	34
3.2.2. Gene sets evaluation	36

4. Publications	38
4.1. Publication 1: BiCoN: network-constrained biclustering of patients and omics data	38
4.2. Publication 2: On the limits of active module identification	48
4.3. Publication 3: Machine learning for deciphering cell heterogeneity and gene regulation	61
5. General Discussion and Outlook	72
5.1. PPI networks as a prior knowledge source	72
5.2. Unsupervised learning approaches have a potential to overcome PPI biases . .	73
5.3. Algorithmic roadblocks	74
5.4. Outlook	76
A. Appendix	81
A.1. Assessment of BiCoN with the AMI testing suite	81
Acronyms	82
References	83

1. General Introduction

1.1. Motivation

Rich molecular data availability provides invaluable opportunities to improve human health and make medicine more precise, personalized, and safe. While data from different molecular levels (multi-omics data, e.g., transcriptomics, proteomics, etc.) is available, every single level provides a limited understanding of the undergoing pathological processes. Two strategies are often considered together or separately to systematically understand a disease: integration of multi-omics data and integration of prior information relevant to the whole population. The prior information consists of, for instance, the knowledge about disease pathways, metabolic pathways, protein structures, or protein interactions. Multi-omics data integration presumes the availability of different omics data types for the same set of patients. Multi-omics data acquisition can be very expensive and therefore less common in practice. On the other hand, prior information integration usually does not require additional costs as various databases provide this information publicly.

Both strategies are often applied in Systems Medicine [1, 2, 3]. Systems medicine is a field of study that looks at the human body as a whole [4]. To obtain the holistic perspective, Systems Medicine relies on complex molecular interactions within the human body that can mechanistically explain patients' phenotype. Mechanistic explanations are essential for moving on from symptom-based disease definitions that are often unable to offer disease treatment, but only partly successful symptoms management [5]. Extraction of the mechanistically connected molecular entities capable of phenotype explanation is often referred to as disease modules mining tasks [6, 7]. A disease module can consist of various entities such as genes, proteins, metabolites, and some others. The connectivity requirement implies that the module is not only an indicator of a specific process but the molecular mechanism that explains the undergoing pathological process. This mechanism might be further used in clinical practice for diagnostics, risk prediction, and drug development. Despite the vast potential of disease module mining, there are currently no gold standards methods to decipher molecular data and extract a disease mechanism. Nevertheless, many methods attempt to extract disease mechanisms, but they have certain limitations related to the fact that disease module mining is algorithmically and conceptually challenging [8].

Bridging Systems medicine with Artificial Intelligence (AI) is vital for large-scale disease module mining. Therefore, the main goal of the Dissertation is to explore AI methods' contribution to Systems Medicine. In particular, I will focus on whether the integration of protein-protein interaction (PPI) prior information and bulk transcriptomics studies can increase the quality of the disease module discovery. The underlying hypothesis is that the integration of PPI networks can provide a mechanistic explanation of differences in

transcriptomic profiles between different conditions and thus contribute to understanding the condition on the molecular level.

The first publication [9] provides a new method (BiCoN) that performs disease module mining and simultaneous patient clustering into clinically-relevant groups, using transcriptomics data and PPI networks. The simultaneous search for disease modules and corresponding patient clusters allows BiCoN to perform patient stratification from a mechanistic perspective. Clustering and disease module extraction are usually done sequentially (first clustering and then PPI analysis), and thus the results are more likely to be driven by batch effects and noise in transcriptomics data [9]. BiCoN is the first unsupervised disease module mining algorithm that is capable of reproducing known disease subtypes as well as novel, clinically relevant patient subgroups [9]. BiCoN performance demonstrated two essential points on which further research was built: PPI networks indeed make results more robust to noise, and thus the developed algorithm demonstrated superiority to the state-of-the-art clustering and biclustering methods (i), cell-composition might be a strong confounder and drive patients separation into high and low immune response groups rather than into different cancer subtypes (ii).

The first conclusion (i) has led to further questions about information gain provided by PPI networks. Despite the wide application of PPI networks in bioinformatics, their exact contribution is poorly understood. To address this issue, I developed the Active Module Identification (AMI) testing suite [10] to systematically assess the influence of PPI networks on the disease module mining task. The analysis was performed by running multiple available tools for the identification of disease modules in PPI networks. The networks were randomized to a different degree, measuring differences in tools' results on randomized and original networks. Several different conditions and PPI networks were tested to ensure that the observed effects are consistent across all available PPI networks and are not dependent on a particular disease. The comparison questioned the biological value of PPI networks and suggested potential bias originating from highly studied proteins.

The second conclusion (ii) suggested that cell type composition is an important confounding factor in the unsupervised analysis, and bulk studies do not fully account for it. Given the conclusions from the AMI testing suite, I explored methods for single-cell multi-omics analysis and integration. Bridging data of different origins and scales requires comprehensive machine learning methods. Thus the third publication [11], was developed to review the state-of-the-art machine learning methods in multi-omics data integration on a single-cell level.

The performed review suggested many opportunities to further increase the quality of disease module discovery methods. Prior information is not limited to PPI data but also might include regulatory interactions, metabolic networks, and other information obtained from curated databases. The reliability of complex algorithms results can be increased when evidence is observed on several molecular levels and/or supported by available database entries. Given the exponential development of machine learning methods and, particularly, methods that aim to network analysis (such as graph neural networks), it is of paramount importance to employ these methods to advance human health and disease approaches.

1.2. Outline

The background Chapter presents essential concepts from molecular biology and AI. In section 2.1 I describe the role of different molecular layers in an organism and how they can be affected by diseases. Next, section 2.2 elaborates further on the advantages of integration of multiple molecular layers and the use of prior biological knowledge. Popular algorithms and the largest multi-omics data repositories are also described.

To understand the necessary computational background, section 2.3 explains the basics of machine learning, statistics, and algorithmics and their application in computational biology. The next section (section 2.3) discusses the role of Systems Medicine and its potential for disease mechanisms extraction.

The discussed topics provide the essential background for the Methods chapter (chapter 3), where the algorithmic framework of BiCoN and the AMI testing suite is introduced. Next, chapter 4 provides summaries of all three publications and precisely describes my contribution. The full versions of the papers are embedded in the text.

In the Discussion (chapter 5), I describe the main roadblocks of Systems Medicine. In particular, I focus on limitations of PPI networks usage for Systems Medicine tools and assessment of those tools in terms of reproducibility and interpretability. I also suggests different opportunities to overcome the described constraints and provide conclusions about the conducted Ph.D. project.

2. Background

2.1. Molecular Biology and Human Disease

James Watson and Francis Crick discovered the structure of deoxyribonucleic acid (DNA) in 1953 [12] (based on data from Maurice Wilkins and Rosalind Franklin [13]). Three years later, Crick proposed the "central dogma of molecular biology" that describes how genetic information flows from DNA to ribonucleic acid - RNA (through transcription) and then from RNA to proteins (through translation) [14] (Figure 2.1). Thus the dogma suggests that DNA has all information needed to make functional products (proteins). Proteins are biological molecules that keep cells in an organism functional. Proteins consist of amino acids. The instruction to build a protein from amino acids is transcribed from DNA using messenger RNA (mRNA). Then ribosomes translate mRNA into proteins.

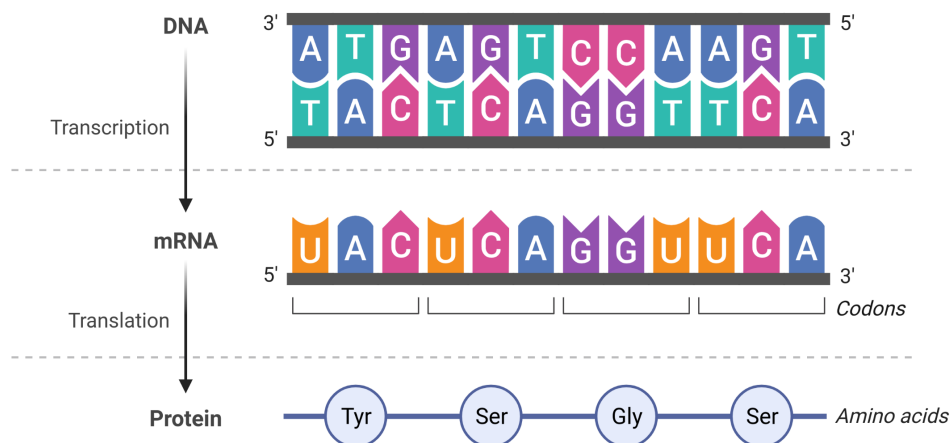


Figure 2.1.: The central dogma of biology. Genetic information flow goes from DNA to mRNA through transcription and then from mRNA to proteins through translation.

This model allowed molecular biologists to better understand the functioning of living organisms and how perturbations at one molecular level can have effects on the whole system and drastically influence a phenotype. Mutations in DNA are often considered as an example of how information flow can be disturbed. While some mutations do not have known influence on the protein sequence, others can largely affect protein function by letting it lose or gain additional functions [15]. Loss-of-function implies that less of a protein is

produced or its function has been compromised. For instance, a loss-of-function mutation in the CFTR gene, which is involved in the production of sweat, mucus, and digestive fluids [16] leads to the development of cystic fibrosis. When CFTR is not functional, body liquids become thicker, affecting the function of lungs, pancreas, kidneys, and intestines. Excess of mucus in the lungs leads to breathing difficulties and frequent lung infections and can even lead to the necessity of a lung transplant. Cystic fibrosis is a monogenic diseases (i.e., caused by mutations in one gene), as well as sickle cell anemia, Huntington’s disease, polycystic kidney disease, and many others. The majority of diseases are not monogenic but have a more complex nature [17]. They are caused by a combination of genetic variations that can occur in various locations and due to different environmental factors. Such diseases include cancer, asthma, mental disorders, infertility, and many others.

2.1.1. Molecular data and technologies

In this section I discuss different levels of molecular data following the central dogma of molecular biology. I start with discussing the DNA level (genome) and then move down to RNA (transcriptome), then to proteins and metabolites. Additionally, we discuss omics data that goes beyond the central dogma of molecular biology such as epigenomics. All described molecular layers are represented in Figure 2.2 as well as their proximity to genome and the extend of environmental influence.

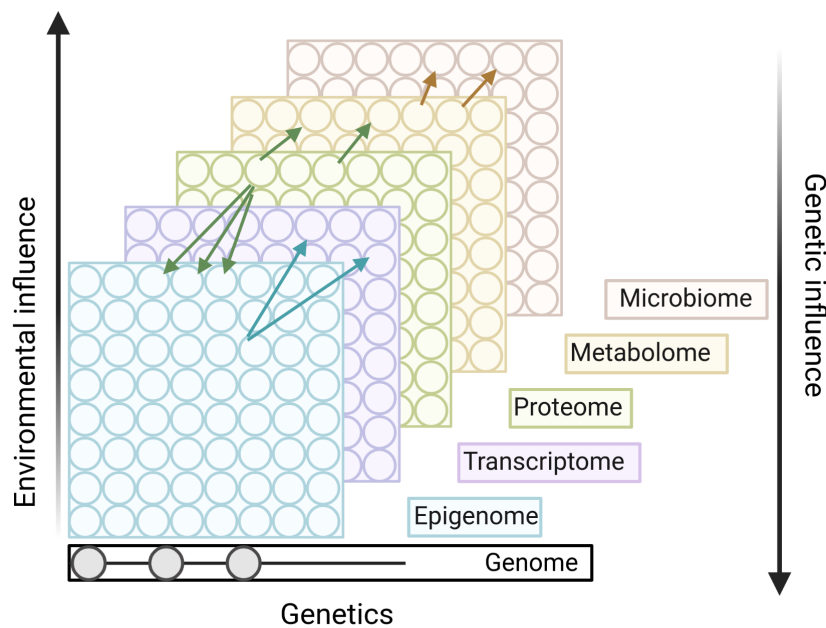


Figure 2.2.: The interplay between various omics layers. Molecular entities are represented with circles and colourful arrows represent some examples of possible interplay between the layers. The thick black arrows represent the influence of environmental factors and genetics.

Genomics

DNA sequencing All DNA contained in one cell is called a genome. A genome of a particular person is approximately 99.8% identical to all other humans [18] and understanding the variations in that other 0.2% is important to discriminate health and disease [19]. Modern sequencing methods allow for efficient analysis of the whole genome instead of focusing on single genes. These methods allow for finding variations that can lead to a disease or protect from one [20]. We can separate possible genomic variation into two groups: Single Nucleotide Variations (SNVs) and Structural Variations (SVs) [21]. SNVs are characterized by either single nucleotide variations (also called Single nucleotide polymorphisms- SNPs) or small insertions and deletions (indels). SVs are more complex and can represent large (100-1k bp) insertions/deletions, gene copy-number variations (CNVs), inversions, and even sequence relocation to another chromosomal region. Both SNVs and SVs can occur in coding and non-coding regions of the genome. A variation in a coding region usually affects the protein sequence while alterations in non-coding regions affect gene expression and splicing [22].

With sequencing costs going down from 100 million dollars per genome (year 2001) to 1000 dollars (year 2015), genomics analysis was experiencing rapid growth at the beginning of the twenty-first century [23]. Availability of full-genome sequences led to the development of Genome-Wide Association Studies (GWAS) that aim to find variations at a single position in DNA (SNPs) associated with particular traits or conditions. GWAS revealed insights in conditions like Alzheimer's disease by identifying disease-associated genes [24] and loci associated with increased risk of developing the condition [25]. GWAS aim to provide a comprehensive catalog of SNPs-condition associations, making the connection of genotype to phenotype relatively straightforward. Despite the promise of GWAS, many limitations have also been identified, such as difficulties in studying rare variants [26], statistical challenges when testing for over millions of SNPs [27], and underrepresentation of racial and ethnic minorities [23].

Based on whole-genome sequencing, CNV analysis determines repetitions of some genome sections. These repetitions can vastly vary among individuals and can also be a reason for many genetic diseases [28]. Studies of CNVs demonstrate that CNVs can be used for diagnostic purposes, such as efficient stratification of patients with gastric cancer into HER2-positive and HER2-low patients and thus determining those patients who can benefit from HER2 inhibitors-based therapy [29].

Classic technologies to capture genetic variants include sanger sequencing [30], microarrays [31] and Next Generation Sequencing (NGS) [22]. Sanger sequencing is performed base-by-base of a given locus that represents a researcher's interest. DNA microarrays are using a set of predefined oligonucleotide probes that hybridize DNA. The oligonucleotide probes can be distributed over the entire genome or be concentrated in an area of interest. Finally, NGS allows splitting the genome into pieces that are subsequently sequenced and then aligned to a reference genome. NGS is considered to be a much more sensitive technology compared to microarrays.

Transcriptomics

RNA sequencing During *transcription*, an RNA copy of a gene sequence is created. This copy is called messenger RNA (mRNA) and its main purpose is to direct the synthesis of a particular protein that the gene sequence encodes. Other commonly studied types of RNA are transfer RNA (tRNA) and ribosomal RNA (rRNA). tRNA and rRNA are present in all organisms, and together with mRNA, they participate in protein formation [32].

Similar to genome capturing technologies, it is also possible to measure the whole transcriptome using microarray, and sequencing technologies [33]. Microarrays have a clear advantage in terms of the price of an experiment but require pre-designed probes, and if a transcript is not included in the designed probe set, it will not be captured by the microarray [34]. RNA sequencing, on the other hand, takes advantage of the fragmentation of RNA and then aligns the fragments to a reference sequence [35]. Gene expression is measured based on the number of reads mapped to each locus in the assembly step.

Post hoc analysis of RNA-seq or microarray data allows performing numerous kinds of quantitative analysis, gaining insights into various biological processes. Several popular analysis perspectives are:

- **Differential expression analysis:** one of the most commonly used methods to determine gene expression differences between conditions [36, 37]. The most traditional application case is a design with two conditions, where one represents a phenotype of interest (case), and another represents healthy donors (control). However, designs with multiple conditions are also possible. The analysis result is usually a list of genes that have significantly different expression patterns in different conditions.
- **Coexpression networks:** coexpression networks are built based on gene-to-gene correlation and can be used to provide functional annotation based on the guilt-by-association principle [38]. Assuming that a gene with an unknown role is densely connected with genes involved in a certain biological process or a pathway, a hypothesis about the unknown gene can be generated.
- **Alternative splicing (AS analysis):** AS is a process that allows a single gene to code for multiple different proteins [39]. It largely contributes to protein diversity as it happens in over 90% of genes [39]. One of the primary goals of AS analysis is to identify how AS events differ between conditions.
- **Variant discovery:** Variant calling in RNA-seq is mainly similar to DNA variant calling, except for covering only expressed regions of the genome [40]. Additionally, RNA-seq based variant calling allows performing allele-specific expression analysis [41].

It is also possible to connect gene expression with genetic variations, bridging RNA-level data and DNA-level data. This type of analysis is commonly referred to as expression quantitative trait loci (eQTL) mapping. We will discuss various other ways of connecting data from multiple omics layers in Chapter 1.2.

Proteomics

Proteome Proteins are often referred to as functional units of a cell as they are required for the regulation and functioning of all bodies, tissues, and organs. Proteins are produced in the process of *translation* when nucleotide triplets are translated into amino acids. Then, primary sequence polypeptides are folded in a particular way to produce a functional protein [33]. While there are 5 nucleotides in DNA and RNA sequences, over 20 different amino acids can make up a protein. Protein formation is also subject to much variability partly due to different folding possibilities, effect of alternative splicing on the primary sequence, and other factors, making proteomics a very challenging field for quantification.

The proteomics field often relies on mass spectrometry technologies (MS) [42] that allow measuring the mass-to-charge ratio of ions to identify and quantify molecules in simple and complex mixtures [43]. Mass-spectrometry allows to perform high-throughput proteome analysis, study complex protein mixtures and perform large-scale protein characterization with high sensitivity. Various other techniques are also available, including chromatography-based (also in combination with MS), enzyme-linked immunosorbent assay for selective protein analysis, protein microarrays or chips for high-throughput and rapid expression analysis, gel-based approaches for separation of complex protein samples [44].

Proteomics analysis tackles a broad set of research questions, including (i) inference of a protein's molecular function, (ii) connection of variations in protein structures and diseases, (iii) drug development, (iv) protein-protein interactions (PPIs), and many others.

With the available high throughput technologies, massive proteomics data have been collected in databases for bioinformatics analysis. The tools that have been developed are aiming to predict and analyze protein interactions, 3D protein structures, protein domains (and their interactions), and motifs [44]. Various alignment tools of protein sequences allowed to establish evolutionary relationships.

Thus, proteomics offers comprehensive approaches to characterize a biological system [45]. One of the very successful recent applications is a structural comparison of coronavirus spike proteins that allowed to drastically improve understanding of the origin and evolution of SARS-CoV-2 virus [46].

Novel applications of Artificial Intelligence for proteomics have led to a breakthrough in the field. A novel approach AlphaFold 2 [47, 48] can predict a 3D protein structure based on a 1d amino acid sequence. This achievement is particularly impressive given that the number of ways a protein could theoretically fold before settling into its final 3D structure is estimated to be approximately 10^{300} [49]. AlphaFold 2 predictions provided a boost for the field in many ways, including the advance of understanding of known diseases and drug design [50].

Protein Interactions A protein's biological function depends on various characteristics such as protein sequence and structure, expression profile, post-translational modifications, intracellular localization, interactions with other proteins, and many others [51]. Up to the 21st century, researchers tended to ignore protein interactions and instead studied individual proteins. Later on it became evident that many proteins require interactions with other proteins to carry out their biological function [51].

Protein interactions are usually separated into two types: stable and transient. Stable interactions form a stable protein complex while transient (or temporal) interactions are involved in various cellular processes, including protein modification, transport, folding, signaling, apoptosis, and cell cycling [51]. Depending on an interaction's type and strength, protein interactions can be measured using co-immunoprecipitation (stable or strong interactions), pull-down assays (stable or strong interactions), crosslinking protein interaction analysis (transient or weak interactions), label transfer protein interaction analysis (transient or weak interactions), far-western blot analysis (moderately stable interactions) [51]. Yeast two-hybrid (Y2H) screens are considered to be the most common approach to determine protein interaction due to its scalability and ability to detect transient interactions (although with a limited power due to a high false positive rate) [52]. Another high-throughput method is the tandem affinity purification run in conjunction with mass spectroscopy (TAP-MS). The classic TAP-MS approach was unable to detect transient interactions, however Worthington et al. demonstrated how the use of chemical crosslinking can be applied to enable transient interactions detection [53].

Protein-Protein Interactions (PPIs) aggregated in a single interactome network represent complex relationships between proteins. Researchers often use these networks to predict disease-associated mechanisms based on the guilt-by-association principle, i.e., physically interacting proteins are likely sharing similar functions and participate in shared biological processes [54]. Various databases [55, 56, 57] store PPIs based on different levels of evidence: experimentally validated interactions, literature-based or *de novo* predicted interactions.

Beyond the central dogma of biology

Epigenetics Only 2-3% of DNA is coding for proteins [58] while the remaining 98% are usually referred to as nonprotein coding RNA (ncRNA). The central dogma of molecular biology tells us how proteins are made, but it is missing the explanation of the purpose of that 98% consisting of ncRNA. Moreover, there is an observation that an organism's complexity is not correlated with the number of protein-coding genes; instead, it is correlated with the relative amount of non-protein-coding DNA regions [58]. These regions can regulate gene expression and thus organize the development and maintenance of complex life. They can also control the timing and rates of the protein manufacturing processes.

Epigenetics can be defined as a study of changes in gene expression that occur without a change in DNA sequence and are meiotically or mitotically heritable [59]. These changes occur between the so-called open and closed frames of chromatin conformation, which, respectively, lead to an increase or a shut down of the transcriptional activity of genes [60]. Epigenetic changes largely depend on environmental factors such as stress, pollution, nutrition, tobacco consumption, and many others.

Studies show that many complex diseases such as schizophrenia can be triggered not only by mutations in protein-coding genes but also ncRNA regulation of disease-related risk genes [58]. There are many other complex diseases in which epigenetic mechanisms play a crucial role: cancer, bipolar disorder, systemic lupus erythematosus, autism, and many others [59].

The primary profiling techniques in epigenetics can be roughly separated into two cate-

gories: techniques for DNA methylation profiling and techniques for chromatin accessibility and histone modification profiling [61]. Several mechanisms that regulate gene expression by DNA methylation include proteins that are involved in gene repression, as well as inhibition of the binding of transcription factor(s) to DNA [62]. The most popular technique to measure DNA methylation is the Illumina Methylation EPIC BeadChip Microarray (previously 27k, 450k and now 850k methylation sites) [61]. Chromatin accessibility is the degree to which DNA is physically accessible [63]. The physical accessibility of DNA is a crucial factor for promoters, enhancers, insulators, and other factors to regulate gene expression. Chromatin accessibility is usually measured with the following techniques: Chromatin Immunoprecipitation (ChIP), Digital deoxyribonuclease (DNase), Nucleosome Occupancy and Methylome sequencing (NOMe-seq), Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) and Chromosome Conformation Capture (3C, 4C, 5C, Hi-C) [61].

The following two sections (Metabolomics and Microbiomics) describe two other omics layers that the Central Dogma of Molecular Biology does not cover. Despite their importance and rapidly developing applications, they are not related to my own work. They are provided here to offer a more complete description of the role of different omics layers.

Metabolomics Metabolomics gives another perspective on how a combination of environmental and genetic factors contribute to health and disease. Measurable changes in human metabolites can occur due to complex or monogenic disorders, and these changes can be tissue-specific or have temporal dynamics [64]. Not only genome perturbations lead to changes in the metabolome, but also the perturbation on the transcriptome and proteome level [65]. Many common human diseases have a distinct metabolic "fingerprint" either on the organism level or on the cell level. The most common example of a metabolic disorder is diabetes, but metabolic perturbations can also cause rare and lethal diseases. An example of such a condition is Gaucher's disease that is caused by a lack of enzymes responsible for the accumulation of glucosylceramide into reticuloendothelial cells [66]. This leads to swelling of the lung, spleen, liver, or other organs and has fatal consequences.

Since metabolomics is a relatively young field, substantial effort is currently being made in order to create databases with associations between metabolite structure and function, biomarker-disease associations, mechanisms of actions, and networks [65].

Microbiomics The whole set of microorganisms in the human gut (or any other location) is called microbiota [67]. All the genetic material within the microbiota is called microbiome. Each person has approximately 10-100 trillion symbiotic microbial cells, and thus the human microbiome consists of the genes these cells harbor [67]. The goals of microbiome studies usually include the identification of environmental and genetic factors that contribute to microbiome composition and the influence of microbiome on the health of a host [68]. Microbiota has influence over a large number of conditions such as infectious diseases [69], liver diseases [70], gastrointestinal malignancy [71], metabolic disorders [72], allergic diseases [73] and even psychiatric diseases [74].

2.1.2. Single cell technologies

Single cell sequencing

Traditional NGS sequencing (bulk sequencing) usually considers an average of a mixture of cells and does not allow to differentiate behavior of different cell subpopulations [75]. Therefore, bulk sequencing has a severe limitation in various scenarios, for instance, when only a small population of cells carries a disease-related signal. In contrast to bulk sequencing, single-cell technologies allow to capture over a million cells per sample which can significantly improve understanding of various diseases [76].

Most prevalent, single-cell sequencing includes single-cell DNA sequencing (to discover mutations and CNVs), single-cell RNA sequencing (scRNA-seq), single-cell DNA methylome sequencing, single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq). For every type of data, various analysis methods have been developed.

scRNA-seq analysis usually aims for the following tasks: cell population identification, regulatory network inference, and trajectories identification [76]. Cell population identification is necessary to separate cell groups into clusters that can be annotated with different cell types. This is usually done using reference marker genes that are known to be expressed for certain cell types. Usually, these methods also perform dimensionality reduction, such as PCA [77], UMAP [78], t-SNE [79] or variational autoencoders [80], in order to visualize cell clusters.

Another large family of methods that is relying on scRNA-seq is network inference methods [81, 82]. They analyze gene co-expression to infer regulatory relationships between transcription factors and their targets. This allows to see regulatory mechanisms for various cell types and learn how those can be disturbed by various diseases.

scRNA-seq also allows performing trajectory inference analysis, which describes the temporal evolution of cells. It can infer shapes of trajectories and align cells on those typologies [83, 84].

Single-cell whole-genome sequencing (scWGS) is performed to identify CNVs and single-nucleotide variations (SNVs) [85]. These methods do not differ significantly from bulk analysis but require modifications to address the low coverage of the genome found in scWGS data [86].

Single-cell epigenomics data, similarly to bulk, is used to identify open chromatin and DNA methylation sites. Compared to bulk data analysis, the main challenge is a very low sequencing depth, making it difficult to robustly identify peaks or methylation sites. Mostly, the methods rely on various data aggregation strategies. For instance, the data can be aggregated among small groups of cells, and then the conclusion (peak presence) is verified for each cell individually [87]. Alternatively, genome binning can be used to identify regions with high read counts that can be considered as potential peaks or methylation sites [88].

Single-cell sequencing is applied to study various conditions [75], but it is particularly beneficial in cancer-related studies as cancer tissues are known to be highly heterogeneous [89]. Therefore, an averaged signal of all cells cannot provide the same level of understanding as single-cell technologies. Promising examples of a single-cell technology application in

cancer include a paper published by Bian et al. [90]. Bian et al. demonstrated a relationship of genomic copy number variation, DNA methylation abnormality, and gene expression change on a single cell level during the occurrence and metastasis of human colorectal cancer.

Another example demonstrated by Nguyen et al. [91, 92, 93] when they used single-cell sequencing to understand the early signs of breast cancer and thus provide criteria for early diagnostics and treatment.

Puram et al. [94] also published their findings in head and neck cell carcinoma, where they defined head and neck tumors based on single-cell sequencing. This subtyping allows one to understand the disease, its causes, metastasis spreading and provides new gene targets to block metastasis development.

Several studies [95, 96] provided an immunological map of liver and lung cancer that revealed tumor heterogeneity, tissue characteristics, and drug target gene expression of immune cells. These findings contributed to understanding of the immune microenvironment of liver and lung cancer which is crucial for biomarker discovery and efficient treatment.

Mass Spectrometry based single cell omics

The section above discussed single-cell sequencing technologies, but not all omics data is measured using sequencing. Proteomics and metabolomics play an essential role in a molecular interplay in health and disease but are often measured using MS technologies [97].

MS analysis allows to identify and characterize known and unknown chemical compounds using their molecular weight [98]. Although advancements in MS technologies led to various scientific discoveries in the last two decades, bulk MS suffers from the same main limitation as bulk sequencing: an averaged mixture of various cells has resolution limitations.

In the last years, MS technologies allowed measurement of spatial proteomics and metabolomics on a single cell scale [99]. For the metabolomics field, it allowed precise measurement of small molecules, lipids, and drugs [98]. Studies demonstrate a crucial role of single-cell metabolomics to understand and prevent viral outbreaks via unveiling the host-virus interactions [100]. Many other applications of single-cell metabolomics are surely coming in the future, but the field still has many technological and computational challenges to overcome [99].

Single-cell proteomics holds an immense promise to revolutionize our understanding of diseases by providing a possibility to identify pathological mechanisms in heterogeneous tissues [101], pinpointing microenvironmental factors that promote or inhibit tumor growth [102] and determining novel cellular subpopulations [103] or developmental trajectories [104, 105]. Technologically, single-cell proteomics is challenging due to absence of protein equivalent to PCR amplification of DNA [106]. Nevertheless, more than 1000 protein groups can now be reliably profiled from single cells, and more than 6000 protein groups can be profiled from samples consisting of just a few hundred cells [105].

Cytometry by time of flight, or CyTOF is another MS-based technique that allows to profile single cells by chemically attaching to them heavy metal isotopes and then use an atomic mass cytometer to detect the time-of-flight (TOF) of each metal [107]. As a part of the Ph.D., I was involved in a project where we used CyTOF to measure platelet marker expression

2. Background

in response to COVID-19 infection [108], as well as after receiving the vaccination against COVID-19 [109]. The studies suggested a pro-thrombotic phenotype in COVID-19 patients and then verified a lack of it for patients that received a vaccine against COVID-19.

2.2. Molecular data integration and interpretation

2.2.1. Goals

Diseases affect various molecular layers of the human body. Independent analysis of each molecular layer usually provides researchers with a list of differences associated with a phenotype in question. These differences can be used as biomarkers, but usage of only one omics data type on its own is risky as we might be able to observe reactive processes rather than causative ones [110]. Multi-omics approaches have the potential to discover not only the consequences of a condition but its whole complex molecular mechanism. Figure 2.2 shows all described molecular levels and also emphasizes how additional insights about the complex nature of a disease can be acquired by combining different omics layers. Learning from these insights is critical if we aim to advance our understanding. Moreover, a single omics dataset can be subject to noise and biases, while incorporating different layers allows for more robust results.

2.2.2. Data repositories

Generation of multi-omics data for a fixed set of samples allows connecting disease perturbation at different molecular levels. Availability of public multi-omics data is of paramount importance as it allows the community to benchmark novel algorithms based on well-studied datasets. Moreover, multi omics-data collection and processing can be costly. Luckily, several large consortiums provide well-curated public datasets that can be freely used for data analysis and algorithms benchmarking.

Most of the large data consortiums are focused on cancer due to its prevalence, lethality, and complex causes that combine environmental and genetic reasons. Several largest consortiums that aim to provide the data publicly will be discussed.

The Cancer Genome Atlas The Cancer Genome Atlas (TCGA; <https://cancergenome.nih.gov/>) aims to collect, analyze and interpret RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and reverse phase protein arrays (RPPA) data from 33 different types of cancer. It is one of the largest data repositories with over 20 000 individual tumor samples. Additionally, TCGA data can be visualized by various independent platforms [111, 112, 113]. Corresponding proteomics data for TCGA samples are available from Clinical Proteomic Tumor Analysis Consortium (CPTAC) (<https://cptac-data-portal.georgetown.edu/cptacPublic/>).

Cancer Cell Line Encyclopedia Cancer Cell Line Encyclopedia (CCLE; <https://portals.broadinstitute.org/ccle>) is a project that aims to provide well-defined and annotated, large-scale cancer cell lines models. CCLE provides metabolomics, proteomics, epigenetic and histone modification, RNA expression, and genetic data for various cancer cell lines. Visualization of the data is possible using DepMap portal [114].

Molecular Taxonomy of Breast Cancer International Consortium Molecular Taxonomy of Breast Cancer International Consortium (METABRIC; <http://molonc.bccrc.ca/aparicio-lab/research/metabric/>) is a project that focuses on the classification of breast tumors into novel data-driven subtypes based on multi-omics data. METABRIC has data about clinical traits, expression, single-nucleotide polymorphism (SNP), and CNV data derived from breast tumors. The data can be explored and visualized using cBioPortal [115].

Therapeutically Applicable Research To Generate Effective Treatments (TARGET; <https://ocg.cancer.gov/programs/target>) aims to provide various multi-omics data for pediatric cancers. TARGET has public and protected datasets of 24 molecular types of cancer. The data include microarray gene expression, CNV, methylation, miRNA data, whole genome, exome, and mRNA sequencing data. The data can also be visualized using Xena browser [113].

2.2.3. Tools

Analysis of single omics data provides valuable insights into ongoing biological processes, but it might not allow capturing a cross-talk between different molecular layers [116]. Different data types can represent several properties of a system, and an independent analysis would not allow capturing a common signal [117]. On the other hand, conjoint multi-omics analysis in a single algorithmic framework brings more interpretability to the analysis but indeed represents a challenging task. The challenges can be split into the following categories:

- Large data volumes: practically all molecular data is highly dimensional, and even a single omics data matrix can take gigabytes of storage [118]. The combination of multiple data sources makes computations, even more time and memory-consuming.
- Different data features: multi-omics technologies aim to measure different molecular entities, which means that reducing to a simple feature \times samples matrix is not always possible. Features are often not easily mapped as they might be genes (genomics/transcriptomics) or genetic coordinates (ATAC-seq, ChIP-seq) or CPGs (methylation) or miRNAs or other molecular features [119].
- Various data distributions: molecular data can come from various data distributions; it can be continuous (gene expression), discrete (CNV), and even binary (SNPs). Combining the data such that the most variable data types do not overshadow the least variable ones might be a very challenging task [119].
- Results interpretation: biological interpretation of algorithmic results is often challenging even for single omics analysis. Interpretation of multi-omics analysis results requires a deep understanding of undergoing biological processes and conditions [120].

Despite the mentioned above challenges, multi-omics data integration still attracts researchers that develop algorithms to unravel the mechanistic aspects of the information flow in a cell. Subramanian et al. [121] collected information about various multi-omics approaches and separated them into three categories based on their purpose: disease subtyping, disease

insights, and biomarker prediction. Subramanian et al. described 34 multi-omics methods most of which aggregate CNV, gene expression, DNA methylation, and sometimes also miRNA and protein expression.

Methods covered by Subramanian et al. are still primarily not developed for combining single-cell/bulk data with epigenomics data. To fill this gap, the third publication of this Dissertation [11] provides a review of multi-omics methods that employ epigenetics data on bulk and single-cell level.

Lastly, another aspect that was not previously covered is microbiome-metabolome relationships. Several studies demonstrated influence of microbiome on metabolome and general health of a host [122, 123]. In particular, MiMeNET [124] is a neural networks based approach that provides insights into the causes of dysregulations in disease by analyzing microbe-metabolite interaction. Furthermore, to study potential host-microbe interplay, methods like AMON (Annotation of Metabolite Origins via Networks) help to evaluate pathway enrichment of host versus microbial metabolites [125].

2.3. Artificial intelligence in biology

Artificial intelligence is often portrayed as a set of technologies that mimic natural human or animal intelligence. By intelligence we often assume problem-solving skills, abilities to learn from mistakes, and the ability to generalize previous experience to new problems [126].

The growth of biological data was nearly exponential starting from the end of the 20th century [127]. High throughput technologies made it practically impossible to analyze and interpret molecular data without advanced algorithmic techniques [128]. One of the first computational problems that arose from the development of sequencing techniques was multiple sequence alignment [129]. It concerns pairs of highly similar sequences (i.e., originated from the same genome) that deviate in some locations due to mutations. The goal is to align sequences such that the maximal number of base pairs is matched. Multiple sequence alignment is an example of an implicit hitting set problem, which is classified as a polynomial-time problem, meaning that its running time is upper bounded by a polynomial expression in the size of input of the algorithm [129]. Polynomial run-time usually means that searching for an exact solution is not feasible within a reasonable time frame. Therefore, for this task (and many other biological problems), so-called *heuristic* approaches must be used. Heuristic algorithms allow finding nearly optimal solutions relatively fast due to tuning an algorithm to expected characteristics of the data [129].

2.3.1. Classic heuristic approaches

Heuristic algorithms are often used for NP-hard (not computable in polynomial time) problems when an exact solution extraction is not feasible under the given constraints [130]. Therefore, heuristic approaches attempt to limit the solution search space by incorporating prior information about the nature of a problem. The main limitation of heuristic approaches is the impossibility of proving that the solution is optimal. Therefore the "goodness" of a solution largely depends on the correctness of initial assumptions. For validation of heuristic approaches, researchers often use bootstrap-based statistical techniques to verify that a produced solution is significantly better than a random one.

According to Blum and Roli [131], heuristic algorithms can be split into the following categories:

- Nature-inspired vs. non-nature-inspired: some heuristic algorithms were inspired by biological processes; for instance, genetic algorithms are inspired by natural selection, ant colony optimization (ACO) algorithms are inspired by the behavior of ants in a search for the shortest path to food. Other methods are not nature-inspired, e.g., simulated annealing is inspired by annealing in metallurgy. Simulated annealing attempts to decide at every time point which movements will lead the system to the lowest energy state.
- Population-based vs. single point search: based on the number of solutions that develop over an algorithm run. Some methods, such as local search, attempt to find a

development trajectory for a single solution. Other algorithms like ACO initialize a set of solutions and iterate until their convergence to a single nearly optimal solution.

- Dynamic vs. static objective function: based on the ability of an algorithm to change the objective function during the optimization process. This behavior might be desirable to avoid stagnation at a local minimum.
- One vs. various neighborhood structures: based on usage of a single search space or an ability to switch between several ones.
- Memory usage vs. memory-less methods: based on how much information an algorithm can utilize from previous iterations to make the next step. This can vary from simple Markov models when only one last step is remembered to a complex long-term memory that is often possible when using neural networks.

First heuristic algorithms in computational biology were developed for sequence analysis (FASTA - Lipman and Pearson 1985 [132], Clustal - Higgins et al. 1988 [133], BLAST - Altschul et al. 1990 [134]).

Another common domain for heuristic algorithms is the analysis of protein-protein interaction (PPI) networks [129]. A common assumption behind PPI networks is that the set of proteins with a similar function is expected to induce a densely connected subgraph. In contrast, the set of proteins occurring in a signal transduction pathway is expected to induce a simple path graph with high-weighted edges [129, 135, 136, 137]. PPI analysis can be framed in multiple optimization problems; here are several examples:

- Maximum clique problem: find the largest fully connected set of proteins.
- Dense subgraph problem: find the largest densely connected set of proteins.
- Heavy path problem: find a simple path of a given length such that edges have the highest possible weight.

There are many other applications of heuristic algorithms in modern computational biology, including one of the papers published as a part of the dissertation (BiCoN [9]) — more details on it will be provided in the Methods section.

2.3.2. Machine learning

Development of Machine Learning

Artificial intelligence (AI), Machine Learning (ML), Artificial Neural Networks (ANNs), and Deep Learning (DL) are terms that are often used interchangeably. In fact, the terms can be displayed as a hierarchy as shown in the Figure 2.4. To be more precise, distinguishing AI from ML, AI usually presumes any kind of *intelligent* behavior that a machine can resemble. At the same time, ML algorithms are supposed to learn how to be *intelligent* based on the given data and not explicitly given rules [138].

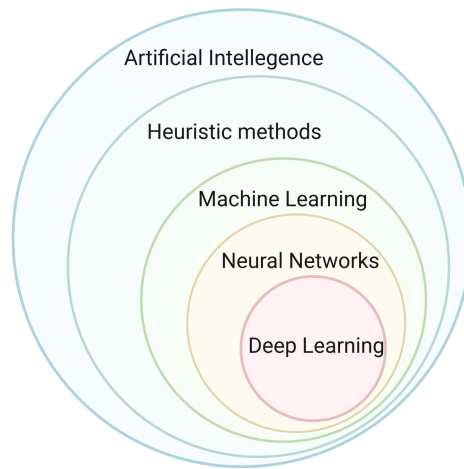


Figure 2.3.: Hierarchy of the discussed Artificial Intelligence (AI) related terms. AI is an umbrella category that covers all intelligent machine behavior. Heuristic methods assume the usage of environmental information in order to speed up computations. Machine Learning (ML) usually implies learning from the data. Neural Networks (NN) are ML algorithms that use a particular architecture that resembles a human brain to learn complex relationships in the data. Finally, Deep Learning is a type of NN with several hidden layers that make a network capable of learning very complex patterns in data.

Neural networks The beginning of neural networks is commonly attributed to the development of the perceptron in 1957 [139]. Perceptron was the first neural network that consists of input values, weights, biases, summation, and activation function [139]. At first, the input values are multiplied by their weights (random at first iteration), then all these values are summed up together, and the sum is scaled to the necessary range by an activation function. After the prediction is made, the deviation of the prediction from a target class is used to readjust weights. A very simple version with no biases is shown in Algorithm 1.

When the perceptron was developed, it was considered to be a groundbreaking achievement, but soon several limitations of the method were acknowledged [140]. First, the perceptron only worked for a binary classification case, but even more restrictive was the fact that the perceptron only was capable of separating linearly separable classes. The latter was a direct consequence of the fact that the perceptron had only one set of weights (i.e., one layer of weights between input and output). Therefore only linearly separable classes could have been handled accurately. This made the perceptron unusable in many complex tasks such as object recognition, until multilayer perceptron (MLP), was developed later in 1960, which allowed learning even non-linearly separable classes stratification [140]. MLP took advantage of organizing single perceptrons into multiple layers to address non-linearity in the data and recognize complex patterns. Moreover, different activation functions were used to scale the output for binary classification and multi-class classification and regression (prediction of a continuous variable). Another essential step was the introduction of backpropagation that

Algorithm 1 Perceptron Learning Algorithm

```
 $P \leftarrow$  inputs with label 1  
 $N \leftarrow$  inputs with label 0  
Initialize  $w$  randomly  
while no convergence do  
  Pick random  $x \in P \cup N$   
  if  $x \in P$  and  $w \cdot x < 0$  then  
     $w = w + x$   
  end if  
  if  $x \in N$  and  $w \cdot x \geq 0$  then  
     $w = w - x$   
  end if  
end while  
the algorithm converges when no more improvements to predictions can be made
```

allowed to adjust the layers of neurons by correcting them with respect to their previous mistakes.

Even though backpropagation is still used to train Artificial Neural Networks (ANN), back in 1970th the field was stagnating until massive data sets and computational power became available to researchers [141]. Only in 1990th ANN research started flourishing again, discovering novel algorithms and applications such as speech recognition, natural language processing, image analysis, and computational biology.

Statistics Many common machine learning methods originate from discoveries made by statisticians and mathematicians starting from the 18th century. Thus 1763 Bayes' Theorem was published as a part of a book "*An essay towards solving a problem in the doctrine of chances*" [142], and it allowed to estimate a probability of an event based on prior knowledge about other conditions, relevant for the event. Bayes' Theorem laid the foundation for many modern machine learning techniques such as Bayes classifier and Bayesian neural networks.

Later, in 1805 Adrien-Marie Legendre described the least square method, that described a line fitting through a set of observation, allowing to make predictions about specific variable behavior, based on a set of conditions [143]. This is a widely used technique for the least square regression, commonly used due to its interpretability and speed.

In 1967, the nearest neighbor algorithm was published, and it allowed to assign an unclassified sample point to a class based on its similarity to a set of previously classified points [144]. The nearest neighbor algorithm was the first step towards pattern recognition. It resulted in many modern machine learning techniques such as the nearest neighbor classification and the nearest neighbor clustering for pattern discovery with unknown classes.

Many other famous and widely used approaches that shaped modern machine learning have been developed in the 90s, such as the Random Forest algorithm [145], Support-Vector Machines [146] and Boosting methods [147]. These methods significantly contributed to

classification for complex, multidimensional, and non-linearly separable data.

Modern approaches and applications in Computational Biology

Modern machine learning algorithms are usually split into three categories: supervised learning (or semi-supervised), unsupervised learning, and reinforcement learning [148] (Figure 2.4).

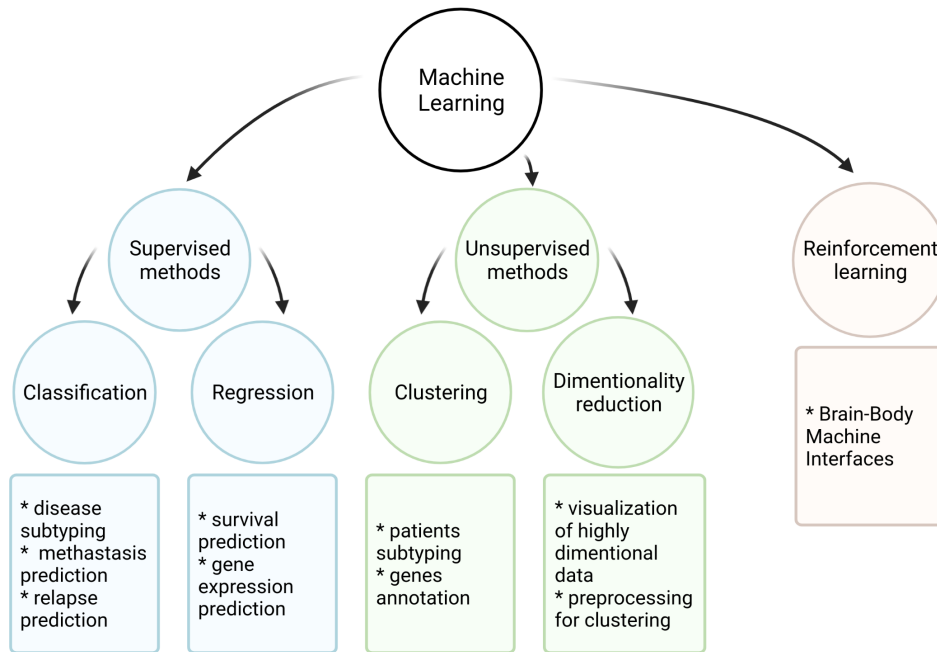


Figure 2.4.: Main categories of machine learning approaches and their application in biomedicine

Supervised learning Supervised learning presumes a prediction task which can be a classification or regression [148]. In computational biology, classification might be applied to predict risk and diagnosis [149], early or late-stage relapse [150], disease subtyping [151] and many other applications. Regression is usually used to predict continuous variables such as gene expression [152] or survival time [153].

Unsupervised learning Unsupervised learning is applied when labels can not be used or not needed for the analysis [148]. Unsupervised learning is often applied in the data exploration stage when researchers aim to get more insights and understanding of their high-dimensional data. Dimensionality reduction techniques such as PCA [154], UMAP [155], or t-SNE [156] are used to represent high dimensional data in a low dimensional space in order to visualize the data structure [157]. Clustering is used to separate samples in a

data-driven way (without labels). This is useful for cell annotation [158], patient stratification [159] and gene clustering [160].

Reinforcement learning Reinforcement learning forms an interplay between an AI and a real or simulated environment [148]. Certain interactions get reinforced, and others "punished" based on a user-established "reward policy". Reinforcement learning is commonly used in computer vision [161], portfolio optimization [162], traffic control [163] and many other settings. Since reinforcement learning usually presumes a continuous input data stream, its application in computational biology is limited. The most common cases are Brain-Machine Interfaces [164] and recommendations based on Electronic Health Records [165].

2.3.3. Artificial neural networks in biomedicine

Separation of the essential information from all available information is necessary for any decision-making process. In biomedicine, this task can be executed manually in some cases (analysis of radiology images, for example). However, for high dimensional numerical data, annotation with a bare eye is not achievable. Not to mention that even when a manual annotation is feasible, it is still a very resource-demanding task.

The ability of ANNs to extract informative features from complex, noisy, high-dimensional data is one of the main reasons why ANNs are so beneficial for the biomedical field [166]. With numerous applications to medical image analysis, genomic sequencing and gene expression analysis, protein structure prediction, and many other fields, ANNs are quickly becoming one of the main tools for biomedical data analysis.

ANNs can be separated into different categories based on the architecture they are using and the type of input data:

- Convolutional Neural Networks (CNN): CNNs can be perceived as a form of regularized MLPs. By design, MLP are fully connected and therefore quite sensitive to noise and prone to overfitting (do not generalize well on previously unseen data) [167]. CNNs exploit hierarchical patterns in the data and try to split them into smaller patterns using filtering approaches. CNNs are shift and rotation invariant, and this makes them extremely useful for extracting features from imaging data [168]. An example of such an application can be the detection of brain hemorrhage based on cross-sectional CT image [169].
- Autoencoders (AE): AEs are neural networks that aim to compress an input and then restore it as close to the original as possible [170]. The compressed representation is commonly used for dimensionality reduction of images, sequences, and numerical data. The restored output can be used for imputation of missing values and denoising. Concrete examples include denoising of medical images [171], biomedical image feature extraction [172] and imputation of missing values in single-cell RNA-seq datasets [171].
- Recurrent neural networks (RNN): RNNs are usually applied to data with temporal or sequential structure [173]. While MLP does not consider the order in which a network

processes samples, RNNs assume that output for a currently considered sample is directly related to the output of a previous sample. RNNs vary in memory capacity, e.g., the number of previous samples that influence the current sample. In the biomedical field, RNNs are often used to analyze unstructured natural language texts in order to provide patients with automated recommendations [174] or to extract medical events from Electronic Health Records [175]. Another common application of RNNs concerns sequencing data for various tasks, including identification of short viral sequences from metagenomes [176] and detection of DNA base modifications [177].

- **Graph Neural Networks (GNN):** GNNs are applied to graphs to exploit topological relationships between entities. GNNs are used to predict missing links, classify nodes or classify graphs themselves. Missing link prediction was successfully shown for miRNAs and diseases association prediction [178] or for scRNA-seq based cell interaction prediction [179] and many other applications [180]. An example of graph classification can be a framework from Gligorijević et al. for protein function prediction [181].
- **Transformers:** Transformers were developed to improve the performance of RNNs on sequence-to-sequence prediction and to replace the memory mechanism with an attention mechanism. ANN uses attention mechanism to remember which part of the sequential information is important. Thus, transformers embed input and output sequences, apply attention mechanism to identify the important fragments and only then actually use classic MLP layers to predict the output. Transformers are also used in bioinformatics, for instance, to predict genome-wide regulatory elements based on up and downstream nucleotide contexts [182].

2.3.4. Common challenges in machine learning

Machine learning (ML) is a complex discipline that requires knowledge of statistics, mathematics, programming, and an in-depth understanding of the application domain. ML research is often criticized for its complexity and lack of transparency to people outside of the field and even other researchers [183, 184]. The main challenges that ML researchers are facing are:

- **Poor quality of data:** to make reliable predictions, the model must be trained on data representative of the modeled system. If the data is subject to biases and errors, the same biases will be propagated to the model.
- **Underfitting of training data:** underfitting means that the model cannot understand the pattern in the data. Figure 2.5 A demonstrates an example of underfitting when the true relationship between x and y is quadratic, yet a linear function was fitted.
- **Lack of training data:** the relationship can not be modeled reliably without enough observations (Figure 2.5 B).
- **Overfitting of training data:** overfitting happens when the model is caught up on minor deviations in the data and unable to make reliable predictions for unseen data

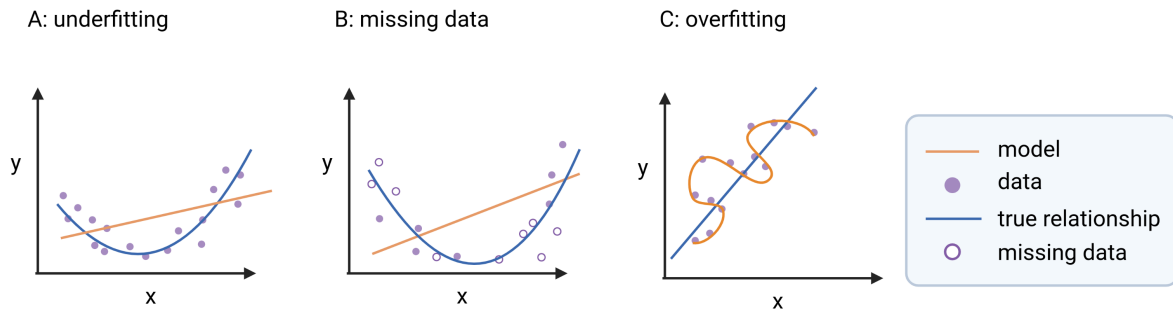


Figure 2.5.: Common issues with ML models. A: underfitting means that the complexity of the data is underestimated; B: missing data leads to an inability of a model to understand patterns in the data; C: overfitting means that the model is unable to generalize.

(Figure 2.5 C). This situation is usually characterized by excellent performance on the training data and very bad performance on new data, i.e., the model is unable to generalize the learned information and memorizes the training data instead.

- Resource demanding: while simple statistical models are easy to implement and use, deep learning methods might require a lot of computational resources, making the training process extremely time and money consuming.

An additional challenge that is particularly relevant for the biomedical field is a so-called "*curse of dimensionality*", i.e., the inability of ML models to perform adequately in the presence of multi-dimensional data [185]. In the biomedical field, one sample is often characterized by thousands of genes or proteins, metabolites, or other molecular entities usually referred to as "features". When the number of features is equal to or larger than a number of samples, many models can not obtain a robust solution and avoid overfitting. Multiple strategies can be applied to combat this issue, such as feature selection, data augmentation, or the use of methods that do not exploit all features simultaneously (like random forest or gradient boosting).

2.4. Network and Systems Medicine

Two philosophical positions: *reductionism* and *holism* have been subjects to heated arguments in the scientific community for at least 30 years [186]. *Reductionism* presumes that complex behavior can be explained by examining simpler behavior patterns of separate components. *Holism*, on the other hand, presumes that a system's complex behavior is more than just a sum of its components, and individual components analysis might not have any value. Even though *reductionistic* approaches lead to many breakthroughs in molecular biology in the 20th century, an increasing number of novel research findings have shown that the molecular-reductionist approach cannot fully explain the complexity of biological systems [187]. Network and Systems Medicine offer *holistic* approach to human health and gained popularity in the 21st century.

Network medicine is a field that applies graph theory (or network science) to study complex biological relationships. The term was first introduced by Albert-László Barabási. Barabási describes several examples of "network thinking" that led to novel scientific insights [188]. One of the examples discusses that a social network (e.g., friends, family members, neighbors, etc.) has a larger influence on the probability of becoming obese compared to a genetic influence [189]. The reconstruction of the social network allowed to conclude that chances of obesity for a friend of an obese person increase by 171%. The authors also studied the effect of having an obese sibling, neighbor, or friend of a friend and concluded that obesity forms dense network communities. The role of a person's social network structure is even more prominent when it comes to infectious diseases such as influenza and HIV.

2.4.1. Molecular networks

Social networks are representing a very high level of possible interactions. However, we can increase the focus down to disease networks [190], organ networks [191] and ultimately down to molecular networks.

The knowledge about genes that were perturbed by disease is not enough to understand the mechanism of the disease. On the other hand, networks allow understanding how the perturbed genes are related to each other and other genes and thus provide a more holistic understanding of a molecular system. [188].

A simplified overview of molecular networks is shown in Figure 2.6. Apart from the edge types shown in Figure 2.6, various other connections are possible (enhancer interactions, metabolic feedback, etc.), and this makes network biology an extremely heterogeneous and complex subject [192].

The following types of networks are often researched in the context of network biology:

- DNA interaction networks: the dogmatic view on transcription states is that gene promoters drive transcriptional initiation, while most enhancers are localized in the proximity of gene promoters and are rarely beyond gene boundaries [193]. The most recent research shows that many DNA binding transcription factors bind to genomic loci that can be far away from their regulated genes [194]. The DNA looping model

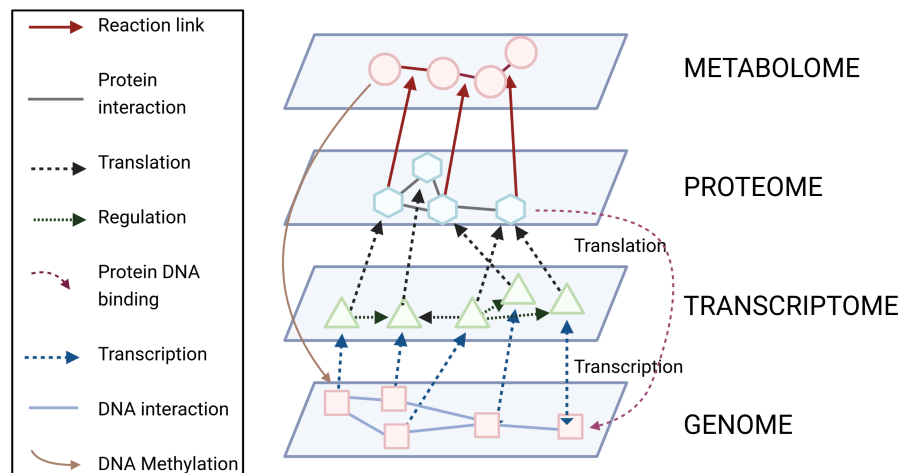


Figure 2.6.: A simplified overview of molecular interactions on different cellular levels.

was thus proposed to explain long-distance enhancer-promoter interactions. Another perspective on DNA-DNA interaction include SNP-SNP interaction - an effect when a single SNP does not affect a phenotype, but a combination of SNPs might have a significant effect [195]

- **Regulatory networks:** Regulatory networks usually consist of genes and transcription factors (TF) where TFs can activate or inhibit gene expression. TFs are also subject to regulation. Cellular function is controlled by gene regulation, and therefore regulatory networks have a remarkable explanatory power for analysis of complex human traits [196].
- **Protein-Protein Interaction networks:** PPI networks depict interactions between proteins based on physical contacts. The interaction might include electrostatic forces, hydrogen bonding, and the hydrophobic effect [197]. PPI network analysis is widely used in interpretation of phenotype-relevant gene sets, as it allows to map genes in question onto the interactome and extract an entire interacting mechanism using guilt-by-association principle [198].
- **Metabolic networks** consist of small molecules (metabolites) and enzymes (proteins) where enzymes catalyze biochemical reactions [199]. Metabolic networks show all biochemical reactions that enzymes can catalyze in a cell. Metabolic pathways and regulatory interactions can be mapped onto the networks.
- **Association networks:** Association networks can model any kind of relationship based on similarity measures and not direct evidence of a physical connection. Examples of association networks include gene co-expression networks and protein similarity networks [192].

2.4.2. De novo endophenotyping

Two important terms that are crucial for this section are an *endophenotype* and a biomarker. Endophenotype is a stable phenotype with a clear genetic connection [200], while a biomarker is "an indicator of medical state observed from outside the patient" [201]. Many studies are

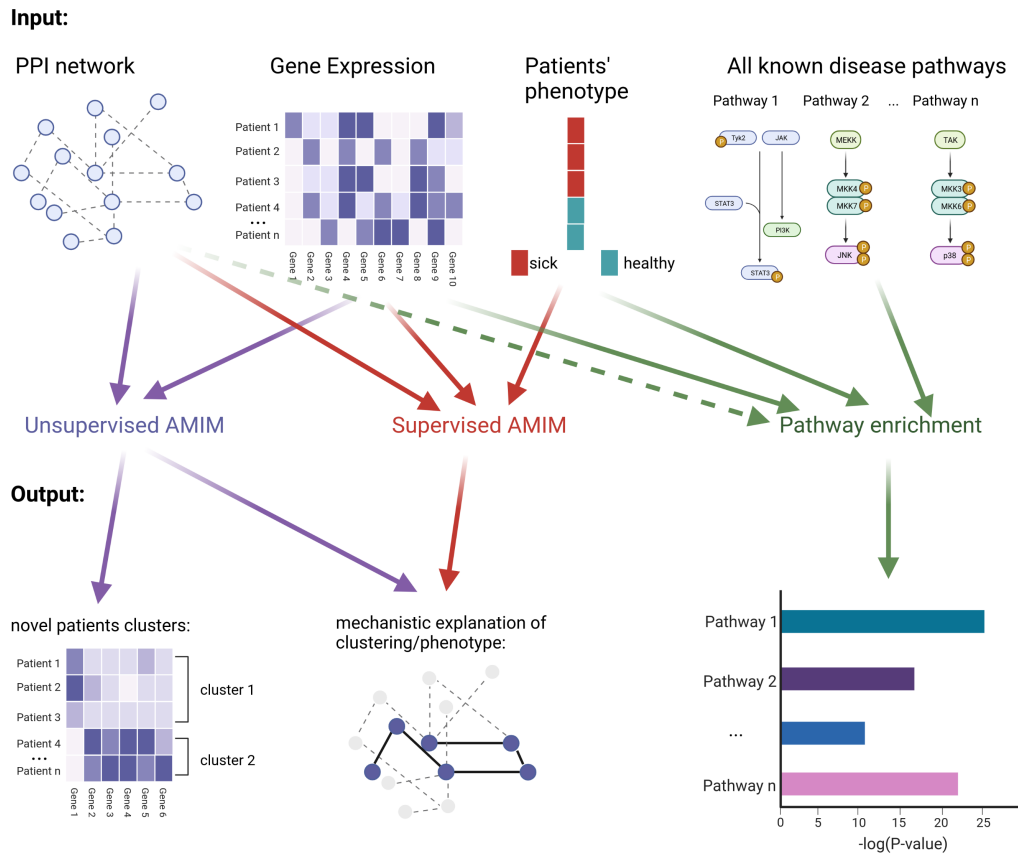


Figure 2.7.: Active Module Identification Methods (AMIMs) and pathway-based analysis tools stratification based on input and output options (solid lines). Dashed lines indicate optional input.

focused on finding stable biomarkers that allow stratifying patients into different disease subtypes, different survival probability groups, response to treatment, and other characteristics [202, 203, 204]. A set of genes that can serve as biomarkers of a particular process is commonly referred to as gene signatures. Despite a vast promise of gene signatures discovery for high throughput molecular data, they can be inconsistent across data sets [205]. Moreover, some studies show that random signatures are capable of providing same-quality patient stratification as published gene signatures [206]. This effect can be explained by the fact that some phenotypes might lead to immense changes in the transcriptome [207]. Therefore, signal-carrying genes suitable for patient stratification might not be directly related to a

phenotype itself.

In order to derive more reliable signatures, many researchers try to integrate prior biological information and increase the meaningfulness of the results. As a part of the PhD project, I also performed a review of methods that approach to solve this problem by integrating gene expression/methylation/CNV and other phenotype-specific data with prior information in a form of a network [208]. The methods were separated into two groups: those that attempt to use *known* phenotype-related pathways (i), and those that use PPI network as a source of prior information (ii). The first approach (i) uses known disease-associated pathways as features. This analysis is usually limited to only known pathways, and therefore, it does not allow the extraction of novel disease-related pathways. The second set of approaches (ii) here and further will be referred to as Active Module Identification Methods (AMIMs). AMIMs have a clear advantage over pathway-based methods: they consider the whole interactome as a possible solution space and, therefore, aim to discover functionally connected genes relevant to the phenotype in question. Numerous studies showed how AMIMs allow gaining valuable biological insights in various conditions such as diabetes mellitus [209], chronic obstructive pulmonary disease, and idiopathic pulmonary fibrosis [210], asthma, and many others [211].

2.4.3. Active Module Identification

AMIMs aim to extract condition-specific modules from large molecular networks. Over the years, a plethora of AMIMs was developed [212, 213, 214, 215, 216]. Since AMIMs usually attempt to solve an NP-hard task, they require heuristic algorithms to reduce the runtime. Batra et al. [8] have stratified the methods based on the employed algorithmic strategy: (1) aggregate score optimization methods, (2) module cover approaches, (3) score propagation approaches, and (4) clustering-based approaches. Aggregate score optimization methods aim to map predefined activity scores (such as fold change, for instance) to nodes or edges and retrieve a subnetwork with the highest scores. Module cover approaches employ a statistical test to find a relevant set of genes and then use various algorithmic techniques to extract the most relevant subnetwork. Score propagation approaches also use predefined scores for nodes/edges, but then they defuse them through the network topology using random walk procedures or diffusion-flow. Then the subnetwork with the highest weights is extracted. Clustering-based approaches are based on clustering of a network or cuts into densely connected components.

Evaluation of AMIMs is very challenging due to a lack of the gold standard. Batra et al. [8] developed an artificial gold standard for methods evaluation, but naturally, data simulation procedures are limited by our current understanding of biological processes. To perform an objective evaluation, we require a complete gene set that is certainly known to be associated with a specific condition. While molecular pathways such as KEGG [217, 218, 219] are often used to evaluate AMIMs, it is important to acknowledge that they are not complete.

Supervised and unsupervised approaches

Active module identification is often perceived as a supervised analysis task where patients' omics data represents a feature matrix, PPI network plays the role of regularizer (by restricting the search space), and a phenotype is the target variable. For example, Grand Forest approach [220] is using classic supervised learning technique random forest [221] (RF) where tree nodes represent the genes (restricted by PPIs) and RF aims to select a set of nodes that can predict the phenotype of patients in the best possible way. Supervised approaches like Grand Forest are widely used by the community and discussed in the previously mentioned review paper by Lazareva et al. [208]. A limitation of those methods is that phenotypic information might be incomplete, missing, or not supported by the data. The described methods do not allow data-driven patient stratification, which can also be perceived as simultaneous unsupervised analysis of patients' omics data and PPI networks. Unsupervised active model identification has the potential to discover novel patients subgroups along with the molecular mechanism that explains the stratification.

In Figure 3.2, a clear difference is demonstrated between unsupervised AMIMs, supervised AMIMs and pathway-based methods.

2.5. Objective

The section provided the necessary background to understand all needed components for Systems Medicine approaches for disease module mining. The introduction to molecular biology section described what genetic information flow (Figure 2.1) is and what molecular levels are observed in the flow (Figure 2.2). Next, I discussed the connection between molecular layers and how they can be modeled as a complex network (Figure 2.6) and examined computational methods that are essential for data analysis and integration. Finally, the Active Module Identification task was discussed to elaborate on why it is essential and what are the main roadblocks in the process.

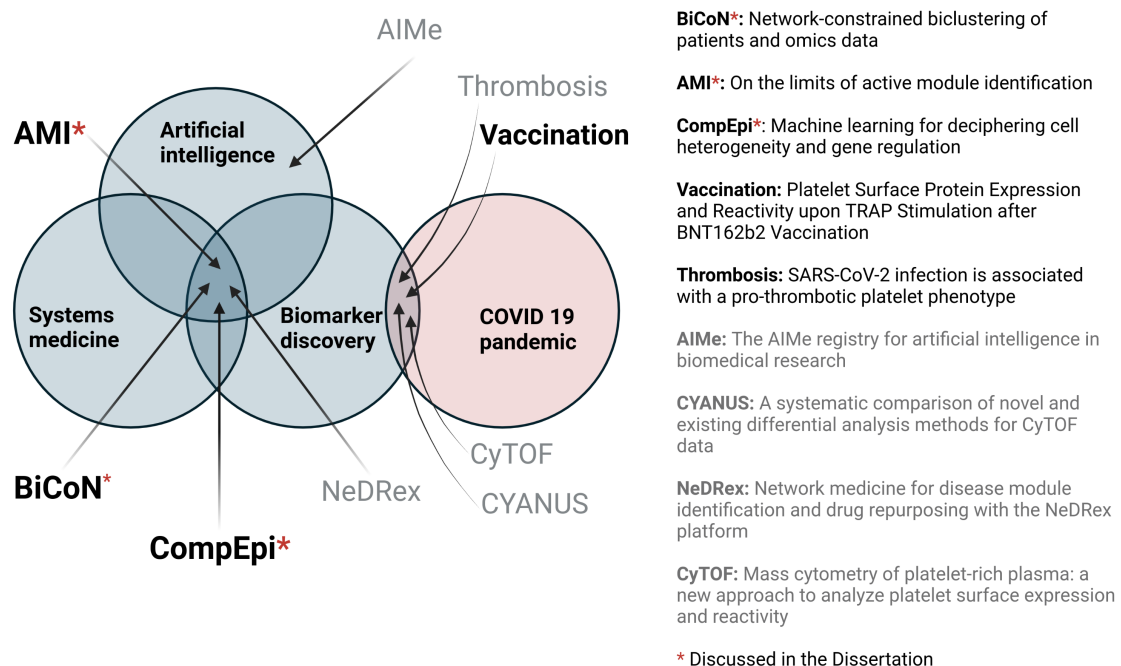


Figure 2.8.: Overview of published peer-reviewed research. Papers marked with a red star are first author research articles described and provided in the Dissertation. Papers are written with black font, but without a star, are the first author papers that will not be further discussed. Papers written in a grey font are not first-author papers.

The main objective of the Dissertation is to bridge Artificial Intelligence with a Systems Biology perspective for robust and meaningful disease modules identification based on PPI networks and transcriptomics data. Therefore, three papers are presented (section 4) that attempts to quantify the role of biological networks for robust patients stratification and disease modules extraction. The first publication - an unsupervised AMIM BiCoN that simultaneously clusters patients and discovers molecular mechanisms by aggregating patient specific omics data with PPI networks [9]. Second, inspired by the promising results of the first approach, an investigation of PPI networks' quantitative properties was made to understand the biological value of general AMIMs results [10]. The third publication reviews

machine learning-based methods capable of using single-cell epigenomics data to increase the resolution of multi-omics analysis [11].

In the Figure 2.8 the contribution of the conducted research is summarized. All produced papers can be separated into three categories: first author papers with full text provided in the Dissertation section; first author papers that are not fully discussed in the Dissertation, other papers written during the Ph.D. that are still relevant for the discussed topic. Unplanned research relevant to the COVID-19 pandemic is also mentioned since two years of the Ph.D. project happened during the global pandemic. Certain adjustments to the initial plan have to be made to contribute to understanding of hypercoagulation risks during the infection and after the vaccination. This research will not be discussed in further detail but available in open access [108, 109]. The full list of published papers is given in chapter 5.4.

3. General Methods

3.1. BiCoN algorithmic framework description

The main algorithmic framework of the first publication [9] is based on Ant Colony Optimization algorithm, applied to a heterogeneous graph. The graph contains two types of nodes: gene nodes and patient nodes; and two types of edges: gene-to-patients (through expression) and gene-to-gene (through protein interaction). The main goal of the algorithm is to find 2 (n in a general case) subnetworks that maximize expression differences between clusters of patients. The framework is summarized in Figure 3.1.

3.1.1. Metaheuristic approach

Heuristic algorithms are often applied to find subnetworks with certain qualities in a large graph when the search space is too large to enumerate all possible options. As discussed in the Introduction (subsection 2.3.1), a plethora of heuristic methods have been developed to address computationally complex tasks. Thus, a metaheuristic framework has been developed that uses Ant Colony Optimization for the search space exploration and Local Search to ensure local optimality of the final solution.

Ant Colony Optimization

Ant Colony Optimization (ACO) is a nature-inspired framework often used when a problem can be reduced to finding an optimal path in a graph [222]. As we aim to discover gene subnetworks optimal for patient clustering, ACO was a natural choice for the task.

The ACO algorithm in its classic form usually starts with computation of the initial transition probability matrix, which is defined as a combination of heuristic information (prior information defined by a user) and a pheromone matrix which is random at the beginning. Then the following steps are executed:

1. A set of semi-random solutions are computed based on the transition probability matrix. These solutions are further referred to as "ant walks";
2. Each of the ants walks is evaluated with respect to an objective function;
3. The pheromone matrix is recomputed based on the scores of the ant walks;
4. The transition matrix is also recomputed with respect to the new pheromone matrix, such that the most successful transitions are promoted and the least successful are downgraded.

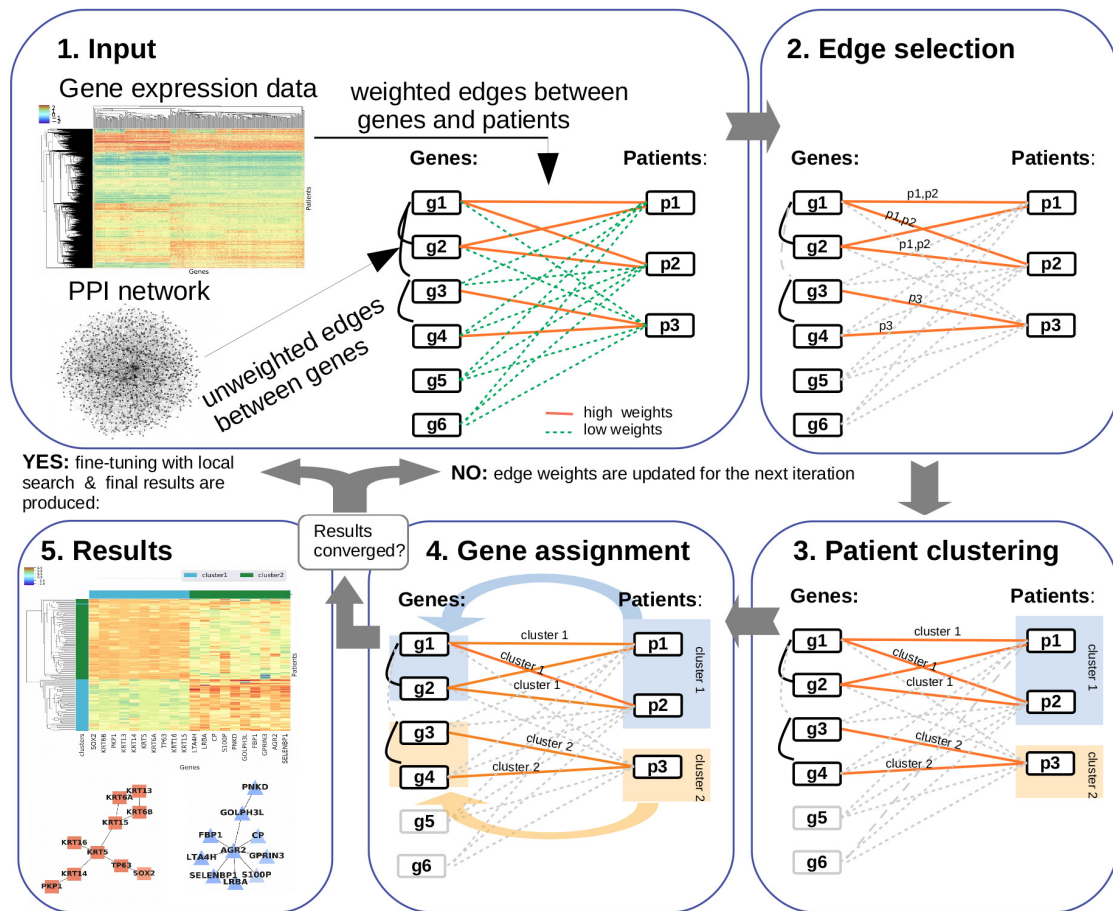


Figure 3.1.: Algorithmic framework of BiCoN. Described in *Bioinformatics* as "(Step 1) Gene expression data X is converted to a bipartite graph B and PPI interactions (black) are added as additional edges between genes to form a joint graph J . (Step 2) ACO determines the most relevant features for each patient where edges are annotated with patient IDs. The selected genes are used for patient clustering in Step 3. Next, (Step 4) genes are reassigned from individual patients to their corresponding clusters. Multiple possible solutions are computed in parallel and then evaluated and reinforced. As a result (Step 5), patients are stratified-based only on subnetworks that can be interpreted as disease mechanisms"

The steps are executed until the full convergence, i.e., until no more improvements are possible. For the full adaptation of the ACO to BiCoN, please refer to the supplementary material of the corresponding publication [9].

Local Search

Local Search is a heuristic approach that uses an assumption that a solution to a computationally difficult problem can be found by application of small local changes to a given solution (which is usually random at the start) until no more local improvements are possible [223]. Local Search has been known to perform well with Ant Colony Optimization algorithms [224] as ACO allows to explore very large search spaces but might be oblivious to finding the best possible solution in the most promising region of a graph.

Once ACO derives promising subnetworks, Local Search explores whether addition, substitution, or deletion of individual nodes to the solution, can improve the objective function value. The search goes on until no more local improvements are possible.

3.2. Active module identification methods evaluation

The main goal of the AMI testsuite [**ami**] is to establish a testing strategy for AMI methods. The expected behavior for such a method is that biologically meaningful results can be achieved only with the real PPI network, while network perturbations will lead to a decline in results quality. To systematically evaluate the results of AMI methods, the AMI suite performs assessments of the derived gene sets for ten different methods. Namely, ClustEx2 [212], COSINE [225], DIAMOnD [213], DOMINO [214], GiGA [226], GXNA [227], KeyPathwayMiner [215, 228, 229], GrandForest [220], Hierarchical HotNet [216] and NetCore [230]. Several different conditions were analyzed: non-small cell lung cancer, amyotrophic lateral sclerosis, ulcerative colitis, Chron's disease and Huntington's disease. All datasets were acquired from Gene Expression Omnibus [231]. The testsuite is summarized in Figure 3.2.

3.2.1. PPI networks and randomizations

Five widely used PPI networks were used to evaluate method's performance: BioGRID [232], APID [233, 234], STRING [56] with high confidence interactions only (score ≥ 0.7), HPRD [235] and IID [57] with experimentally validated interactions only. We performed five randomizations of the original networks to evaluate the expected drop in the algorithm's performance. The randomizations were performed such that they gradually alter more qualities of the original PPI networks. Summary of the alternations is provided in Table 3.1, and detailed explanation is given in the original publication [10] and cited without alterations below:

REWired: degree preserving generator. Repeatedly swaps pairs of edges and non-edges to produce random networks whose degree sequences are identical to the degree sequences of the original PPI networks. Preserves the individual node degrees and hence the hub-genes.

EXPECTED_DEGREE: expected degree preserving generator. Creates networks with randomly sampled edges where the sampling probabilities are chosen such that the expected node degrees correspond to the node degrees in the original PPI networks. Preserves individual node degrees and hub-genes in expectation.

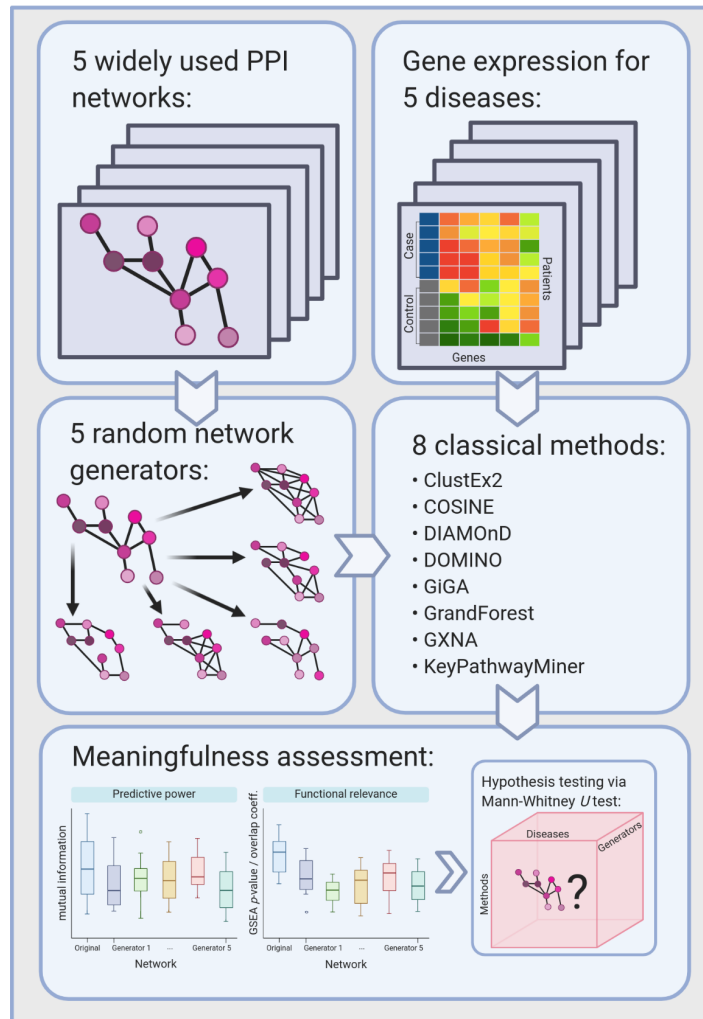


Figure 3.2.: AMI testing suite

SHUFFLED: topology preserving generator. Shuffles the gene IDs. Preserves the degree sequence and the topology but not the individual node degrees and the hub-genes.

SCALE_FREE: scale-free generator. Produces scale-free networks using the Barabási–Albert model. The parameters are chosen such that the numbers of nodes and edges in the random networks match the numbers of nodes and edges in the original PPI network. Preserves neither the topology nor the individual node degrees or the hub-genes, but produces networks that are structurally similar to the original PPI networks, since PPI networks are usually scale-free.

UNIFORM: uniform generator. Produces random graphs using the Erdős–Rényi model. The parameters are chosen such that the numbers of nodes and edges in the random networks matches the numbers of nodes and edges in the original PPI network. The produced networks are very different from the original PPI networks. In particular, their degrees are binomially distributed, whereas PPI networks tend to have power law degree distributions."

Table 3.1.: Summary of the performed randomizations

	Preserves the node degree exactly (preserves hubs)	Preserves the expected node degree (preserves hubs)	Preserves the network node-degree distribution exactly (does not preserve hubs)	Preserves the scale-free node-degree distribution (does not preserve hubs)
REWired	yes	yes	yes	yes
EXPECTED_DEGREE	no	yes	no	yes
SHUFFLED	no	no	yes	yes
SCALE-FREE	no	no	no	yes
UNIFORM	no	no	no	no

3.2.2. Gene sets evaluation

Evaluation of algorithms for biomarker discovery is challenging due to a lack of complete knowledge of disease-associated gene sets. Nevertheless, several approaches can be used as a proxy of relevance to a condition. These approaches can be separated into those based on similarity of the retrieved genes to a known disease-associated pathway and those that can accurately predict a patient's phenotype.

Association with disease-related genes

Disease-associated pathways can be retrieved through different databases such as KEGG [217, 218, 219], which collects pathway maps for various diseases, and DisGeNet [236], which collects information from GWAS catalogs, animal models, and the scientific literature. Gene set over-representation analysis (GSOA) allows to estimate a probability of observing by chance the overlap of a gene set in question with a known disease-associated gene set, given the size of both. GSOA is usually performed based on the hypergeometric test. The test estimates the probability of a certain number of "events" happening given how many events are observed in a population by chance. In the case of gene set overrepresentation analysis, the hit would represent an overlap between a gene set in question and a known disease-associated gene set. Thus, a probability of a gene set s to be associated with a disease-associated pathway d is equal to:

$$p_X(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad (3.1)$$

where N is a population size, i.e. the size of the background gene set from where the subset s was extracted,

K is the size of the disease-associated pathway d ,

n is the size of the gene set s ,

k is the size of the overlap of s and d , i.e. $|s \cap d|$.

For AMI test suite we performed enrichment of KEGG pathways with the retrieved gene modules. Additionally, we also computed overlap between the retrieved gene set and a

disease specific gene set from DisGeNet:

$$o = \frac{k}{\min(K, n)} \quad (3.2)$$

, using the same notation as in Equation 3.1.

Association with the phenotype

Evaluation of a gene set based on its ability to predict patients' phenotype allows estimating a predictive power of the biomarker set. Predictive gene sets are usually called "gene signatures" and might be used to predict disease subtype [237], survival [238] or a treatment outcome [239]. Various supervised machine learning models can be used to make the prediction based on the given gene set.

AMI test suite measures the strength of association using mean mutual information that quantifies how much information can be obtained about one variable (phenotype) by observing the other variables (gene expression). Mutual information is computed between expression of the obtained gene set s and a vector with patient phenotype information (case/control indication) y : $\sum_{g \in s} \frac{MI(x_g, y)}{n}$, where x_g in expression of a gene g .

4. Publications

4.1. Publication 1: BiCoN: network-constrained biclustering of patients and omics data

Citation

The article titled "BiCoN: network-constrained biclustering of patients and omics data" has been published online at OUP Bioinformatics on 26 December 2020 (in print on 15 August 2021).

Full citation:

"Olga Lazareva, Stefan Canzar, Kevin Yuan, Jan Baumbach, David B Blumenthal, Paolo Tieri, Tim Kacprowski, Markus List, BiCoN: network-constrained biclustering of patients and omics data, *Bioinformatics*, Volume 37, Issue 16, 15 August 2021, Pages 2398–2404, <https://doi.org/10.1093/bioinformatics/btaa1076>".

Summary

Unsupervised approaches for patients stratification into clinically relevant groups are necessary for data-driven patient stratification. Label-driven stratification (i.e., supervised learning) is easier to interpret as it uses the known patient groups, but it is oblivious to previously unknown group characteristics. Given that most diseases were defined before the availability of rich molecular data and therefore are based on symptoms, it is of paramount importance to contribute to novel, data-driven disease definitions.

We developed BiCoN to derive stable patient phenotypes from molecular data by aggregating patient-specific data (e.g., gene expression, methylation, copy-number variation) and prior knowledge in the form of Protein-Protein Interaction (PPI) network. BiCoN attempts to cluster patients and extract connected subnetworks from PPI network simultaneously. This allows BiCoN to infer molecular mechanisms that explain patient stratification directly.

BiCoN results were verified using Breast and Lung Cancer data and have shown robust reproduction of known molecular subtypes and potential novel subtypes that might have clinical relevance. The results were also exhaustively compared to known clustering and biclustering methods. Our benchmarking results demonstrated that BiCoN outperforms biclustering methods in a task of patient subtyping and outperforms clustering and biclustering methods in batch effect and noise robustness.

BiCoN achieves its performance by using Ant Colony Optimization to find nearly-optimal gene subnetwork that allows to cluster patients in the best possible way. In the final step,

we apply Local Search to fine-tune the solution and allow local changes to the subnetwork that might improve the objective function score. The two-step optimization function ensures result robustness and quality.

Availability

BiCoN is available and maintained as a Python package <https://pypi.org/project/bicon> and a web-interface <https://exbio.wzw.tum.de/bicon>. The data described in the manuscript is publicly available: Non-Small-Cell Lung cancer dataset is available through GEO: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219> and the breast cancer dataset (BRCA) is a part of The Cancer Genome Atlas <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>.

Contribution

I had the leading role in algorithmic framework design, wrote all the code for the algorithm and experiments, obtained and processed the data, and ran and adjusted competing methods for a fair comparison. I, together with Dr. M. List, co-supervised K. Yuan for the web interface development. I also wrote the first draft of the manuscript and generated all the figures. All project-related activities were supervised by Prof. T. Kacprowski, Prof. J. Baumbach, and Dr. M. List. Prof. D.B. Blumenthal provided his feedback on the algorithmic framework and the written manuscript. Prof. T. Kacprowski and Dr. M. List revised the manuscript. All authors provided their feedback on the final manuscript.

Rights and permissions

The original article is embedded with permission of Oxford Academic Press. All rights belong to Oxford Academic Press.

Additional supplementary material

Supplementary data are available at Bioinformatics online <https://doi.org/10.1093/bioinformatics/btaa1076>.

Systems Biology and Networks

BiCoN: Network-constrained biclustering of patients and omics data

Olga Lazareva^{1,*}, Stefan Canzar², Kevin Yuan¹, Jan Baumbach¹, David B. Blumenthal¹, Paolo Tieri^{3,4}, Tim Kacprowski^{1,5†}, Markus List^{1, †}

¹ Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich 80333, Germany

² Gene Center, Ludwig-Maximilians-University of Munich, Munich 81377, Germany

³ CNR National Research Council, IAC Institute for Applied Computing, Via dei Taurini 19, Rome, Italy

⁴ Data Science Program, La Sapienza University of Rome, Rome, Italy and

⁵ Division of Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Brunswick, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Unsupervised learning approaches are frequently employed to stratify patients into clinically relevant subgroups and to identify biomarkers such as disease-associated genes. However, clustering and biclustering techniques are oblivious to the functional relationship of genes and are thus not ideally suited to pinpoint molecular mechanisms along with patient subgroups.

Results: We developed the network-constrained biclustering approach BiCoN (Biclustering Constrained by Networks) which (i) restricts biclusters to functionally related genes connected in molecular interaction networks and (ii) maximizes the difference in gene expression between two subgroups of patients. This allows BiCoN to simultaneously pinpoint molecular mechanisms responsible for the patient grouping. Network-constrained clustering of genes makes BiCoN more robust to noise and batch effects than typical clustering and biclustering methods. BiCoN can faithfully reproduce known disease subtypes as well as novel, clinically relevant patient subgroups, as we could demonstrate using breast and lung cancer datasets. In summary, BiCoN is a novel systems medicine tool that combines several heuristic optimization strategies for robust disease mechanism extraction. BiCoN is well-documented and freely available as a python package or a web interface.

Availability and Implementation: PyPI package: <https://pypi.org/project/bicon>

Web interface: <https://exbio.wzw.tum.de/bicon>

Contact: olga.lazareva@tum.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomarkers are essential for stratifying patients for diagnosis, prognosis, or treatment selection. Currently, individual or composite molecular biomarkers based on, e.g., expression, methylation, mutation status, or

copy number variation are used. Biomarker discovery has greatly benefited from supervised methods that identify molecular features that have a strong association with disease-relevant variables such as drug response, relapse, survival time, or disease subtype. However, supervised methods are strongly biased by our current understanding of diseases, in particular by disease definitions that were established before rich molecular data became available. While classical unsupervised methods such as clustering have been successfully applied in the past, e.g., to reveal gene signatures

†Joint last authorship

predicting breast cancer subtypes (Parker *et al.*, 2009; Nielsen *et al.*, 2010), they group patients based on the entire molecular profile, and overlook meaningful differences limited to a subset of genes.

Biclustering aims to discover rows in a matrix which exhibit similar behaviour across a subset of columns and *vice versa* (Hartigan, 1972). It is suited for identifying disease-associated genes from gene expression data while stratifying patients at the same time (Prelic, 2006). As an NP-hard problem (Tanay *et al.*, 2002), biclustering is typically solved via heuristics. A gene expression matrix describes the expression of genes (rows) across samples (columns), which can reflect individual patients, time points or conditions. In patient stratification (i.e. splitting patients into clinically relevant subgroups), samples typically stem from different patients with a disease phenotype. In gene expression data, a bicluster defines a set of genes and a set of patients for which these genes are co-expressed (Cheng and Church, 2000). Gene co-expression does not imply a direct functional connection and, hence, genes identified by biclustering are often difficult to interpret. In contrast, molecular interaction networks such as protein-protein-interaction (PPI) networks capture direct and functional interactions.

Many diseases are caused by aberrations in molecular pathways or modules of functionally related genes (Berg *et al.*, 2002). This suggests to focus on gene modules for delivering more interpretable and robust mechanistic explanations of disease phenotypes. Network enrichment methods leverage prior information of molecular interactions for identifying gene modules as subnetworks (Batra *et al.*, 2017). Gene modules are robust features for classification and disease subtyping (Alcaraz *et al.*, 2017). Few methods exist that can utilize molecular interaction networks along with gene expression for patient stratification. Two integer linear programming methods were suggested (Yu *et al.*, 2017, Liu *et al.*, 2014) both of which rely on the GeneRank (Morrison *et al.*, 2005) algorithm to incorporate network information. GeneRank depends on a parameter θ describing the influence of the network whose choice is not straight-forward and was shown to have a notable impact on the results (Yu *et al.*, 2017). While these methods propagate the gene expression signal among the connected genes in a network, they generally do not produce connected subnetworks. Thus, they are not suited for discovering disease modules with mechanistic interpretation. To overcome this issue, we present BiCoN, a tool that accepts gene expression data as input and stratifies patients into two subgroups while identifying, for each group, a subnetwork of genes that can be interpreted as a shared molecular mechanism. In contrast to the classical definition of biclustering, BiCoN extracts a fixed number of non-overlapping biclusters, which are connected in a molecular interaction network. BiCoN delivers meaningful results on real-world datasets on par with other state-of-the-art methods. We have validated our results on breast cancer (TCGA Pan-Cancer) and non-small cell lung carcinoma (NSCLC) datasets (Rousseaux *et al.*, 2013) and found that BiCoN is robust to batch effects and delivers biologically interpretable mechanistic insights into disease subtypes.

2 BiCoN Approach

2.1 Problem statement

BiCoN aims at stratifying patients into two subgroups while extracting two sets of genes which are connected in a molecular interaction network and show opposite behaviour (i.e. similar to conventional differential expression analysis). The resulting subnetwork can thus be interpreted as a biological function jointly carried out by these genes which is active in one patient group and inactive in the other one. This assumption is reflected in our objective function and formally described below.

Consider a matrix of expression values $X^{n \times m}$ with n genes and m patients as well as $G = (V, E)$, a molecular interaction network of gene

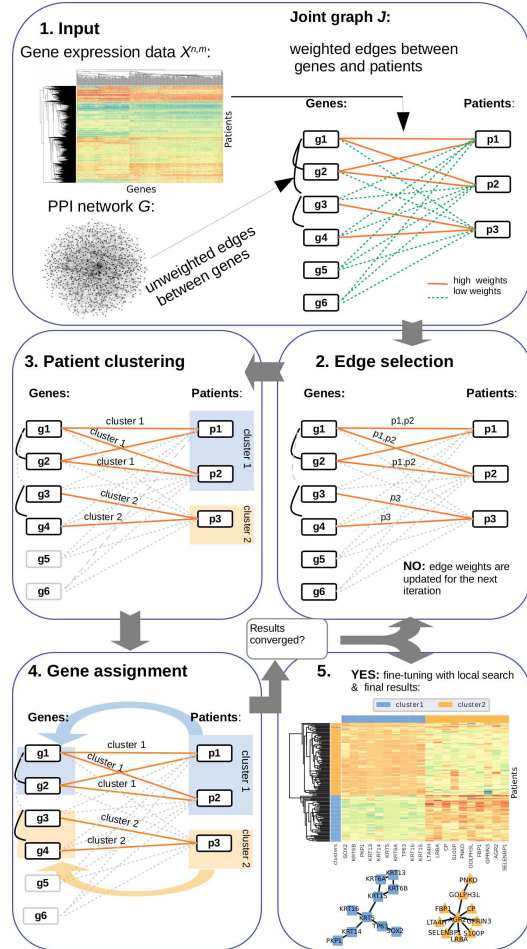


Fig. 1: The algorithmic framework of BiCoN. (1) Gene expression data X is converted to a bipartite graph B and PPI interactions (black) are added as additional edges between genes to form a joint graph J . (2) ACO determines the most relevant features for each patient where edges are annotated with patient ids. The selected genes are used for patient clustering in step (3). Next, (4) genes are reassigned from individual patients to their corresponding clusters. Multiple possible solutions are computed in parallel and then evaluated and reinforced. As a result (5), patients are stratified based only on subnetworks that can be interpreted as disease mechanisms.

set V and protein-protein or gene-gene interactions E . We further consider P as the set of m patients (samples) and construct a complete bipartite graph $B = ((V, P), E_w)$ with genes V and patients P as node types connected by weighted edges E_w . Edge weights reflect the expression strength for a given patient from expression values $X^{n \times m}$. We construct a joint graph $J = ((V, P), (E, E_w))$ by mapping G onto B via the shared genes in V . Our goal is to partition P into clusters P_1, P_2 , and to find 2 connected subnetworks $G_1(V_1, E_{,1}), G_2(V_2, E_2)$ each of minimal size L_{min} and of maximal size L_{max} . Size constraints can be adapted by users

to the expected size of the molecular pathways, i.e. small subnetworks will represent more specific and large subnetworks more general molecular functions or biological processes. Thus, we aim to derive patient groups (clusters P_1 and P_2) which are characterised by maximally differential expression in the extracted subnetworks:

$$f(X, V, P, c) = \sum_{(i,j) \in (1,2), (2,1)} w_i (\bar{X}[V_i, P_i] - \bar{X}[V_i, P_j]) \quad (1)$$

Where $\bar{X}[V_i, P_j]$ is the average expression of genes of module i for patients in cluster j , w_i is a weight for $G_i(V_i, E_i)$ which penalizes too small or too large, disconnected solutions:

$$w_i = \begin{cases} \frac{|LCC_{G_i(V_i, E_i)}|}{L_{min}} & \text{if } |LCC_{G_i(V_i, E_i)}| \leq L_{min} \\ \frac{L_{max}}{|LCC_{G_i(V_i, E_i)}|} & \text{if } |LCC_{G_i(V_i, E_i)}| \geq L_{max} \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Where $|LCC_{G_i(V_i, E_i)}|$ is the size of the largest connected component (LCC) in a subnetwork $G_i(V_i, E_i)$. Thus, w_i is always equal to 1 if the size of LCC corresponds to the user defined L_{min} and L_{max} and $w_i < 1$ means that the obtained solution does not fit into the desired range. Smaller w_i means larger deviation from a user’s preferences. The motivation for implementing a fuzzy threshold is that the user-selected L_{min} and L_{max} parameters may not always lead to a viable solution, i.e. if $w_i < 1$ BiCoN could not find any differentially expressed subnetworks of the selected size in the given data.

To obtain more than two clusters, BiCoN can in principle be applied recursively to further split clusters as also shown in the application in Section 4.2.

2.2 BiCoN algorithm

BiCoN is a heuristic algorithm that finds differentially expressed subnetworks that can mechanistically explain patient stratification. This combinatorial problem can be addressed by various metaheuristic frameworks such as e.g. Genetic Algorithm (Banzhaf *et al.*, 1998) or Swarm Intelligence (Eberhart and Kennedy, 1995). We have chosen Ant Colony Optimization (ACO) (Stützle, 2009) as the main framework that performs exploration of the search space and Local Search (Aarts *et al.*, 2003) to ensure local optimality of the final solution. The combination of ACO and Local Search was shown to be very efficient in finding near-optimal solutions to hard combinatorial optimization problems (Stützle and Hoos, 1999) and leads to significant improvements compared to ACO or local search alone (Stutzle and Hoos, 1997). As we already had good prior experiences with ACO on similar problems (Alcaraz *et al.*, 2012) we expected that combination with local search will lead to high quality results.

ACO is a nature-inspired probabilistic technique for solving computational problems which can be reduced to finding optimal paths through graphs. We use ACO to identify a set of relevant genes for each patient which we subsequently aggregate into a global solution. A full description of the algorithm and the pseudo-code can be found in the Supplementary Material, section "Algorithm description". We also describe the full workflow on Figure 1. Briefly, ants travel the joint graph J in three phases which are repeated until convergence:

1. An ant performs a random walk within nodes that are highly connected to a patient-node and makes greedy choices according to the objective function (Equation 1) by choosing genes which are most relevant to a

patient (orange edges on Figure 1 step 2). The probability of selecting a gene for a certain patient depends on the combined information from gene expression values (which are encoded in the heuristic information matrix) and the ant’s “memories” on whether the choice of this gene has led to a quality solution in the previous rounds (pheromone matrix). More details about the implementation can be found in Supplementary Material, section “Algorithm description”.

2. The selected genes are then used for clustering patients with the k-means algorithm where $k = c = 2$ (step 3, Figure 1). Relevant genes for each patient cluster are extracted at step 4 (Figure 1). A candidate solution is evaluated by the objective function score.
3. The best solution is used for updating the pheromone and probability matrices for the next iteration.

When the best solution is obtained we perform local search for possible local improvements, i.e. iteratively apply changes to subnetworks (such as node insertion, deletion or substitution) and keep changes that lead to objective function maximization. This allows us to retrieve robust and stable solutions as well as to ensure local optimality.

Even though BiCoN uses several hyperparameters, our experiments have shown that those do not have a large impact on the results and the optimal combination is determined automatically based on the dimension and distribution of the expression matrix. Therefore, the user only has to specify the desired size of the solution subnetworks (L_{min} and L_{max}).

3 Methods

3.1 Data collection and processing

3.1.1 Gene expression data

TCGA breast cancer data was obtained through the UCSC Xena browser (<https://xenabrowser.net/>). The NSCLC dataset (accession number GSE30219, (Rousseaux *et al.*, 2013) was obtained using GEO2R (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>). Both datasets were retrieved together with the corresponding metadata which contained annotated cancer subtypes.

For the NSCLC dataset, gene probes were mapped to Entrez gene IDs. If multiple probes corresponded to a single gene, the median value was used. We applied a \log_2 transformation to account for skewness of the data. Data was z-score transformed to indicate the magnitude of changes in gene expression in individual samples and conditions compared to the background. In most gene expression datasets, a majority of genes is lowly expressed and does not vary to a larger extent. To account for this and to improve run-time, BiCoN filters out genes with a small variance preserving only the n most variant genes (here $n = 3000$).

3.1.2 Molecular interaction network

We used physical and genetic protein-protein interactions (PPI) in *H. Sapiens* from BioGRID (version 3.5.176). The network consisted of 343,563 unique interactions between 16,830 genes.

3.2 Simulation of Batch Effects

Batch effects are technical variations that have been introduced by external factors during handling of the samples (e.g. personnel effects, environmental conditions, different experiment times) (Luo *et al.*, 2010; Goh *et al.*, 2017). While some of those effects can be minimised, batch effects are still almost inevitable in practice (Chen *et al.*, 2011). Many methods have been proposed for removing batch effects from data (Lazar *et al.*, 2012). However, removing batch effects may also remove biologically relevant group differences from the data. Batch effect

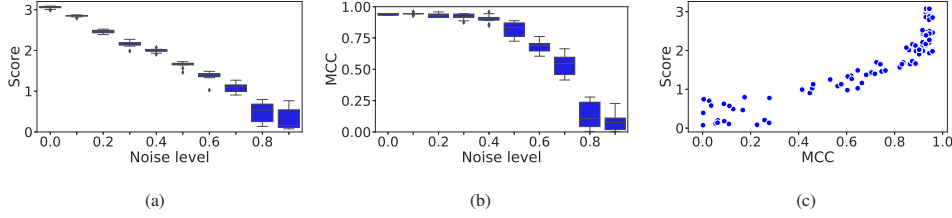


Fig. 2: Robustness analysis. **a)** Objective function score versus the percentage of noisy data. **b)** Matthews Correlation Coefficient (MCC) with respect to the known classes versus the percentage of noisy data. **c)** Correlation of objective function scores and MCC.

correction methods that are designed to retain group differences can lead to exaggerated confidence in downstream analyses (Nygaard et al., 2016). In unsupervised analysis, this issue is critical, since we, by definition, do not know the relevant sample or patient groups *a priori*.

To demonstrate that BiCoN is robust to batch effects, we simulate data using a linear mixed effect model. We consider two variables: *cluster* and *batch*. The variable *cluster* indicates whether a gene is part of the foreground ($cluster = 1$ or $cluster = 2$) or the background ($cluster = 0$), i.e. it is not differentially expressed. The variable *batch* indicates the study or batch of expression values ($batch = 1$ or $batch = 2$). The expression values are simulated as follows:

$$g_i = 1 + 2 \times batch + 2 \times cluster + \gamma_1 \times cluster + \gamma_2 \times batch + \varepsilon_i \quad (3)$$

where the first part of the equation ($1 + 2 \times batch + 2 \times cluster$) is fixed and shared by all genes. Errors ε_i are independent and identically distributed (with zero mean). The random effects parameters γ_1 and γ_2 follow a bivariate normal distribution with zero mean, and variance 1 and 2 respectively, i.e. the technical variance is twice the biological variance.

The network was simulated as three disjoint Barabasi-Albert graphs (one for each of genes biclusters and one for background genes) (Barabási and Albert, 1999) which were connected by random edges until they have reached the same density as the BioGRID network (0.0013). Barabasi-Albert graphs have similar node degree distribution as the BioGRID network and were thus considered suitable for the simulation study.

3.3 Benchmarking

To show how BiCoN results compare to commonly used clustering and biclustering algorithms, we selected several popular biclustering and clustering methods (listed in Supplementary Table S3) and performed multiple assessments:

- To show how BiCoN can recover PAM50 annotated breast cancer subtypes (using TCGA data as a source), we computed Jaccard index (an intersection of two sets over the union) between the known subtypes and the resulting patients clusters/biclusters.
- To show how BiCoN can handle batch effect in comparison to other methods, we simulated data as described in section 3.2 and computed the overlap between known classes of patients and the resulting clusters/biclusters. To avoid favouring the assumption of genes connectivity used by BiCoN, we also repeated the simulation such that the signal-carrying foreground genes are randomly distributed over the network.

As a metric for comparison we used Jaccard index rather than MCC as it allows to measure relationship between resulting biclusters and the

actual classes even when the patients biclusters overlap and do not include all patients. All data was normalized and processed as described in Section 3.1 for all methods (including BiCoN).

Even though we use classical clustering methods for benchmarking, we emphasise key differences between the suggested approach and classical clustering. BiCoN extracts biological mechanisms that explain patient stratification. Even though subnetworks extraction after clustering of patients is feasible, to our knowledge there is no gold standard for this procedure. While it is possible to extract subnetworks and disease mechanisms subsequent to clustering or by relying on known disease subtypes (Alcaraz et al., 2017), we argue that such clusters are driven by global differences and not by the activity of a single disease mechanism. Hence, extracting disease mechanisms along with patient stratification is better suited to identify patient subgroups affected by key disease mechanisms. Moreover, clustering performed on the whole genome is also not advisable as the use of multidimensional data can lead to multiple negative effects, which are often referred to as “curse of dimensionality” (Thangavelu et al., 2019).

For all selected algorithms, we chose parameters that maximize performance for each of the methods.

4 Results and Discussion

We evaluated BiCoN on simulated and real data with respect to the robustness of patient clustering and gene selection as well as robustness to batch effects. Furthermore, two application cases illustrate the practical use of BiCoN.

4.1 Noise robustness

To introduce varying levels of noise to a data set, we randomly select between 0 and 90% of the genes and randomly permute their expression values. A noise level of 0.1 means that the expression vectors of 10% of genes were permuted. For each noise level, we average results over 10 independent runs.

We use the NSCLC data set with two annotated subtypes as gold standard: adenocarcinoma and squamous cell carcinoma. As evaluation metrics, we consider the value of BiCoN objective function as well as Matthews Correlation Coefficient (MCC) (Matthews, 1975) between the proposed clusters and cancer subtype labels. The latter is meant to demonstrate that BiCoN is able to recover cancer subtypes while inferring a mechanistic explanation for the subtype differences. For all described results, we retain the 3000 most variant genes and set parameters $L_{min} = 10$ and $L_{max} = 25$ to control the size of the solution.

Figure 2(a) shows a consistent decline in the objective function with increasing noise, indicating that the algorithm is reacting reasonably to

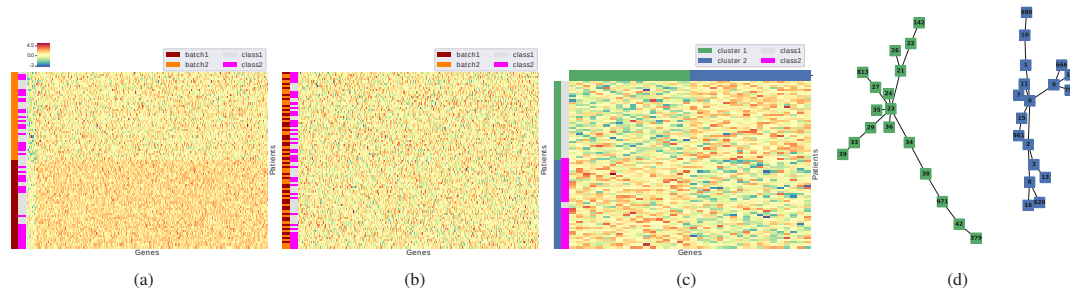


Fig. 3: **a)** Hierarchical clustering of two data sets with different distributions due to batch effects. **b)** The merged data sets after z-scores normalization. The batch effect vanishes, but the disease phenotype is still not distinguishable. **c)** BiCoN is able to recover the initial disease phenotypes with Jaccard index of 0.92 (in average after 10 runs) while extracting the 40 foreground genes out of 1000 background genes. **d)** The resulting subnetworks for two corresponding patient clusters.

the decline in data quality. Figure 2(b) shows that the algorithm is able to recapture the cancer subtypes almost perfectly (average MCC higher than 0.9) up to a noise level of 0.5 where 50% of the data have been permuted. Figure 2(c) shows a strong positive correlation between the objective function value and MCC, which confirms that the objective function is high when cancer subtypes are well separated.

4.2 Batch effect robustness

BiCoN is a graph-based method and, hence, it is not as strongly affected by the global distribution of expression values as classical clustering methods. Pre-processing methods that scale data to a certain range enforce it to have certain mean and variance (e.g. z-scores) or make the distribution more symmetrical (e.g. \log_2 transformation) are not ideal for batch effect correction as they do not differentiate between signal and noise. In this scenario, a graph-based method benefits from the assumption that the joint signal of the genes in a subnetwork is stronger than that of individual genes.

To study if BiCoN can indeed tolerate batch effects, we simulate gene expression data (see Methods for details) where we introduce a batch effect with a larger variance than for the group difference. Our aim is to show that BiCoN can leverage the network to recover the signal even if it is overshadowed by batch effects.

We have simulated expression data for 2×20 foreground genes (two biclusters) and for 1000 background genes. We also tested the performance with 2×30 , 2×40 and 2×60 foreground genes.

Figure 3(a) shows that the batches differ in their distribution, causing hierarchical clustering to group samples by batch rather than by disease phenotype. Figure 3(b) shows that differences due to batch effects are eliminated after z-score normalization. We can also see that the difference between the sample groups is now lost and can not be recovered by hierarchical clustering. Figure 3(c) shows that in spite of this noise, BiCoN can recover the disease phenotype together with the foreground genes. Thus, when two datasets can be normalized separately (e.g. z-scores are applied to each dataset), BiCoN is uniquely suited to cluster patients where individual gene modules are disturbed. Even when the signal is obscured by batch effects, the functional connection of solution genes in the network (Figure 3 (d)) helps to robustly recover the signal.

To show how BiCoN results align with other clustering and biclustering methods, we have simulated 10 datasets with batch effect and evaluated the performance. To make sure that we do not put BiCoN in favour by enforcing connectivity of genes, we also performed simulations with a

single Barabasi-Albert graph, where foreground genes were randomly distributed (Figure 4).

Among the considered biclustering algorithms (Supplementary Table S3), only Bimax was capable of finding any clusters, while Plaid and QUBIC could not find any structure in the given data regardless of chosen parameters and therefore was excluded from further assessment. The experiments showed that even though the quality of the results drops when the foreground genes are not directly connected, BiCoN still performs significantly better than other methods. The simulated network had power-law node degree distribution which means that the network diameter is rather small and therefore many foreground genes are still reachable through hub-nodes even when they are not directly connected. Thus, BiCoN performance dropped when using random networks (due to the noise of the hub nodes) but still outperformed other methods that are not network-restricted.

4.3 Application to TCGA breast cancer data

We applied BiCoN to the TCGA breast cancer dataset. We expected BiCoN to be able to recover known subtypes assigned via the PAM50 gene panel (Parker *et al.*, 2009; Nielsen *et al.*, 2010) which is a gold standard in breast cancer subtype prediction. For the analysis, we focused on patients with the most common molecular subtypes, luminal (estrogen-receptor and/or progesterone-receptor positive) and basal (hormone-receptor-negative and HER2 negative).

As a proof of concept, we first showed that BiCoN can separate patients into the two clinically well distinguishable subtypes luminal and basal. Next, we applied BiCoN separately for patients with luminal and basal subtype to investigate how patients are stratified in a more challenging scenario. For each subgroup, we ran the algorithm 10 times and selected a solution with the highest score based on the previous observation that the highest objective function score corresponds to the highest correlation between the resulting biclusters and the expected patient groups. We conducted gene set enrichment using genes from both subnetworks together using the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) as a background. We used the same hyperparameters as for our previous analysis: 3000 the most variant genes and $L_{min} = 10$ and $L_{max} = 25$ to control the size of the solution.

4.3.1 Luminal versus basal separation

As expected, the separation between patients with luminal and basal breast cancer subtypes is straightforward. The clusters correspond to the subtype

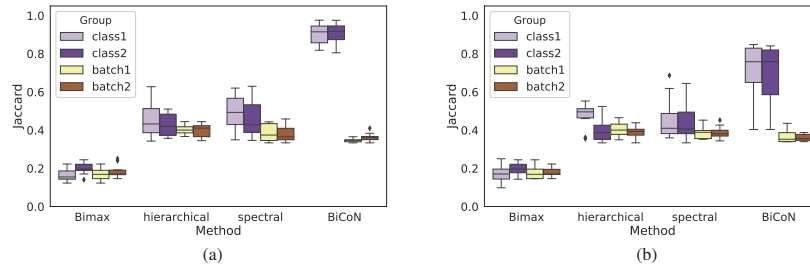


Fig. 4: Jaccard indices between the patients clusters and actual subgroups (class 1 or class 2) as well as with batches of patients (batch 1 and batch 2) for 10 simulated datasets with a strong batch effect. **a)** When foreground genes are connected in a network, BiCoN clusters patients almost perfectly based on the actual signal. **b)** When the foreground genes are randomly distributed in the network, BiCoN still achieves higher performance than other methods that were capable to find any clusters. Plaid and QUBIC were not able to find any clusters and were excluded from further assessment.

labels and the separation between patients groups matches the PAM50 classification (average Jaccard index is equal to 0.99, Supplementary Figure S1 (a)). BiCoN not only performs as well as methods like hierarchical clustering (Figure 5, where the Jaccard index is 0.96 for the luminal and basal subtypes) but also yields two differentially expressed subnetworks (Supplementary Figure S1 (b)). The extracted subnetworks explain subtype differences with a vastly lower number of genes than a classical clustering method while offering a mechanistic explanation of subtype differences. Note that while BiCoN restricts genes inside a bicluster to be connected, it does not impose any relationships between two biclusters. As a consequence, it is possible that the resulting subnetworks overlap.

In contrast to methods yielding gene signatures such as PAM50, BiCoN focuses on revealing specific pathways. Enrichment analysis of cancer-related pathways (Supplementary Material Figure S6) confirms strong association of the resulting genes with breast cancer subtype-specific signalling, in particular estrogen signaling pathway (adjusted p-value = 0.018) and ErbB signaling pathway (adjusted p-value = 0.025).

Random-walks on scale-free networks are biased towards hub nodes since these have a high degree (Gillis et al., 2014). BiCoN avoids this hub bias as it performs random walks on the joint graph of a PPI and expression data which is not scale-free. Consequently, the selected nodes have approximately the same degree distribution as the input network (Supplementary Materials Figure S4).

4.3.2 Luminal patient stratification

Next, we consider only patients that were originally classified as luminal subtype to see if we can further stratify them into subtypes luminal A and luminal B which are known to be difficult to separate on the level of gene expression. Here, our solution does not agree with the PAM50 classes (Figure 5, mean Jaccard index 0.49 for the luminal A (lumA) and luminal B (lumB) subtypes), although we observe two clearly separable groups and that most of the luminal B patients were placed in cluster 1 (Supplementary Figure S2). We hypothesize that contributions of the tumor-microenvironment may explain the observed clusters. To test this hypothesis, we used the signature-based deconvolution method xCell (Aran et al., 2017) to estimate contributions of 64 immune and stromal cell types in the two clusters. xCell summarizes the contribution of tumor-infiltrating leukocytes to the microenvironment via aggregated scores such as an immune score, a stromal score and a microenvironment score. Clusters reported by BiCoN show significant differences between cell types scores. The strongest difference between patients is found in

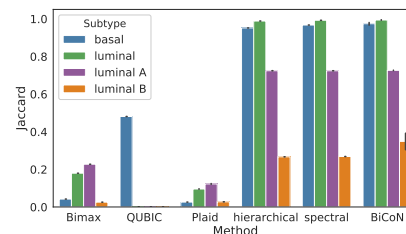


Fig. 5: TCGA breast cancer subtypes identification by various algorithms (for 10 runs). Jaccard index was computed as a best match between produced patients clusters and the known breast cancer subtypes for BiCoN and other well-known clustering and biclustering algorithms. BiCoN shows performance which is comparable with other clustering algorithms while also revealing functionally connected subnetworks which explain the phenotype.

the stromal score ($-\log_{10}$ p-value is over > 55), hematopoietic stem cells ($-\log_{10}$ p-value > 50) and CLP cells ($-\log_{10}$ p-value > 50). See Supplementary Figures S7(a), S8(a) for details. These results indicate that some of the luminal A and luminal B patients share similar tumor microenvironments and, consequently, the further stratification of luminal subtypes is not straightforward. These results are corroborated by other studies which investigate immune-related subtypes of luminal breast cancer (Zhu et al., 2019; Jiang et al., 2020).

4.3.3 Basal patients stratification

Bertucci et al., 2012 characterised basal, also known as triple negative, breast cancer as the most challenging breast cancer subtype with poor prognosis despite relatively high chemosensitivity. Currently, there is no targeted therapy and no routine diagnostic procedure specifically for this subtype. Although no clinically relevant subgroups of the basal subtype are known, BiCoN achieved a clear separation into two subgroups (Supplementary Figure S3).

Derived subnetworks show robust correlation with immune system response functions which is reasonable given that tumour samples are infiltrated with leukocytes. All 3 enriched pathways (Primary immunodeficiency, Hematopoietic cell lineage, B cell receptor signaling

pathway) have a direct connection to the immune response (Figure S5(a) Supplementary Material). Molecular function enrichment also confirms the relation between the selected genes and immune response (Figure S5(b) Supplementary Material). Cell type deconvolution analysis with xCell shows a high correlation of the clusters with aDC, CD4+ memory T-cells, B-cells, CD8+ T-cells and other immune response related cells (Supplementary material Figures S7(b) and S8(b)). Similar to the results in luminal patients, our results indicate that basal breast cancer patients can be clustered by the contribution of tumor-infiltrating leukocytes, which is a clinical key factor for prognosis and treatment via immunotherapy.

5 Conclusion and Outlook

Classical biclustering methods were shown to perform sub-optimally when non-intersecting, large patient subgroups are of interest as is often the case in patient stratification. Clustering methods, on the other hand, are more suited for this task, but they use the whole gene set and do not provide a mechanistic explanation of patient stratification (Figure 5). Therefore BiCoN is uniquely suited to cluster patients along with extracting fixed-size subnetworks capable of mechanistically explaining the patient stratification. Moreover, simultaneous clustering of gene expression and networks makes BiCoN robust to noise and more robust to batch effect than typical clustering and biclustering methods.

BiCoN leverages molecular interaction networks in the analysis of gene expression data to faithfully produce known subtypes as well as novel, clinically relevant patient subgroups, as we could demonstrate using data from TCGA. We stress that BiCoN and the concept of network-constrained biclustering are not limited to gene expression data or protein-protein interaction networks. We plan to apply BiCoN to other types of omics data such as DNA methylation, copy number variation or single nucleotide polymorphisms. We envision BiCoN to be useful for single-cell RNA-seq data for uncovering differences in signalling between clusters of cells and for the discovery of novel cell types. BiCoN, which is available as a web-interface and a PyPI package, has great potential to enhance our understanding of diseases, cellular heterogeneity and putative drug targets.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

We thank Quirin Heiß for his contributions to the source code of the web-interface and Hoan Van Do for a fruitful discussion about the algorithm

Funding

This work was supported by the Bavarian State Ministry of Science and the Arts as part of the Bavarian Research Institute for Digital Transformation (bidit) [to OL]; H2020 project RepoTrial [777111 to JB and TK]; VILLUM Young Investigator grant [to JB]; COST CA15120 OpenMultiMed [to OL].

References

Aarts, E. *et al.* (2003). *Local search in combinatorial optimization*. Princeton University Press.
 Alcaraz, N. *et al.* (2012). Efficient key pathway mining: combining networks and omics data. *Integrative Biology*, **4**(7), 756–764.
 Alcaraz, N. *et al.* (2017). De novo pathway-based biomarker identification. *Nucleic Acids Res.*, **45**(16), e151.

Aran, D. *et al.* (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**(1), 220.
 Banzhaf, W. *et al.* (1998). *Genetic programming: an introduction*, volume 1. Morgan Kaufmann San Francisco.
 Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, **286**(5439), 509–512.
 Batra, R. *et al.* (2017). On the performance of de novo pathway enrichment. *NPJ systems biology and applications*, **3**(1), 6.
 Berg, J. *et al.* (2002). Defects in signaling pathways can lead to cancer and other diseases. *Biochemistry. 5th Edition*. New York: WH Freeman, Section, **15**.
 Bertucci, F. *et al.* (2012). Basal breast cancer: a complex and deadly molecular subtype. *Curr. Mol. Med.*, **12**(1), 96–110.
 Chen, C. *et al.* (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS one*, **6**(2).
 Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, **8**, 93–103.
 Eberhart, R. and Kennedy, J. (1995). Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks*, volume 4, pages 1942–1948. Citeseer.
 Gillis, J. *et al.* (2014). Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of proteomics*, **100**, 44–54.
 Goh, W. W. B. *et al.* (2017). Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.*, **35**(6), 498–507.
 Hartigan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, **67**(337), 123–129.
 Jiang, J. *et al.* (2020). Tumour-infiltrating immune cell-based subtyping and signature gene analysis in breast cancer based on gene expression profiles. *Journal of Cancer*, **11**(6), 1568.
 Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.
 Lazar, C. *et al.* (2012). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, **14**(4), 469–490.
 Liu, Y. *et al.* (2014). A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC bioinformatics*, **15**(1), 37.
 Luo, J. *et al.* (2010). A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *The pharmacogenomics journal*, **10**(4), 278–291.
 Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**(2), 442–451.
 Morrison, J. L. *et al.* (2005). GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.
 Nielsen, T. O. *et al.* (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.*, **16**(21), 5222–5232.
 Nygaard, V. *et al.* (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, **17**(1), 29–39.
 Parker, J. S. *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**(8), 1160–1167.
 Prelic, B. S. & Zimmermann, P. (May 2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.
 Rousseaux, S. *et al.* (2013). Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Science translational medicine*, **5**(186), 186ra66–186ra66.

- Stützle, T. (2009). Ant colony optimization. In M. Ehrgott, C. M. Fonseca, X. Gandibleux, J.-K. Hao, and M. Sevaux, editors, *Evolutionary Multi-Criterion Optimization*. Springer Berlin Heidelberg.
- Stutzle, T. and Hoos, H. (1997). Max-min ant system and local search for the traveling salesman problem. In *Proceedings of 1997 IEEE international conference on evolutionary computation (ICEC'97)*, pages 309–314. IEEE.
- Stützle, T. and Hoos, H. (1999). The max-min ant system and local search for combinatorial optimization problems. In *Meta-heuristics*, pages 313–329. Springer.
- Tanay, A. et al. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18 Suppl 1**, S136–144.
- Thangavelu, S. et al. (2019). Feature selection in cancer genetics using hybrid soft computing. pages 734–739.
- Yu, G. et al. (2017). Network-aided Bi-Clustering for discovering cancer subtypes. *Sci Rep*, **7**(1), 1046.
- Zhu, B. et al. (2019). Immune gene expression profiling reveals heterogeneity in luminal breast tumors. *Breast Cancer Research*, **21**(1), 147.

4.2. Publication 2: On the limits of active module identification

Citation

The article titled "On the limits of active module identification" has been published online at Briefings in Bioinformatics on 29 March 2021 (in print on September 2021).

Full citation:

"Olga Lazareva, Jan Baumbach, Markus List, David B Blumenthal, On the limits of active module identification, Briefings in Bioinformatics, Volume 22, Issue 5, September 2021, bbab066, <https://doi.org/10.1093/bib/bbab066>."

Summary

Identification of active modules in PPI networks based on gene expression data is often used to find mechanisms potentially responsible for a given phenotype. The connection of transcriptome with proteome appears intuitively correct and, as shown in the first publication (BiCoN), increases the computational robustness of results. An important question to those results is whether they were achieved due to the biological value of PPI networks or due to possible technical or selection bias. As PPI networks are widely used in many developed Systems Medicine approaches, it is of paramount importance to investigate possible biases and determine informational gain provided by PPI networks.

We developed a testing framework that allowed us to disentangle various properties of PPI networks (such as diameter, node degree distribution, hub nodes distribution, and others) and evaluate their influence on the results of active module identification methods (AMIMs). Several *meaningfulness scores* have been developed to systematically evaluate results in terms of functional gene relation to a phenotype in question, ability to predict patient phenotype, and survival prognosis.

We have established that most AMIMs do not produce more meaningful results on actual PPI compared to PPIs with random edges and preserved node degrees. This conclusion implies that AMIMs are mostly oblivious to biological knowledge coming from PPIs and prioritize proteins based on their degrees. The consequence of this conclusion is a necessity to reconsider the field of active module identification and employ the developed testing procedure to eliminate systematic biases when using biological networks.

Availability

As described in the publication: "The KEGG pathways were obtained from KEGG: <https://www.genome.jp/kegg/disease/>. BioGRID (v3.2.149), APID (v1.0), STRING (version 11.0) and HPRD (release 9) as well as DisGeNET (v7.0) were obtained using nDEx [240, 241, 242]. The IID network (v2018-11) was downloaded from <http://iid.ophid.utoronto.ca/>. All gene expression datasets and corresponding metadata were retrieved from Gene Expression Omnibus [231], using the GEO2R R interface (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>). The entire

test-suite (Python environment, tool executables, PPI networks, expression datasets) is available at <https://github.com/dbblumenthal/amim-test-suite/>."

Contributions

I developed the hypothesis and a general testing strategy, processed the data and collected, aggregated, and analyzed methods results. Together with Prof. D.B. Blumenthal, I wrote the code for the testing framework and implemented the Active Module Identification Methods. Prof. D.B. Blumenthal, Dr. M. List and Prof. J. Baumbach supervised the project. All authors contributed to writing of the manuscript and provided their feedback. I finalized the manuscript and produced the figures.




Rights and permissions

The original article is embedded with permission of Oxford Academic Press. All rights belong to Oxford Academic Press.

Additional supplementary material

Supplementary data are available online at Briefings in Bioinformatics <https://doi.org/10.1093/bib/bbab066>.

On the limits of active module identification

Olga Lazareva , Jan Baumbach, Markus List [†] and David B. Blumenthal [†]

Corresponding author: David B. Blumenthal, Chair of Experimental Bioinformatics, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany, phone: +49 8161 71 2712; E-mail: david.blumenthal@wzw.tum.de

[†]Joint senior authors.

Abstract

In network and systems medicine, active module identification methods (AMIMs) are widely used for discovering candidate molecular disease mechanisms. To this end, AMIMs combine network analysis algorithms with molecular profiling data, most commonly, by projecting gene expression data onto generic protein–protein interaction (PPI) networks. Although active module identification has led to various novel insights into complex diseases, there is increasing awareness in the field that the combination of gene expression data and PPI network is problematic because up-to-date PPI networks have a very small diameter and are subject to both technical and literature bias. In this paper, we report the results of an extensive study where we analyzed for the first time whether widely used AMIMs really benefit from using PPI networks. Our results clearly show that, except for the recently proposed AMIM DOMINO, the tested AMIMs do not produce biologically more meaningful candidate disease modules on widely used PPI networks than on random networks with the same node degrees. AMIMs hence mainly learn from the node degrees and mostly fail to exploit the biological knowledge encoded in the edges of the PPI networks. This has far-reaching consequences for the field of active module identification. In particular, we suggest that novel algorithms are needed which overcome the degree bias of most existing AMIMs and/or work with customized, context-specific networks instead of generic PPI networks.

Key words: active module identification; de novo network enrichment; network and systems medicine; systems biology

Introduction

Because of massive advances in high-throughput technologies, large amounts of gene expression data have become available over the past decades. This has raised hopes to identify new molecular mechanisms that might provide valuable insights into cellular function and the pathobiology of diseases [1–3]. However, gene expression data tend to be overdetermined and noisy and, as a result, the discovery of disease genes via purely statistical means is often unstable, since the reported genes are often just surrogates of the actual disease genes and hence functionally not necessarily related to the disease of interest [4, 5].

To mitigate these problems, active module identification methods (AMIMs) leverage additional biological knowledge encoded in protein–protein interaction (PPI) networks [6–9]. These methods project gene expression data onto PPI networks and then use network algorithms to identify disease modules consisting of small subnetworks. This dramatically decreases the size of the search space and prioritizes disease modules consisting of functionally related genes, which, in turn, positively affects both stability and functional relevance of the discovered modules [10]. AMIMs have been successfully used for providing novel pathobiological insights into complex diseases such as pulmonary arterial hypertension [11], coronary heart

Olga Lazareva is doctoral fellow at the Bavarian Research Institute for Digital Transformation and a PhD candidate at the Technical University of Munich. **Jan Baumbach** is professor and chair of Computational Systems Biology at the University of Hamburg. He obtained his PhD in Computer Science from Bielefeld University.

Markus List obtained his PhD at the University of Southern Denmark and worked as a postdoctoral fellow at the Max Planck Institute for Informatics before starting his group Big Data in BioMedicine at the Technical University of Munich.

David B. Blumenthal is postdoctoral fellow at the Technical University of Munich. He obtained his PhD in Computer Science from the Free University of Bozen-Bolzano.

Submitted: 11 January 2021; Received (in revised form): 29 January 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

disease [12], diabetes mellitus [13], liver fibrosis [14], chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis [15], as well as asthma [16].

Despite these impressive results, there is increasing awareness in the field that the combination of gene expression data and PPI networks is subject to technical and literature bias. PPI networks suffer from technical bias [17] e. g. since the ‘bait’ proteins used for measuring new interactions often have significantly more interactions. Moreover, literature bias [18], where research focuses on proteins with already known characteristics (e. g. biological function), leads to a strong correlation between the number of studies conducted on a protein and the protein’s degree in the PPI network.

The node degree distribution of PPI networks typically follows a power law. As a consequence, perturbances of cellular programs (e. g. via mutations or other mechanisms) typically have a cascading effect when observed on the level of gene expression. As a result, differential gene expression analysis often reveals hundreds or thousands of genes to be disease-associated. By projecting these noisy gene expression data on PPI networks with a small diameter, disease-associated genes can easily be combined into subnetworks or disease modules, most of which may not contain a single disease-causative gene. Although such network modules may be well suited as robust biomarkers for a disease, they may be less suited to pinpoint a disease mechanism.

To account for network-related biases, some recently proposed methods such as Hierarchical HotNet [19] and NetCore [20] integrate data and network randomization steps into their workflows. These permutation-based methods extract subnetworks whose associations with the disease are significantly stronger in the original PPI networks than in the randomized counterparts. Levi et al. further reported that gene ontology enrichment of several state-of-the-art AMIMs on randomly permuted input data produced similar results, questioning the context-specificity of existing AMIMs. To address this issue, Levi et al. [21] propose a new method DOMINO.

Although the effect of random permutations of the input omics data was systematically tested by Levi et al. [21], the question if AMIMs also benefit from the biological knowledge captured in PPI networks remains unanswered (cf. Figure 1). In this study, we close this gap. For this, we developed a test suite for AMIMs, which studies the effect of different types of network randomization on the results. Our test suite, which is openly available at <https://github.com/dbblumenthal/amim-test-suite/>, expects a network and expression data (or input that can be derived from expression data) as input and produces a set of candidate disease modules as output. These modules are then evaluated using mutual information (MI) and gene set enrichment analysis (GSEA) with known disease signatures (see ‘Methods’ for details). Since further AMIMs can easily be integrated by implementing a well-defined interface, our test suite can be used not only to reproduce the results reported in this paper, but also to objectively test novel AMIMs with respect to their robustness against network randomization.

In a large-scale empirical evaluation on gene expression data for five different diseases, we ran eight classical and two permutation-based AMIMs on five different widely used PPI networks as well as on randomized counterparts generated by five different random network generators (more than 10 000 runs in total). The most striking result of our analysis is that all except one of the tested AMIMs did not yield significantly more meaningful subnetworks if run on the original PPI networks than if run on random networks with matching node degrees. Most

AMIMs hence pick up on the number of interactions a protein is involved in, but do not benefit from the biological knowledge captured in the PPIs themselves.

The remainder of this paper is organized as follows: In the ‘Results’ section, we briefly describe the protocol implemented by our test suite and present the results of our analyses. In the ‘Discussion’ section, we discuss the implications of our findings for the field of active module identification. In the ‘Methods’ section, we provide a more detailed description of our test protocol and also elaborate on how developers of new AMIMs can use our test suite to evaluate their methods.

Results

Test protocol

Figure 2 visualizes our protocols for method evaluation (cf. ‘Methods’ section for details). We selected eight classical AMIMs, also referred to as *de novo* network enrichment tools in the literature [6] (ClustEx2 [22], COSINE [23], DIAMOnD [24], DOMINO [21], GiGA [25], GXNA [26], KeyPathwayMiner [27–29] and GrandForest [30]) and two permutation-based methods (Hierarchical HotNet [19] and NetCore [20]). Although the classical methods were run with the full protocol (Figure 2A), we used a subset of the protocol for the two permutation-based methods (Figure 2B) since their runtime prohibits large-scale evaluation.

For the full protocol, we compared five widely used PPI networks (BioGRID [31], APID [32, 33], STRING [34], HPRD [35] and IID [36]), as well as gene expression and case/control data for five complex diseases: amyotrophic lateral sclerosis (ALS), non-small cell lung cancer (LC), ulcerative colitis (UC), Chron’s disease (CD) and Huntington’s disease (HD). For ALS and LC, we had access to survival data that we used for an additional evaluation. Moreover, we used five different random network generators, which produce randomized networks that preserve selected properties of the original PPI networks. For each PPI network, we generated 10 randomized counterparts with each generator. We then ran each classical AMIM on each of the 1275 network-disease pairs.

For each subnetwork produced for a network-disease pair, we measured two dimensions of meaningfulness: Firstly, predictive power quantified as mean MI [37] with (i) the phenotype and (ii) the survival data. Secondly, functional relevance quantified via (i) GSEA [38] w. r. t. Kyoto Encyclopedia of Genes and Genomes (KEGG) [39] pathways associated with the disease of interest and (ii) overlap coefficient w. r. t. disease-associated DisGeNET [40] gene sets. Finally, we used the one-sided Mann-Whitney U-test to assess whether the results obtained for the original PPI networks were significantly better than the results obtained for the randomized counterparts. Note that since AMIMs are intended for discovering yet unknown disease modules, the four meaningfulness scores employed in this paper should not be viewed as direct measures of performance but rather as proxy indicators for biological plausibility of the results.

For the slower permutation-based methods, we employed a restricted protocol using only the smallest PPI network (HPRD), the two smallest gene expression datasets (CD and HD) and the degree preserving network generator REWIRED. We selected this generator, because it produces the randomized networks that are most similar to the original PPI networks. We ran both permutation-based methods on each network-disease pair (in total 22 runs per method) and used the one-sided one-sample t-test to assess whether the subnetworks obtained for the original PPI networks were significantly more meaningful than the

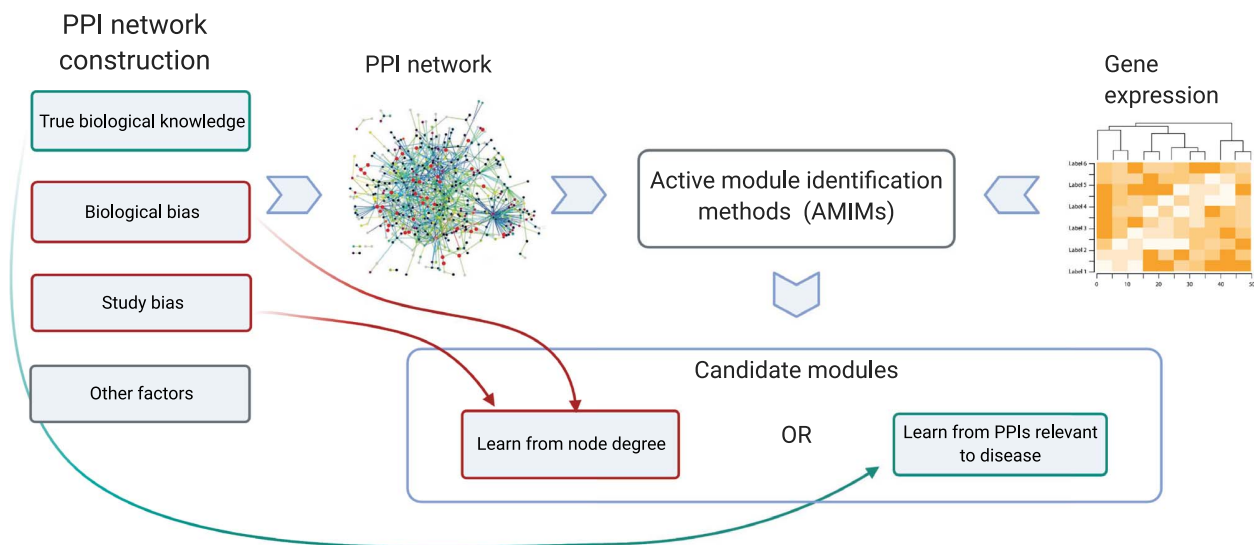


Figure 1. The limitation of AMIMs motivating this study. Since PPI networks suffer from technical and study bias, they usually contain hub-nodes with very high node degrees. In this study, we test the hypothesis whether AMIMs merely learn from the node degrees instead of exploiting the PPIs relevant to the disease of interest.

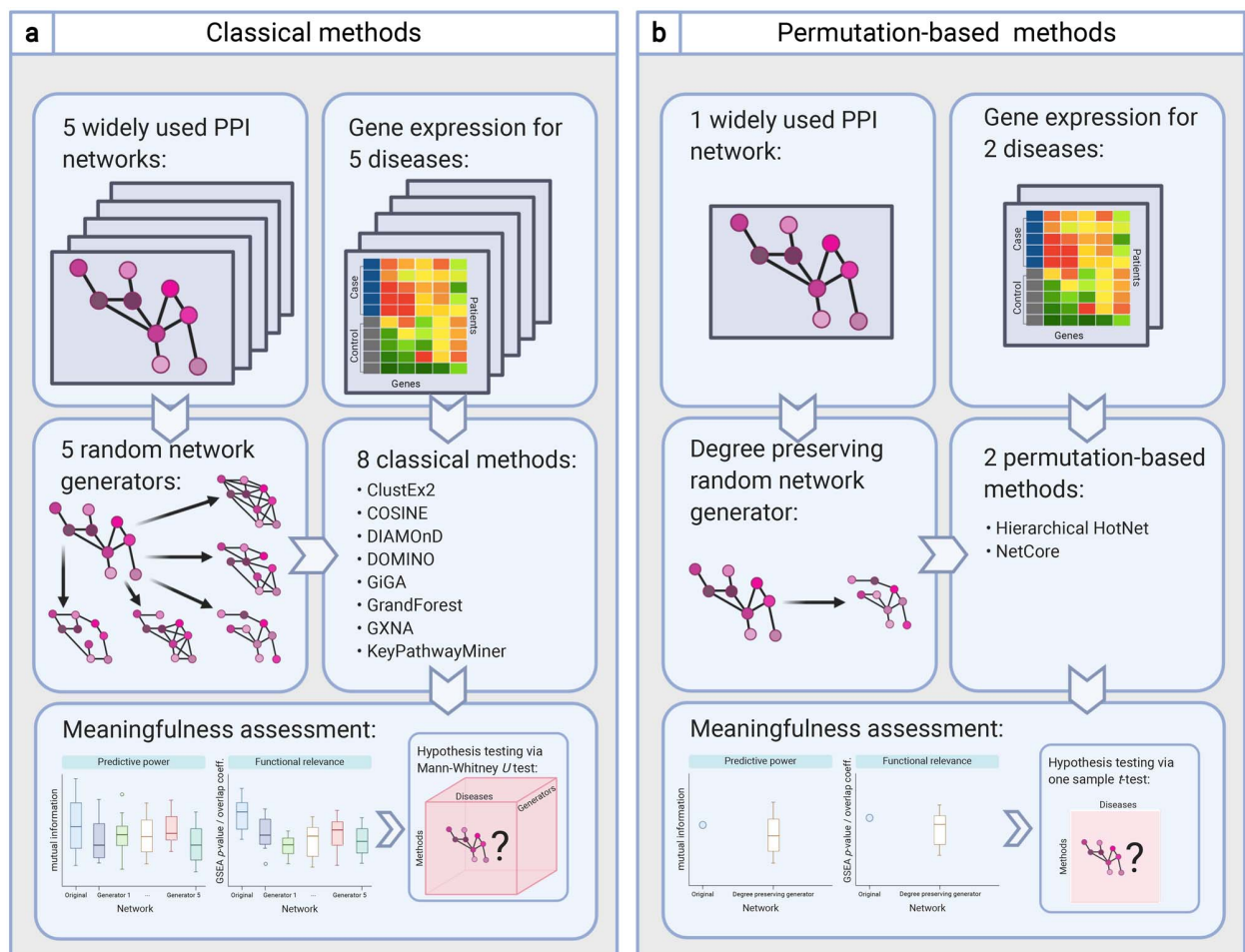


Figure 2. Test protocols employed in this study. (A) Large-scale protocol for classical methods. (B) Restricted protocol for slow permutation-based methods.

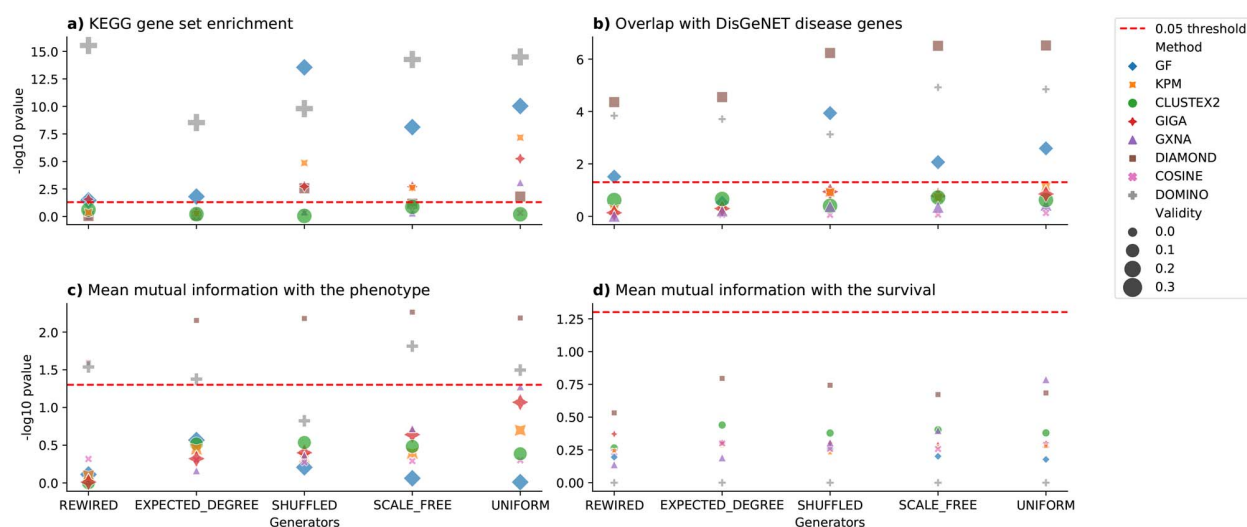


Figure 3. Log-transformed P-values for all classical AMIMs and all random network generators computed with the one-sided Mann–Whitney U-test. For each AMIM and each meaningfulness score, we computed a validity score (range from 0 to 1) as the fraction of the original network/condition pairs where the AMIM yielded a score $\geq \tau$ on the original PPI network. For the log-transformed GSEA P-values, we employed the cutoff $\tau = -\log_{10} 0.05$; for all other scores, we used the cutoff $\tau = 0.2$. The larger the validity scores, the larger the corresponding semi-transparent shapes.

subnetworks obtained for the random networks with prescribed node degrees.

Results for classical methods

Figure 3 visualizes the P-values obtained when comparing the results on the original PPI network to those on randomly generated networks for eight classical AMIMs separately (cf. Supplementary Figure 1 for visualizations of the distributions of the meaningfulness scores).

For the two scores quantifying predictive power (i. e. mean MI w. r. t. phenotype and survival times), we observe that, for most AMIMs, the scores of the candidate disease modules obtained on the original PPI networks are not significantly better than the scores obtained when using random graphs generated by any of the generators. This is the case even for the UNIFORM generator that produces networks that are structurally very different from the original PPI networks. For mean MI w. r. t. survival times (Figure 3D), no AMIM reaches the significance threshold of 0.05. For mean MI w. r. t. the disease phenotypes (Figure 3C), DIAMOND produces significant results compared with all random network generators but its solution on the original PPI receives a validity score of 0.0 (i. e. there was not a single original network-disease pair for which DIAMOND computed a candidate module whose mean MI with the phenotype reached 0.2). Notably, DOMINO produced significantly better solutions compared with all random network generators but SHUFFLED. DOMINO results are also slightly more meaningful, as they have a validity score > 0.0 . Most of the tested classical AMIMs hence fail to exploit the biological knowledge encoded in generic PPI networks for mining disease modules with high predictive power. DOMINO is the only tool to show potential w. r. t. the phenotype albeit with very low predictive power where the validity score does not exceed 0.1. All tools fail to produce disease modules that are predictive of survival time.

The two scores quantifying functional relevance (GSEA P-values w. r. t. disease-associated KEGG pathways and overlap coefficients w. r. t. disease-associated DisGeNET gene sets) present a different picture. Here, we observe that most

methods produce significantly more meaningful results on the original network compared with the SHUFFLED, SCALE_FREE and UNIFORM generators. However, when compared with structurally similar networks generated by the REWIRED and the EXPECTED_DEGREE generators, only DOMINO shows good performance. For KEGG gene set enrichment (Figure 3A), GrandForest and DOMINO reach the significance threshold, whereas DIAMOND and DOMINO do so for DisGeNET enrichment (Figure 3B) when compared with the two degree-preserving generators REWIRED and EXPECTED_DEGREE. Notably, DOMINO is the only tool to produce very significant results on degree-preserving random network generators. However, the validity scores are low in all cases and never exceed 0.3. Our results hence indicate that although most AMIMs are guided toward functionally relevant disease modules, the interactions themselves seem to be largely irrelevant.

To evaluate the effect of the five original PPI networks and the gene expression datasets for the five diseases, we also split the results along the PPI network dimension and along the disease dimension (cf. Supplementary Figures 2 and 3 for visualizations of the distributions of the meaningfulness scores). Figures 4 and 5 visualize the obtained P-values. These results suggest that HPRD is the best performing network in terms of KEGG gene set enrichment and DisGeNET overlap. This finding may be explained by the fact that HPRD is the smallest and least frequently updated network and contains mostly well-studied proteins that are more likely to overlap with KEGG pathways or DisGeNET genes.

We also observe that, in terms of functional relevance (especially DisGeNET overlap), the results for the CD dataset were much better than for the other datasets. This may be due to the fact that inflammation is a well-understood process and the DisGeNET disease gene annotation for CD is therefore better suited compared with other diseases. Note that the same argument does not apply to the UC dataset, since DisGeNET only reports on the more general inflammatory bowel disease as a proxy (cf. ‘Methods’ for details).

The results reported above suggest that, except for DOMINO, the tested AMIMs largely learn from the degree distributions rather than exploiting the biological knowledge encoded in the

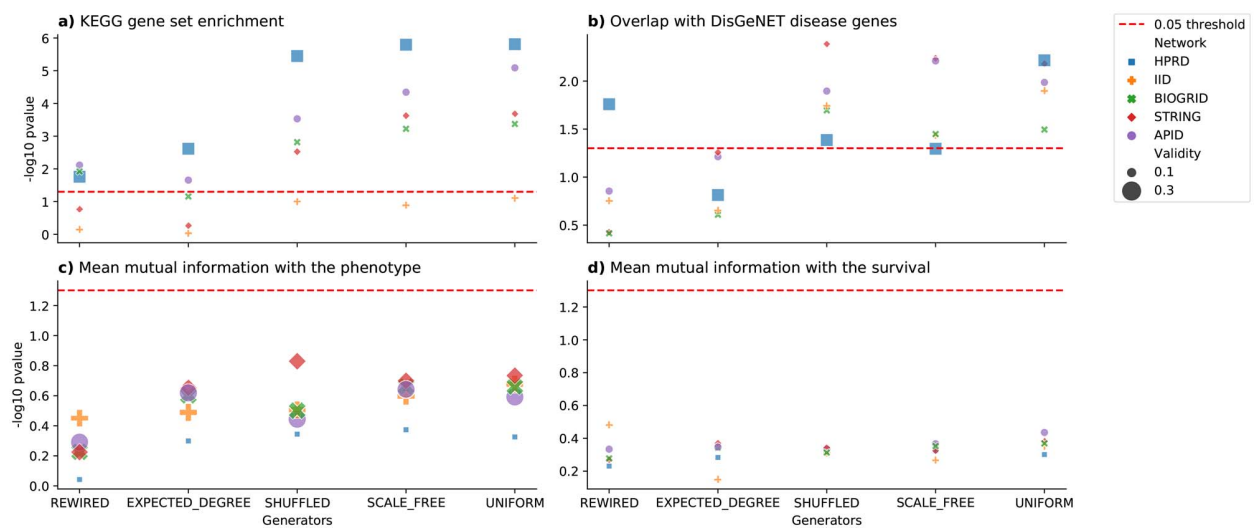


Figure 4. Log-transformed P -values for all PPI networks and all random network generators computed with the one-sided Mann–Whitney U-test.

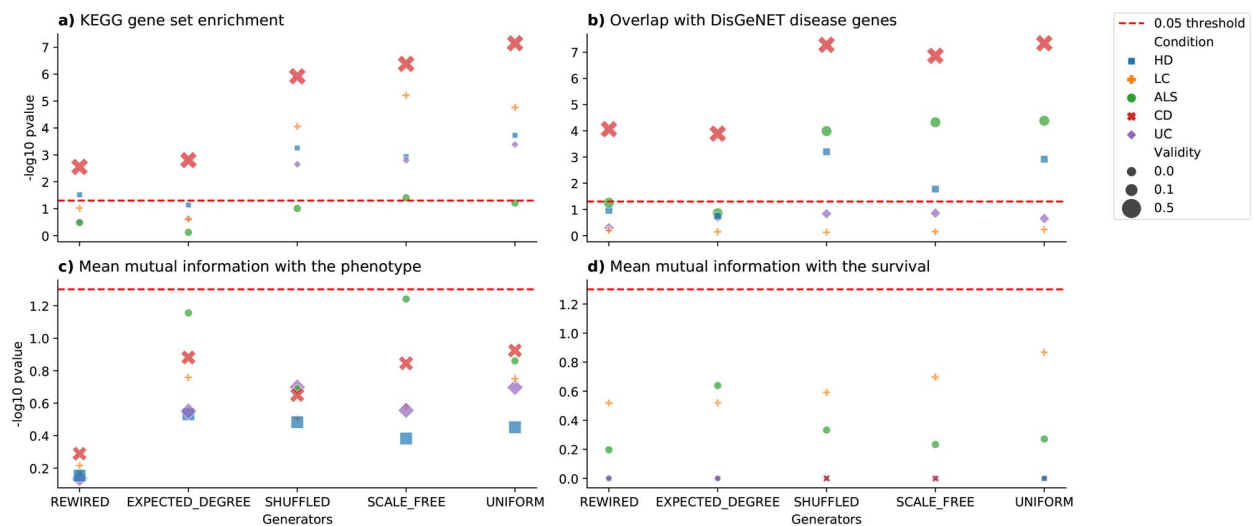


Figure 5. Log-transformed P -values for all diseases and all random network generators computed with the one-sided Mann–Whitney U-test.

interactions themselves. Figure 6 shows the outcomes of further analyses we carried out to find possible explanations for these results. The first interesting finding is that the topologies of the active modules DOMINO and COSINE computed on the original PPI networks are different from the topologies of the other AMIMs' modules (Figure 6A): DOMINO and COSINE's modules tend to have larger maximum pairwise distances, i. e. they tend to include fewer hub-nodes that would ensure a high connectivity. This is reflected by the fact that the mean degrees of the result sets and the two scores quantifying functional relevance are less strongly correlated for DOMINO and COSINE than for the other AMIMs (Figure 6E). These observations indicate that DOMINO and COSINE are less influenced by the node degrees than the other AMIMs. Although we expected this finding for DOMINO, it is somewhat surprising for COSINE. One possible explanation is that COSINE performed poorly even on the original PPI networks and hence neither learned from the node degrees nor from the interactions effectively.

We also observe several global trends in the results of the full protocol, which indicate that when aggregating across all tested AMIMs, the degrees on the genes contained in the

result sets are predictive of KEGG gene set enrichment P -value and DisGeNET overlap: Firstly, the mean degrees drop very significantly only for the SHUFFLED, the SCALE_FREE and the UNIFORM generators (Figure 6B). This reflects the results visualized in Figures 3–5, where significant drops in performance compared with the original PPI networks were observed mostly for these generators. Secondly, both the negative log-transformed KEGG gene set enrichment P -value and the DisGeNET overlap coefficient increase with increasing mean degrees (Figure 6C and D). Thirdly, we observe a very strong global correlation between the Mann–Whitney U-test P -values for the mean degrees, on the one side, and for two measures quantifying functional relevance, on the other side (last column in heat map in Figure 6E).

Results for permutation-based methods

Figure 7 shows the results for the two permutation-based AMIMs NetCore and Hierarchical HotNet (cf. Supplementary Figure 4 for visualizations of the distributions of the meaningfulness scores).

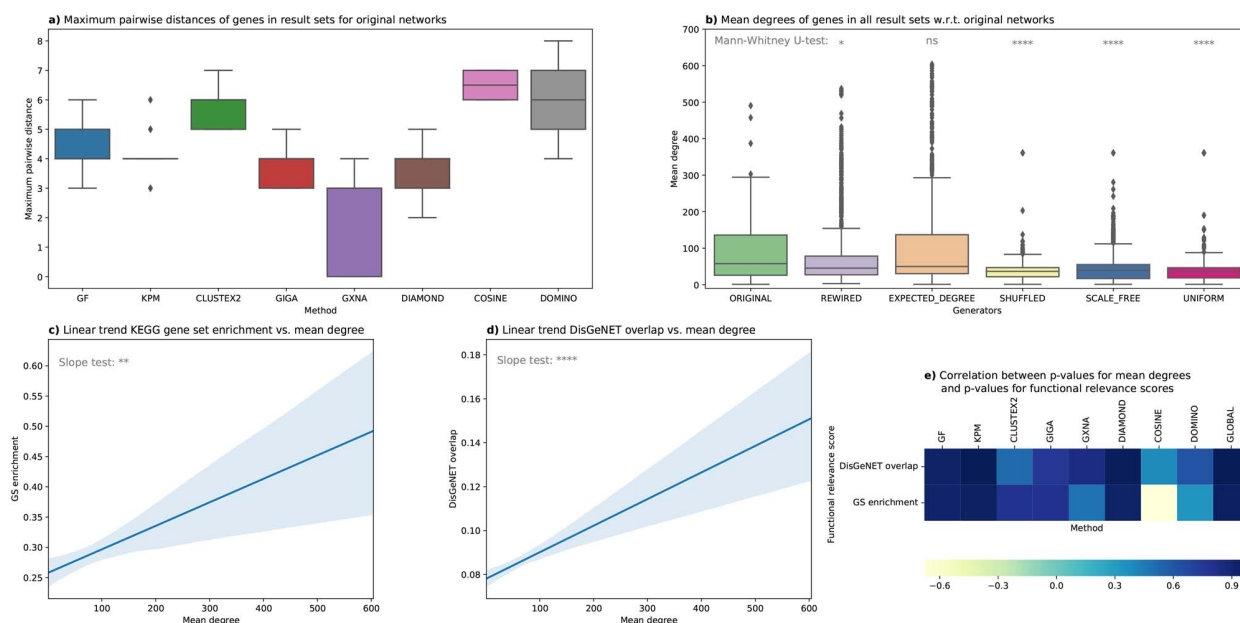


Figure 6. Detailed analyses explaining the results for the functional relevance scores. (A) Maximum pairwise distances of genes contained in result sets for original PPI networks for each AMIM. (B) Mean degrees in original PPI networks of genes contained in result sets for each generator. (C) Linear trend of KEGG gene set enrichment P -values versus mean degrees in original PPI networks aggregated across all generators and AMIMs. (D) Linear trend of DisGeNET overlap coefficient versus mean degrees in original PPI networks aggregated across all generators and AMIMs. (E) AMIM-specific and global correlation coefficients between Mann-Whitney U-test P -values for mean degrees in original PPI networks, on the one side, and the two functional relevance scores, on the other side (cf. 'Methods' for details).

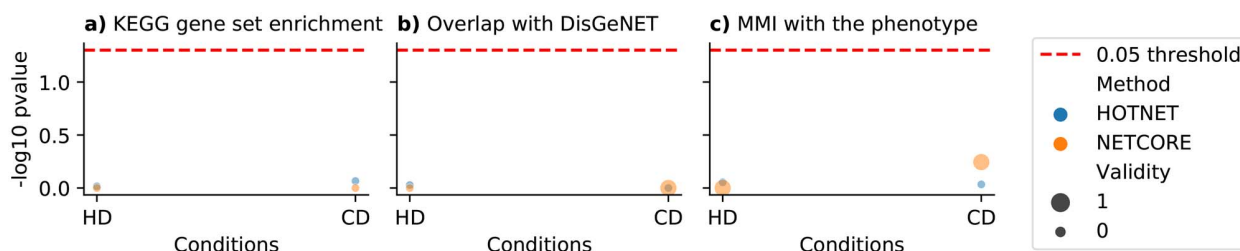


Figure 7. Log-transformed P -values for permutation-based AMIMs computed with the one-sided one-sample t -test. Mean MI w. r. t. survival times is not reported, because no survival data are available for the HD and CD datasets employed by the restricted protocol. The validity scores are binary here, because the restricted protocol uses only one original PPI network and the P -values are computed separately for the two diseases.

Recall that, because of their high computational costs, these methods were run with the restricted protocol visualized in Figure 2B, which only uses one PPI network (HPRD), two diseases (CD and HD), and one random network generator (REWIRE). Surprisingly, the results for the permutation-based methods are not better than the results for the classical AMIMs reported in the previous subsection: Both NetCore and Hierarchical HotNet clearly fail to reach the significance threshold of 0.05 for all three meaningfulness scores.

Discussion

It is commonly believed that prior biological knowledge captured in PPI networks can be leveraged for extracting functionally and mechanistically interpretable disease modules by AMIMs. However, an open question in the field is what characteristic of a PPI network makes these methods successful. Since PPI networks are known to suffer from a considerably node degree bias, we hypothesize here that AMIMs may use the node degree as prior

information rather than the connectivity of the network i. e. the biological knowledge captured in the interactions themselves. To test this hypothesis, we compared 10 state-of-the-art AMIMs on original as well as randomly generated networks. Although a few methods produced meaningful results w. r. t. functional enrichment, none of the methods produced disease modules with appreciable predictive power w. r. t. to both phenotype as well as survival. This demonstrates that results of AMIMs are not directly suited for such tasks without further refinement through e. g. supervised machine learning as shown by Alcaraz et al., where disease modules were used successfully as features for disease subtyping in a random forest classifier [10].

To investigate which network properties are exploited by AMIMs, we compared the results against different types of random network generators. Our results clearly show that most methods do not yield more meaningful candidate disease modules on randomized networks if these are constructed such that the (expected) node degrees match the node degrees of the original networks. Unexpectedly, permutation-based methods

that include steps to correct for PPI network characteristics in their workflow did not produce more meaningful results on the original PPIs.

Only one tested tool benefits from the PPI networks

The only tool to produce more meaningful results on the original network was the recently proposed method DOMINO [21]. Interestingly, DOMINO's development was motivated by the observations that existing methods are not sensitive to permutations of the input data. Our finding that most existing methods are also not sensitive to network randomization suggests that these two issues are related. Considering the small diameter of a PPI network, AMIMs sensitive to high-degree nodes are likely to produce subnetworks that are enriched for similar biological functions. Given this, we see several possibilities to advance the field, namely (i) further algorithmic improvements to overcome PPI network biases, (ii) the integrative use of complementary omics data that increase the signal to noise ratio, which is inherently low in gene expression data and (iii) the use of more fine-grained tissue-specific, condition-specific or even personalized networks.

Further algorithmic improvements are needed

The encouraging results of DOMINO indicate that algorithmic improvements to overcome the node bias of PPI networks are possible. From an algorithmic point of view, DOMINO differs from all other tested AMIMs in that it discards some of the disease-associated genes in a partially unsupervised manner. We hypothesize that this is the key to DOMINO's success, because it makes hub-genes other AMIMs include into their modules to connect the disease-associated genes less attractive for DOMINO (cf. Figure 6A). Consequently, we expect algorithmically improved AMIMs to be either partially unsupervised such as DOMINO or even fully unsupervised [41]. Importantly, newly developed AMIMs need to be tested with respect to their sensitivity to network randomization. One way to do this systematically is to evaluate them with the test suite presented in this paper.

Different types of omics data and context-specific networks should be considered

For this study, we evaluated the performances of AMIMs when run on gene expression data and PPI networks, which is currently the most common use case. Consequently, our findings are restricted to this use and cannot be generalized to other types of omics data and biological networks. In fact, we expect that using different types of omics data and context-specific networks introduces new opportunities for AMIM users and developers. For instance, promising directions for future research include using microbiome data in combination with metabolic networks, using DNA methylation data with gene regulatory networks [42], or inferring condition- or tissue-specific gene regulatory networks from expression data [43]. Next-generation AMIMs might even integrate the inference of context-specific networks with disease module mining. Although all of these strategies come with their own challenges and limitations, we believe that they could help to overcome some of the biases of PPI networks (especially, the literature bias).

Quantitative measures of functional relevance need to be used carefully

The most widely used method for quantitatively assessing the functional relevance of candidate disease modules is to compare them against known disease-associated genes. In fact, we also follow this strategy in our test suite (recall that we use KEGG GSEA *P*-value and DisGeNET overlap as our measures of functional relevance). However, this approach is severely limited and biased by our current knowledge. In particular in the light of the results shown here, it must hence always be kept in mind that such quantitative measures are at best proxy indicators for functional relevance. Alternatively, the simulation of synthetic gold standard datasets could be considered, but this approach is limited by our understanding and assumptions on network and disease module characteristics [6].

Cross-disciplinary research is key to success

Since quantitative measures of functional relevance are biased, it is unlikely that simply reporting on disease modules will yield novel insights into complex diseases. Interestingly, studies that report successful applications of active module identification are usually co-authored by cross-disciplinary teams of researchers that include not only bioinformaticians but also domain experts for the disease of interest [11–16]. We argue that this is no coincidence and promote cross-disciplinary research. To this end, AMIM developers should follow best practices for developing usable software [44], allowing domain experts without a background in computer science to run the tools on their data and to leverage their domain knowledge in the interpretation of the results. Ideally, such interfaces should follow the expert-in-the-loop paradigm and provide functionality for all three steps of active module identification (data integration, network construction, disease module mining). To the best of our knowledge, such an integrated active module identification platform is available only for COVID-19 [45].

Conclusions

A plethora of tools for identifying disease modules via the integration of gene expression data and PPI networks have been developed over the years. Here, we could show conclusively that most AMIMs do not produce more meaningful results on the original compared with randomized PPI networks in which the (expected) node degrees do not change. Our results indicate that classical but also supposedly bias-aware AMIMs extract disease modules based on the node degree rather than benefiting from the interactions of the nodes. Only a single recently proposed method, DOMINO, showed significantly better results on the original PPI network, suggesting that the development of better algorithmic approaches as well as less biased, context-specific networks are urgently needed to provide the biomedical community with the necessary tools to deliver on the promises that the field of active (disease) module identification and *de novo* network enrichment made almost two decades ago.

Methods

PPI networks and random network generators

We ran our test protocol on five widely used PPI networks: BioGRID [31], APID [32, 33], STRING [34] with high confidence interactions only (score ≥ 0.7), HPRD [35] and IID [36] with experimentally validated interactions only. Key properties of the

PPI networks are summarized in Supplementary Table 1. All networks have one giant largest connected component with a very small diameter. Note that although BioGRID, APID, STRING and IID are continuously updated, HPRD is no longer maintained and has not been updated since 2010. However, HPRD is still useful for our study, because it is smaller and focuses on well-studied interactions. Moreover, some of the tested AMIMs were designed with HPRD in mind, which was the largest network available at the time of implementation.

We used five different random network generators, which were chosen to produce randomized networks that preserve selected properties of the original PPI networks:

REWired: *degree preserving generator.* Repeatedly swaps pairs of edges and non-edges to produce random networks whose degree sequences are identical to the degree sequences of the original PPI networks [46, 47]. Preserves the individual node degrees and hence the hub-genes.

EXPECTED_DEGREE: *expected degree preserving generator.* Creates networks with randomly sampled edges where the sampling probabilities are chosen such that the expected node degrees correspond to the node degrees in the original PPI networks [48, 49]. Preserves individual node degrees and hub-genes in expectation.

SHUFFLED: *topology preserving generator.* Shuffles the gene IDs. Preserves the degree sequence and the topology but not the individual node degrees and the hub-genes.

SCALE_FREE: *scale-free generator.* Produces scale-free networks using the Barabási-Albert model [50]. The parameters are chosen such that the numbers of nodes and edges in the random networks match the numbers of nodes and edges in the original PPI network. Preserves neither the topology nor the individual node degrees or the hub-genes, but produces networks that are structurally similar to the original PPI networks, since PPI networks are usually scale-free [51, 52].

UNIFORM: *uniform generator.* Produces random graphs using the Erdős-Rényi model [53]. The parameters are chosen such that the numbers of nodes and edges in the random networks matches the numbers of nodes and edges in the original PPI network. The produced networks are very different from the original PPI networks. In particular, their degrees are binomially distributed, whereas PPI networks tend to have power law degree distributions [51, 52].

Expression, phenotype and survival data

For testing we considered gene expression datasets for five different diseases: ALS, non-small cell LC, UC, CD and HD. For all datasets, case/control phenotype data are available, whereas for ALS, LC and HD, survival data are also reported. Gene probes were mapped to Entrez gene IDs, and if multiple probes corresponded to a single gene, the median value was used. Key properties of the expression datasets are summarized in Supplementary Table 3.

In the LC dataset, we only considered non-small cell LC patients due to their significant biological difference from small cell LC and the larger number of available samples. For the HD dataset, we preselected samples such that the most distinct gene expression difference is present. To achieve this, we only used samples from caudate nucleus, since this region has been reported to have the largest change in gene expression [54]. As a case group, only patients with Vonsattel grades 2–4 were

considered, whereas samples with Vonsattel grade 0–1 were discarded.

AMIMs and method-specific preprocessing

In the past years, various AMIMs have been presented (cf. Batra et al. [6] for a benchmarking paper and Lazareva et al. [9] for a systematic review). Here, we selected 10 tools, namely ClustEx2 [22], COSINE [23], DIAMOnD [24], DOMINO [21], GiGA [25], GXNA [26], KeyPathwayMiner [27–29], GrandForest [30], Hierarchical HotNet [19] and NetCore [20] (cf. Supplementary Table 4 for details). These tools were selected for three reasons:

- They require expression data and phenotypes or input formats that can be derived from these data.
- They return a gene set representing a candidate disease module.
- They are available online and sufficiently bug-free and documented to allow integration in our test suite.

Hierarchical HotNet and NetCore are permutation-based methods, i. e. they include data or network randomization steps in their workflows to correct for typical PPI network biases. All other tools use the PPI networks without applying any corrections.

To set the hyper-parameters of the AMIMs, we used default values whenever available. For parameters where no default values are provided in the implementations, we used the values chosen in the tutorials, READMEs, or original publications. For tools that return several candidate disease modules, we always used the union of all reported subnetworks. We hence did not carry out hyper-parameter tuning. The reason for this is 3-fold: Firstly, hyper-parameter tuning would have been computationally infeasible, since already without our protocol required more than 10 000 AMIM runs. Secondly, our aim is not to obtain the optimal results but to test if equally good results can be obtained using a random network. Thirdly, because of the large number of AMIM runs, small changes in the results for a specific AMIM have little effect on the overall conclusions. Note, however, that since we did not optimize the tools, our findings should not be interpreted as a benchmark but rather as an evaluation of the effect of network biases on AMIMs.

Although COSINE, GXNA and GrandForest can be run directly on the normalized expression data, the other tools require different input formats. More specifically, ClustEx2, DIAMOnD and DOMINO expect a list of disease-associated seed genes, Hierarchical HotNet and NetCore expect gene scores, GiGA expects a sorted list of genes, and KeyPathwayMiner expects an indicator matrix of genes that are differentially expressed in the case samples.

For each gene g , let \mathbf{x}_g^1 and \mathbf{x}_g^0 be the vectors of expression values for all case and control samples, and $x_{g,s}$ be the expression value for sample s . Furthermore, let n be the number of genes contained in the expression dataset and m be the number of case samples. To derive gene scores, seed genes and sorted gene lists from the expression data, we evaluated the two-sided Mann-Whitney U-test on \mathbf{x}_g^1 and \mathbf{x}_g^0 to obtain P-values P_g of differential expression for all genes g . We then defined gene scores as $-\log_{10}(P_g)$, used all genes g with $P_g < 0.001/n$ as seed genes, and obtained sorted lists of genes by sorting the genes in non-decreasing order of p_g . The indicator matrix $M = (m_{g,s}) \in \{0, 1\}^{n \times m}$ required by KeyPathwayMiner was defined as $m_{g,s} = [|x_{g,s} - \text{mean}(\mathbf{x}_g^0)| > 1.5 \cdot \text{std}(\mathbf{x}_g^0)]$, where s is a case sample, $[\cdot]$ is the Iverson bracket (i. e. $[\text{true}] = 1$ and $[\text{false}] = 0$),

and the operators $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ denote mean and standard deviation, respectively.

Evaluation metrics

Quantitative measures are needed to evaluate how well AMIMs perform on the original and on the randomized PPI networks. That is, we need to quantify the meaningfulness of the gene sets S returned by the tools. For this, we distinguish two dimensions of meaningfulness: predictive power w. r. t. the phenotype and survival time, and functional relevance for the disease of interest.

For quantifying predictive power, we employed MI, which is widely used for selecting features with high predictive power. More precisely, let \mathbf{y} be the vector of case/control disease phenotypes and \mathbf{x}_g be the vector of expression values of all samples for a gene $g \in S$. We computed the mean MI w. r. t. the phenotype $\sum_{g \in S} \text{MI}(\mathbf{x}_g, \mathbf{y})/|S|$ between \mathbf{y} and \mathbf{x}_g across all genes $g \in S$. Analogously, the mean MI w. r. t. the survival times was computed as $\sum_{g \in S} \text{MI}(\mathbf{x}_g, \mathbf{t})/|S|$, where \mathbf{t} denotes the vector of survival times. The larger the mean MI, the stronger the association between the expression data for the genes contained in S and, respectively, the disease phenotypes and the survival times.

To quantify functional relevance, we computed the mean negative log-transformed GSEA P -values between the result sets S and the KEGG [39] pathways related to the disease of interest. The disease-to-pathway mappings are shown in the Supplementary Table 2. Moreover, we computed the overlap coefficients $|S \cap D|/\min\{|S|, |D|\}$ between the results sets S and the disease-associated DisGeNET [40] gene sets D . These gene sets were obtained by taking all genes connected to the condition of interest in DisGeNET. Only for the UC dataset there was no exact match. Therefore, we used genes associated with inflammatory bowel disease of which UC is a subtype. The full DisGeNET diseases IDs mapping to the conditions is shown in the Supplementary Table 2. Note that, for all four meaningfulness scores, larger means better.

Let O be a batch of meaningfulness scores obtained for one of the original PPI networks and R be a batch of scores obtained for randomized counterparts generated by one of the random network generators described above. In the large-scale protocol used for the classical AMIMs (Figure 2A), we used the one-sided Mann–Whitney U-test to assess whether the scores contained in O are significantly larger than the scores contained in R . In the restricted protocol used for the permutation-based methods (Figure 2B), the Mann–Whitney U-test is not applicable, because we have $|O| \leq 4$ for each partitioning of the results (there are only four runs on the original PPI networks). Consequently, we partitioned the results along the methods and disease dimensions to ensure $|O| = 1$ and instead used the one-sided one-sample t -test.

Although the P -values from the one-sided Mann–Whitney U-test and the one-sided one-sample t -test tell us whether the candidate disease modules computed for the original PPI networks are significantly more meaningful than those obtained for the randomized counterparts, they are oblivious to the question if the candidate disease modules for the original PPI networks are sufficiently meaningful in absolute terms. Assume, for instance, that O and R contain negative log-transformed GSEA P -values, that the values contained in O fall into the range $[0.5, 1]$ and that the values contained in R fall into the range $[0.2, 0.5]$. Then the one-sided Mann–Whitney U-test will return a significant P -value, which, however, should be treated with extreme caution because the scores in O are themselves not significant. To account for this fact, we computed a validity score

$\{|\{o \in O \mid o \geq \tau\}|/|O|\}$ for each P -value computed by the one-sided Mann–Whitney U-test and the one-sided one-sample t -test. For the negative log-transformed GSEA P -values, the threshold was set to $\tau = -\log_{10} 0.05$; for all other scores, we used $\tau = 0.2$. In Figures 3 and 7 and Supplementary Figures 1 and 2, the validity scores are visualized as the sizes of the shapes corresponding to the P -values.

Let O be a batch of result sets obtained for one of the original PPI networks, R be a batch of result sets obtained for randomized counterparts, and $\text{avdeg}(S)$ denote the mean degree of a gene set S , computed w. r. t. the original PPI network. For further analyzing the results of the full protocol, we used the one-sided Mann–Whitney U-test to assess whether the mean degrees $\{\text{avdeg}(S) \mid S \in O\}$ of the gene sets for the original networks are significantly larger than the mean degrees $\{\text{avdeg}(S) \mid S \in R\}$ obtained for the randomized counterparts (cf. Figure 6B and E). By splitting along the AMIM dimension, we obtain an array of P -values for each AMIM with entries for each network generator. The correlation coefficients of these arrays with the arrays of AMIM-specific P -values obtained for the meaningfulness scores visualized in Figure 3 indicate to which extent the AMIMs merely learn from the degree distributions of the PPI networks. The larger the correlation coefficient, the stronger the impact of the degrees of the genes contained in the result sets on the meaningfulness scores (cf. Figure 6E).

Implementation

The overall architecture of our test suite is implemented in Python 3 and schematically visualized in Supplementary Figure 5. Each tested AMIM is wrapped into an implementation of an abstract `AlgorithmWrapper` interface. The wrappers run the AMIMs via system calls to the original executables. Graph operations and random network generators are implemented with NetworkX [55] and graph-tools [56]. GSEA is carried out via the GSEAPy interface of the Enrichr API [57], and statistical tests are implemented with SciPy [58].

To reproduce the results reported in this paper, it suffices to execute the top-level Python script `run_tests.py`, which is shipped with our test suite. If developers of new AMIMs would like to use our test suite for evaluating their methods, they can provide a custom implementation of the `AlgorithmWrapper` interface. Our test suite can hence be used to easily benchmark new AMIMs against the 10 pre-implemented existing methods. Our test suite is available at <https://github.com/dbblumenthal/amim-test-suite/>, along with a detailed README and all data needed to reproduce the experiments.

Key Points

- Most AMIMs only learn from the node degrees but not from the biological knowledge encoded in the edges of PPI networks.
- Only the recently presented AMIM DOMINO yields significantly more meaningful disease modules if run on original PPI networks rather than on randomized counterparts with preserved node degrees.
- Better algorithmic approaches and less biased, context-specific networks are urgently needed in the field of active module identification and *de novo* network enrichment.

Availability

The KEGG pathways were obtained from KEGG: <https://www.genome.jp/kegg/disease/>. BioGRID (v3.2.149), APID (v1.0), STRING (version 11.0) and HPRD (release 9) as well as DisGeNET (v7.0) were obtained using nDEx [59–61]. The IID network (v2018-11) was downloaded from <http://iid.ophid.utoronto.ca/>. All gene expression datasets and corresponding metadata were retrieved from Gene Expression Omnibus [62], using the GEO2R R interface (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>). The associated GEO accession codes are shown in Supplementary Table 2. The entire test-suite (Python environment, tool executables, PPI networks, expression datasets) is available at <https://github.com/dbblumen/aml/amim-test-suite/>.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

J.B. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant no. 777111 and grant no. 826078). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. J.B. and M.L. were supported by the German Federal Ministry of Education and Research (BMBF) within the e:Med framework (grant no. 01ZX1908A). J.B. was supported by the German Federal Ministry of Education and Research (BMBF) within the e:Med framework (grant no. 01ZX1910D) and within the CLINSPECT-M framework (grant no. 031L0214A). Contributions by O.L. are funded by the Bavarian State Ministry of Science and the Arts within the framework coordinated by the Bavarian Research Institute for Digital Transformation (bidt, Doctoral Fellow). Figures 1 and 2 were created with [BioRender.com](https://www.biorender.com/).

References

- Perou CM, Srlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000; **406**(6797): 747–52.
- Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014; **511**(7511): 543–50.
- Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; **21**(11): 1350–6.
- van Vliet MH, Reyat F, Horlings HM, et al. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics* 2008; **9**:375.
- Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011; **7**(10):e1002240.
- Batra R, Alcaraz N, Gitzhofer K, et al. On the performance of de novo pathway enrichment. *NPJ Syst Biol Appl* 2017; **3**:6.
- Silverman EK, Schmidt HHHW, Anastasiadou E, et al. Molecular networks in network medicine: development and applications. *Wiley Interdiscip Rev Syst Biol Med* 2020; **12**(6):e1489.
- Maron BA, Altucci L, Balligand J-L, et al. A global network for network medicine. *NPJ Syst Biol Appl* 2020; **6**(1): 29.
- Lazareva O, Lautizi M, Fenn A, et al. Multi-omics analysis in a network context. In: Olaf Wolkenhauer. In: *Systems Medicine*. Oxford: Academic Press, 2021, 224–33.
- Alcaraz N, List M, Batra R, et al. De novo pathway-based biomarker identification. *Nucleic Acids Res* 2017; **45**(16): e151.
- Samokhin AO, Stephens T, Wertheim BM, et al. NEDD9 targets COL3A1 to promote endothelial fibrosis and pulmonary arterial hypertension. *Sci Transl Med* 2018; **10**(445):eaap7294.
- Wang R-S, Loscalzo J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J Mol Biol* 2018; **430**(18, Part A): 2939–50.
- Amitabh Sharma, Arda Halu, Julius L Decano, et al. Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates type 2 diabetes associated genes. *NPJ Syst Biol Appl* 4:25, 2018.
- AbdulHameed MDM, Tawa GJ, Kumar K, et al. Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One* 2014; **9**(11):e112193.
- Halu A, Liu S, Baek SH, et al. Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Hum Mol Genet* 2019; **28**(14): 2352–64.
- Sharma A, Menche J, Chris Huang C, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 2015; **24**(11): 3005–20.
- Stibius KB, Sneppen K. Modeling the two-hybrid detector: experimental bias on protein interaction networks. *Biophys J* 2007; **93**(7): 2562–2.
- Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet* 2015; **6**:260.
- Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* 2018; **34**(17): i972–80.
- Barel G, Herwig R. NetCore: a network propagation approach using node coreness. *Nucleic Acids Res* 2020; **48**(17): e98.
- Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021; **17**(1): e9593.
- Ding Z, Guo W, Gu J. ClustEx2: gene module identification using density-based network hierarchical clustering. In *CAC* 2018; **2018**:2407–12.
- Ma H, Schadt EE, Kaplan LM, et al. COSINE: COndition-specific sub-NEtwork identification using a global optimization method. *Bioinformatics* 2011; **27**(9): 1290–8.
- Ghiassian SD, Menche J, Barabási A-L. A Disease Module detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015; **11**(4).
- Breitling R, Amtmann A, Herzyk P. Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinform* 2004; **5**:100.
- Nacu S, Critchley-Thorne R, Lee P, et al. Gene expression network analysis and applications to immunology. *Bioinformatics* 2007; **23**(7): 850–8.
- Nicolas Alcaraz, Hande Küçük, Jochen Weile, et al. KeyPathwayMiner: detecting case-specific biological pathways using expression data. *Internet Mathematics*, **7**(4): 299–313, 2011.

28. Alcaraz N, Pauling J, Batra R, et al. KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with cytoscape. *BMC Syst Biol* 2014; **8**(99).
29. List M, Alcaraz N, Dissing-Hansen M, et al. KeyPathwayMiner-Web: online multi-omics network enrichment. *Nucleic Acids Res* 2016; **44**(Webserver-Issue): W98–104.
30. Larsen SJ, Schmidt HHHW, Baumbach J. De novo and supervised endophenotyping using network-guided ensemble learning. *Systems Medicine* 2020; **3**(1): 8–21.
31. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019; **47**(D1): D529–41.
32. Alonso-Lpez D, Gutierrez MA, Lopes KP, et al. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res* 2016; **44**(W1): W529–35.
33. Alonso-Lpez D, Campos-Laborie FJ, Gutierrez MA, et al. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database* 2019; **2019**.
34. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; **47**(D1): D607–13.
35. Keshava Prasad TS, Goel R, Kandasamy K, et al. Human protein reference database–2009 update. *Nucleic Acids Res* 2009; **37**(Database issue): D767–72.
36. Kotlyar M, Pastrello C, Malik Z, et al. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res* 2019; **47**(D1): D581–9.
37. Ross BC. Mutual information between discrete and continuous data sets. *PLoS ONE* 2014; **9**(2):e87357.
38. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102**(43): 15545–50.
39. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**(D1): D457–62.
40. Piero J, Ram-rez-Anguita JM, Sach-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020; **48**(D1): D845–55.
41. Lazareva O, Canzar S, Yuan K, et al. BiCoN: network-constrained biclustering of patients and omics data. *Bioinformatics* 2020.
42. Wu J, Gu Y, Xiao Y, et al. Characterization of DNA methylation associated gene regulatory networks during stomach cancer progression. *Front Genet* 2018; **9**:711.
43. Selber-Hnatiw S, Sultana T, Tse W, et al. Metabolic networks of the human gut microbiota. *Microbiology* 2020; **166**(2): 96–119.
44. List M, Ebert P, Albrecht F. Ten simple rules for developing usable software in computational biology. *PLoS Comput Biol* 2017; **13**(1):e1005265.
45. Sadegh S, Matschinske J, Blumenthal DB, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat Commun* 2020; **11**(1): 3518.
46. Gkantsidis C, Mihail M, Zegura EW. The markov chain simulation method for generating connected power law random graphs. In: Ladner RE (ed). *ALLENEX 2003*. SIAM, 2003, 16–25.
47. Viger F, Latapy M. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *J Complex Networks* 2016; **4**(1): 15–37.
48. Chung F, Lu L. Connected components in random graphs with given expected degree sequences. *Ann Combinatorics* 2002; **6**(2): 125–45.
49. Joel C. Miller and Aric A. Hagberg. Efficient generation of networks with given expected degrees. In Alan M. Frieze, Paul Horn, and Pawel Pralat, editors, *WAW 2011*, volume **6732** of LNCS, pages 115–26, Berlin, Heidelberg, 2011. Springer.
50. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science* 1999; **286**(5439): 509–12.
51. Jeong H, Mason SP, Barabási AL, et al. Lethality and centrality in protein networks. *Nature* 2001; **411**(6833): 41–2.
52. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev Genet* 2004; **5**(2): 101–13.
53. Erdős P, Rényi A. On random graphs I. *Publ Math Debrecen* 1959; **6**:290.
54. Hodges A, Strand AD, Aragaki AK, et al. Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet* 2006; **15**(6): 965–77.
55. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkx. In: Varoquaux G, Vaught T, Millman J (eds). *SciPy 2008*. Pasadena, 2008, 11–5.
56. Peixoto TP. The graph-tool python library. *figshare* 2014.
57. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**(W1): W90–7.
58. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020; **17**:261–72.
59. Pratt D, Chen J, Welker D, et al. NDEx, the network data exchange. *Cell Syst* 2015; **1**(4): 302–5.
60. Pratt D, Chen J, Pillich R, et al. NDEx 2.0: a clearinghouse for research on cancer pathways. *Cancer Res* 2017; **77**(21): e58–61.
61. Pillich RT, Chen J, Rynkov V, et al. NDEx: a community resource for sharing and publishing of biological networks. *Methods Mol Biol* 2017; **1558**:271–301.
62. Barrett T, Willhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013; **41**(Database issue): D991–5.

4.3. Publication 3: Machine learning for deciphering cell heterogeneity and gene regulation

Citation

The article titled "Machine learning for deciphering cell heterogeneity and gene regulation" has been published online at Nature Computational Science on 15 March 2021 (in print on March 2021).

Full citation:

"Scherer M, Schmidt F, Lazareva O, Walter J, Baumbach J, Schulz MH, List M. Machine learning for deciphering cell heterogeneity and gene regulation. Nature Computational Science 1, 183–191 (2021). <https://doi.org/10.1038/s43588-021-00038-7>"

Summary

Epigenetics data availability has significantly improved in recent years, making the use of machine learning technologies possible. The limitations of PPI networks pushed towards discovery of stable epigenetic mechanisms to bring additional value and robustness to disease mechanism discovery. Thus, in this publication we aimed to describe state of the art in epigenomics level machine learning and discuss possible future improvements. We have collected information about 53 methods, separated into three different categories:

- methods that aim to address cell-type heterogeneity and in particular contribute to deconvolution of DNA methylation;
- methods for prediction of gene expression using epigenomic data;
- methods that contribute to gene regulation understanding on a single-cell level.

The methods differ in terms of the input data (e.g., open chromatin data, ChIP-seq, DNA methylation, transcriptomics data) and the used computational approaches.

We identified several main limitations of the field: lack of technology for single-cell ChIP-seq analysis, single-cell-based signatures are not yet used for deconvolution analysis, transcriptional and post-transcriptional gene regulation remain a serious challenge in the field as well as the effect of post-translational modifications on the level of proteins. Machine learning field also needs to rise to the challenge and make better use of the available data and known prior information. All in all, we observed a rising demand in the scientific community for robust and interpretable epigenomics data analysis methods that can push the field forward, helping to obtain the essential knowledge about epigenomic regulation.

Contribution

As described in the publication: *"M.L. conceived this Review, supervised and contributed to the writing of the manuscript. M.S., F.S. and O.L. collected information about the tools and databases*

and wrote the corresponding parts of the manuscript. M.H.S., J.B. and J.W. provided critical feedback on the manuscript. All authors discussed the results and contributed to the final manuscript. These authors contributed equally: Michael Scherer, Florian Schmidt, Olga Lazareva. "

In detail: Dr. M. Scherer, Dr. F. Schmidt, O. Lazareva were responsible for collecting information for three main sections: "Dissecting cell-type heterogeneity", "Understanding gene expression using epigenomic data", "Gene regulation on the single-cell level". I contributed to the later section, "Gene regulation on single-cell level". I collected, systemized, and described algorithms and methods for single-cell epigenomics analysis. I also collected information about databases for epigenetic data (Table 1 in the manuscript), wrote the first draft of the section and produced figures. Dr. M. List supervised the project and finalized the manuscript.

Rights and permissions

The original article is embedded with permission of Springer Nature. All rights belong to Springer Nature.

Additional supplementary material

Supplementary data are available online at Nature Computational Science <https://doi.org/10.1038/s43588-021-00038-7>.



Machine learning for deciphering cell heterogeneity and gene regulation

Michael Scherer^{1,2,3,9}, Florian Schmidt^{4,9}, Olga Lazareva^{5,9}, Jörn Walter², Jan Baumbach^{5,6,7}, Marcel H. Schulz⁸ and Markus List⁵✉

Epigenetics studies inheritable and reversible modifications of DNA that allow cells to control gene expression throughout their development and in response to environmental conditions. In computational epigenomics, machine learning is applied to study various epigenetic mechanisms genome wide. Its aim is to expand our understanding of cell differentiation, that is their specialization, in health and disease. Thus far, most efforts focus on understanding the functional encoding of the genome and on unraveling cell-type heterogeneity. Here, we provide an overview of state-of-the-art computational methods and their underlying statistical concepts, which range from matrix factorization and regularized linear regression to deep learning methods. We further show how the rise of single-cell technology leads to new computational challenges and creates opportunities to further our understanding of epigenetic regulation.

DNA is the molecular basis of the genome and relies on a four-letter code of the nucleobases adenine, cytosine, guanine and thymine to store information. These nucleobases are combined with a sugar backbone and a phosphate group, yielding four corresponding nucleotides, which are concatenated to form the DNA. Information units called genes are transcribed into RNA and serve as a blueprint for the assembly of proteins during translation. In contrast to bacteria, which have a small circular genome, eukaryotic cells afford a much larger genome; for example, a human cell fits DNA of 2 m in length into a nucleus of only a few micrometers in diameter¹. This is feasible since only a fraction of the genome is needed at any given time, allowing cells to compact most of the DNA. Compaction is achieved by tightly wrapping DNA around histone protein complexes, known as nucleosomes. Compacted DNA is referred to as chromatin, which can exist in a dense state known as heterochromatin or a more accessible state referred to as euchromatin (Fig. 1a). A complex regulatory machinery can selectively unpack DNA from hetero- into euchromatin. The degree of openness is also referred to as chromatin accessibility and influences the level of gene expression, that is, the amount of messenger RNA (mRNA) that can be translated into protein or serves some other function. Importantly, this allows cells to express the genes they need at the correct dosage¹. The importance of gene regulation cannot be overstated, since only an estimated 2–3% of the human genome is protein coding, leaving the rest for putatively regulatory purposes².

Epigenetics studies the various mechanisms that, without altering the DNA sequence itself, have evolved to regulate access and compaction of DNA, for example, to regulate gene expression. Apart from mechanisms influencing chromatin accessibility, modifications to the DNA itself can have regulatory function. For instance, specific patterns of methyl groups added to cytosines followed by a guanine (CpG dinucleotides) in the DNA, referred to

as DNA methylation, can lead to gene repression. The main field of epigenetic research focuses on DNA methylation, modifications of histone proteins and the regulation of DNA compaction and chromatin accessibility. Once established, epigenetic modifications can be inherited in cell division but they can also be passed on to offspring via trans-generational epigenetic inheritance. Understanding epigenetic changes is central to understanding changes in cellular programs during important processes such as development and aging, but also in diseases. Similarly, epigenetic factors are responsible for repressing the expression of mobile elements of DNA and for the protection of the genomic sequence. For example, DNA methylation can be used to estimate the epigenetic age of a cell using statistical models^{3,4}. The epigenome is severely altered in tumors, which can be leveraged to predict subtypes of, for instance, breast cancer⁵, tumors of the central nervous system⁶, colon adenocarcinoma⁷, sarcoma⁸, and to identify tumors of unknown origin⁹. In several cancer types, such as Ewing sarcoma¹⁰ and glioblastoma¹¹, epigenetic changes have been identified as major drivers of the disease. Furthermore, epigenome-wide association studies (EWAS) have revealed DNA regions associated with, for instance, multiple sclerosis¹², type 1 diabetes¹³ and schizophrenia¹⁴.

Advances in array-based DNA methylation profiling and next-generation DNA sequencing technology have allowed national and international consortia such as ENCODE² and the International Human Epigenome Consortium (IHEC)¹⁵ to gather large-scale datasets to unravel the complexity of epigenetic modifications of different cell types and in the context of diseases. Here, the term epigenomics refers to genome-wide interrogation of diverse aspects of gene regulation. It comprises diverse mechanisms such as DNA methylation measured via, for instance, whole-genome/reduced-representation bisulfite sequencing (WGBS/RRBS)¹⁶, and histone modifications such as H3K9me3 or H3K27ac (ref. ¹⁷)

¹Department of Genetics/Epigenetics, Saarland University, Saarbrücken, Germany. ²Computational Biology Group, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. ³Graduate School of Computer Science, Saarland Informatics Campus, Saarbrücken, Germany. ⁴Genome Institute of Singapore, Singapore, Singapore. ⁵Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. ⁶Computational BioMedicine Lab, Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. ⁷Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany. ⁸Institute of Cardiovascular Regeneration, University Hospital and Goethe University Frankfurt, Frankfurt, Germany. ⁹These authors contributed equally: Michael Scherer, Florian Schmidt, Olga Lazareva. ✉e-mail: markus.list@wzw.tum.de

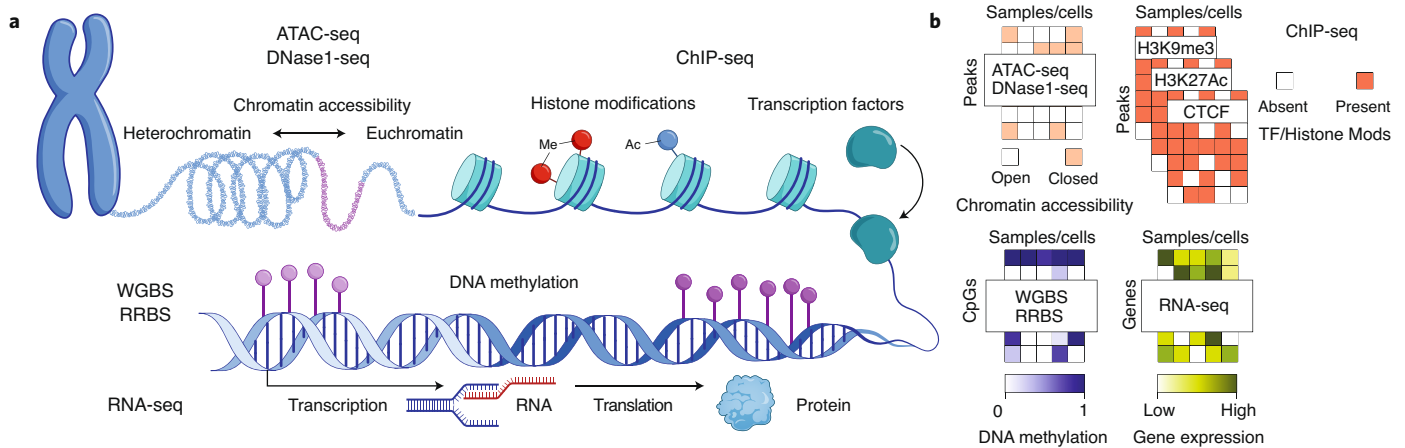


Fig. 1 | Chromatin organization and epigenomic readouts. **a**, DNA stores the genetic information of the cell. It is densely packed in a structure called heterochromatin, where DNA is tightly wrapped around complexes of histone proteins. The degree of compaction is controlled by chemical modifications to the tails of the histones, where DNA is selectively unpacked to increase chromatin accessibility and thus to allow regulatory proteins known as transcription factors to bind to the DNA. These regulate the transcription of RNA, which is then translated into a protein. The activity of DNA-binding proteins is further controlled by chemical modifications of the DNA, for example, via selective methylation of cytosines followed by guanine (methylated CpGs, indicated by full circles attached to the DNA). **b**, Many epigenetic modifications can be detected via next-generation sequencing and other high-throughput techniques. For instance, chromatin accessibility is detected by sequencing DNA fragments that are accessible for DNA-cutting enzymes (ATAC and DNase1-seq). In chromatin immunoprecipitation sequencing (ChIP-seq), antibodies are used to capture DNA-binding proteins together with their related DNA fragments, prior to sequencing. The captured short fragments are then mapped to a reference genome, where their aggregated signal forms peaks. DNA methylation of CpG dinucleotides can be detected after bisulfite treatment (WGBS, RRBS), which converts unmethylated cytosines via uracil to thymine, detectable as a deviation from the expected sequence. Finally, RNA abundances (gene expression) can be measured after converting RNA to DNA and sequencing (RNA-seq). These molecular readouts each yield a feature (position-specific peaks, CpGs or genes) by sample matrix, which is used as input for machine-learning approaches.

measured via, for instance, chromatin immunoprecipitation followed by sequencing (ChIP-seq). Chromatin accessibility can be mapped using methods such as ATAC-seq or DNaseI-seq (Fig. 1). Recently, it has also become possible to map the 3D conformation of DNA to suggest interactions between distal genomic loci^{18,19}. A number of data portals and web resources offer access to raw and processed epigenomic profiles as well as annotations of regulatory DNA regions such as promoters, enhancers and repressors (see Supplementary Table 1 for an overview).

However, due to the diversity of experimental protocols and the complexity of epigenomic modifications, computational methods for epigenomic data analysis are manifold. For an introduction to basic data formats or (pre)processing steps, such as mapping/alignment to a reference genome, peak calling, normalization or differential analysis (peaks or methylation), we refer to the literature^{20–22}. Rather than considering well established methods for annotating chromatin states via hidden Markov models²³ or for functional enrichment analysis^{24,25}, we focus here on machine-learning tools that allow for the interpretation of large-scale epigenomic data in the study of gene regulation and cellular heterogeneity (see Supplementary Table 2 for an overview).

In complex organisms, cells do not act in isolation but interact as part of a community with contributions from different cell types. However, the vast majority of epigenomics data is obtained through bulk profiling, where each sample represents an inhomogeneous mixture of cell states and cell types with unique epigenomic profiles. This introduces additional complexity in the molecular profiles that occlude condition-specific changes. In this Review, we first show how cell-type heterogeneity can be quantified using *in silico* approaches. Next, we consider the use of epigenomic data in models that predict transcriptomic (gene expression) profiles generated through RNA sequencing (RNA-seq, Fig. 1b).

To avoid the computational challenges of analyzing convoluted epigenomic profiles, cell types can be physically split using cell

sorting technologies prior to sequencing. Alternatively, single-cell sequencing technology has matured to a degree that allows studying gene regulation at the resolution of individual cells. In the third part of this Review, we highlight emerging single-cell methods that study gene regulation. Finally, we discuss expected future developments, challenges and opportunities.

Dissecting cell-type heterogeneity

In bulk samples, the data matrices obtained represent a mixture of contributions of individual cell types with potentially distinct epigenomic profiles, which poses challenges for the interpretation of epigenomic datasets. However, the presence and abundance of cell types is in itself an important feature, for instance, for understanding the response of the immune system against pathogens or cancerous cells²⁶. Computational methods, including linear regression, non-negative matrix factorization, and Bayesian approaches, can help to estimate cellular proportions from a mixture of cell types.

DNA methylation is a highly cell-type-specific epigenetic modification, and thus a premier candidate for deconvolving complex tissue samples with a heterogeneous (convoluted) mixture of signals (here, cell types in a bulk tissue) into their basic constituents. While such methods also exist for bulk RNA-seq data^{27,28}, we focus here on DNA methylation. Further work is required for deconvolution of other epigenomic layers such as histone modifications or chromatin accessibility data^{29,30}. Challenges in using epigenomic data beyond DNA methylation for deconvolution include the read-count-based structure of the data, and the dynamic range of the abundance estimates. Notably, DNA methylation heterogeneity is not only driven by (homeostatic) cellular composition, but also by other sources of heterogeneity including pathological cell infiltration (for example, immune cell infiltration into tumors), DNA methylation erosion, and allele- and strand-specific methylation (Fig. 2a).

DNA-methylation-based deconvolution tools use as input a data matrix $D_{m \times n}$ (m CpGs \times n samples, see also Fig. 1b) that is

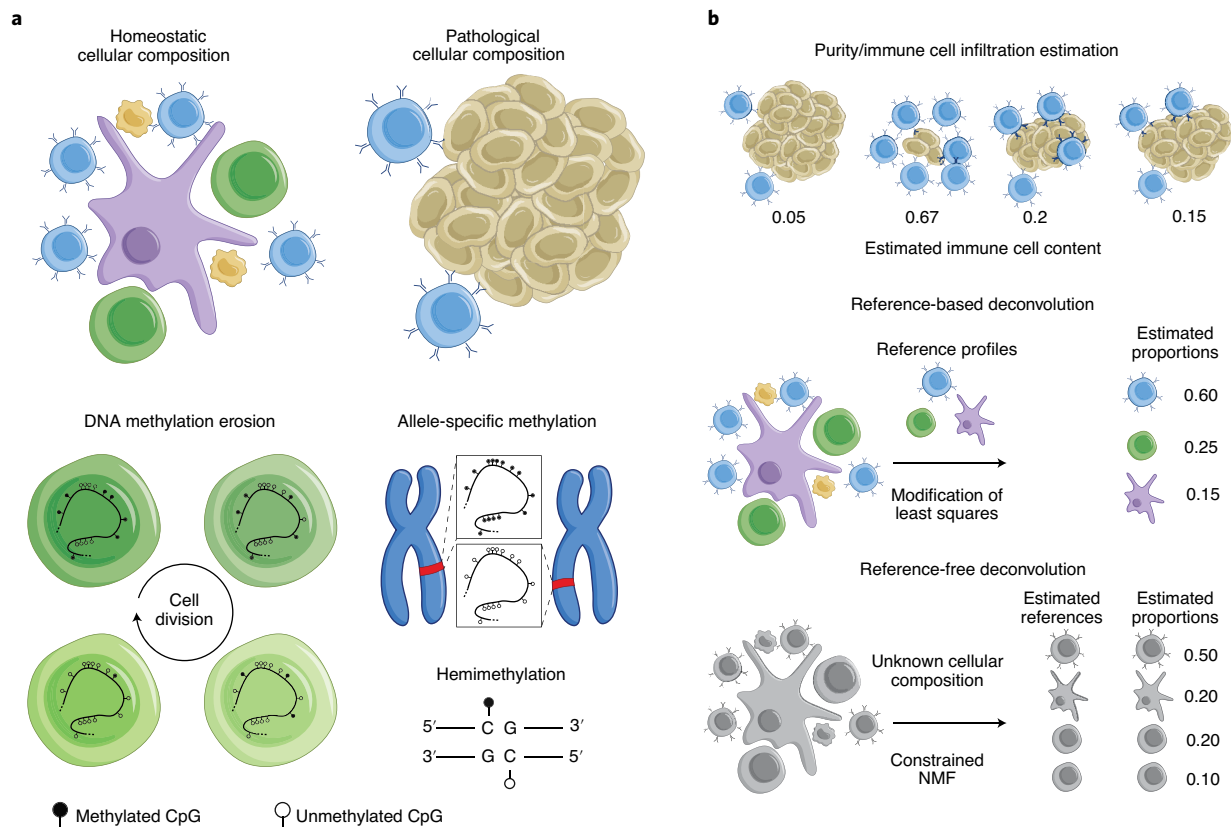


Fig. 2 | Deconvolution of complex DNA methylation data. **a**, Five potential sources of variation in DNA methylation data exist. Homeostatic cellular composition refers to a mixture of cell types (cell-type heterogeneity) with potentially distinct epigenetic patterns in a bulk sample. Pathological cellular composition is the cellular heterogeneity that is caused by the state of a biological system, which includes infiltration of immune cells into tumors. DNA methylation erosion refers to the stochastic loss of DNA methylation during cell division. Allele-specific methylation is the methylation of only one of the two alleles in a diploid organism, and hemimethylation refers to the methylation of only one of the strands of the DNA. **b**, Three classes of deconvolution tools exist: immune-cell infiltration or sample-purity estimation methods return a prediction of the overall immune cell content of a sample. Reference-based deconvolution requires reference profiles for the estimation of cellular proportions and reference-free deconvolution methods do not require such references.

decomposed into the major sources of variation in the data and into their proportions across the samples. The computational approaches can be broadly divided into three categories: (i) purity/infiltration estimation methods (for example, the LUMP estimate³¹), (ii) reference-based, and (iii) reference-free deconvolution tools (Fig. 2b). Tumor purity can be estimated via simple score-based methods that consider CpGs that are specifically unmethylated in immune cells in the tumor sample. We will hence focus on the other two categories that require more complex methods.

Reference-based deconvolution. Rather than adjusting for cellular composition, which is common in EWAS^{32,33}, reference-based deconvolution methods use modifications of linear least squares to estimate the proportions of provided DNA methylation profiles of reference cell types in a given dataset. Given the input DNA methylation matrix and a matrix of reference profiles, the objective of reference-based methods is to determine the proportions of the reference profiles in the observed matrix $D_{m \times n}$. Reference-based methods use constrained projection³⁴, robust partial correlations^{35,36}, least trimmed squares regression³⁷, or other statistical learning methods, such as support vector regression³⁸, to solve this problem. A comparison study across these methods³⁹ found that the constrained projection method implemented in the minfi R-package⁴⁰ had superior performance. From our experience, reference-based methods perform best when high-quality DNA methylation data

exist for purified cell types and when the constituting cell types of a bulk sample are known. This is the case for whole blood samples, where cell-type-specific reference profiles exist⁴¹. In tumor samples, reference-free methods have been used to dissect the cellular composition⁴².

Reference-free deconvolution. By contrast, reference-free deconvolution methods predict both the reference profiles and the proportions of these estimated reference profiles across the samples using unsupervised statistical learning. They are hence ideal to account for cell types where a suitable reference is not available or for samples where the contributing cell types are not known. A common computational approach for this problem is non-negative matrix factorization (NMF), which dissects the input DNA methylation data matrix ($D_{m \times n}$) into the estimated reference profiles ($T_{m \times k}$) and the proportions matrix ($A_{k \times n}$). The profiles $T_{m \times k}$ are often referred to as latent methylation components (LMCs). k is the number of latent components to be determined, m the number of CpGs and n the number of samples. This approach for the analysis of DNA methylation data is implemented in various computational tools^{43–45} (Supplementary Table 2), which formulate a constrained version of NMF. Compared with reference-based deconvolution methods, biological interpretation of reference-free deconvolution results is more challenging and a dedicated preprocessing of DNA methylation data is required. The preprocessing includes

accounting for potential confounding factors, data normalization, and quality filtering⁴². We point out that data processing is crucial for both reference-free and reference-based methods, and that the processing steps including filtering, normalization, and accounting for confounding factors need to be considered. We refer to related publications for details^{42,46,47}.

Hybrid deconvolution methods. There are also computational approaches that borrow concepts both from reference-based and reference-free methods. For instance, semi-reference-free methods⁴⁸ use a Bayesian prior to improve NMF results. Additional frameworks combine differential DNA methylation analysis with deconvolution using standard NMF⁴⁹. Furthermore, the paradigm of a single latent methylation components matrix $T_{m \times k}$ can be replaced by sample-specific components (that is, $T_{m \times k}^1, \dots, T_{m \times k}^n$; one T for each sample), which can be induced by a cell type changing its biological identity in response to a disease. To solve this problem, tensor composition analysis can be used to obtain sample-specific reference profiles⁵⁰. Since confounding factors can strongly influence the DNA methylation landscape, canonical correlation analysis employs multiple DNA methylation data matrices as input to discern technical from biological sources of variation associated with cell-type heterogeneity⁵¹.

Leveraging bisulfite read heterogeneity. Most of the aforementioned methods use as input the DNA methylation data obtained from microarrays, which use color intensities as a readout and return an aggregate methylation state over various cellular states. By contrast, bisulfite sequencing provides sequence information for single molecules encoded in the sequencing reads. Heterogeneity in sequencing reads observed in a biological sample (within-sample heterogeneity) can be leveraged for deconvolution of the cellular composition of the sample. Genome-wide estimates of within-sample heterogeneity can be used to reliably estimate tumor purity or for segmenting the genome into highly and lowly variably methylated regions⁵². Additionally, read-level information can be used for *DBSCAN* clustering according to the cell-of-origin and used for computational deconvolution⁵³.

Understanding gene expression using epigenomic data

Gene expression is a result of a complex machinery that involves chromatin remodeling complexes for opening the chromatin, followed by the recruitment of a transcriptional machinery consisting of proteins referred to as transcription factors (TFs), which act in a cell-type and condition-specific manner. Likewise, histone modifications, DNA methylation and other epigenetic modifications occur in a cell-type and condition-specific context, known as epigenomic signatures⁵⁴. Associating TF binding information and epigenomic data with gene expression readouts thus allows the regulatory role of both epigenetic modifications and of DNA-binding proteins to be elucidated. Furthermore, such models allow researchers to interpret genetic variants in non-coding parts of the genome identified, for instance, with genome-wide association studies. In predictive models, such genetic data is an important feature for improving our understanding of diseases⁵⁵.

A prerequisite for almost all of the gene expression prediction approaches discussed in this section is the association of epigenomic signals to genes to generate a feature matrix X , which is typically used for constructing (regularized) linear or logistic regression models. An important factor to consider is the linking strategy, that is, how to select putative target genes for a regulatory region identified via the epigenomic signal, which is essential for feature generation.

Linking regulatory elements to their putative target genes. Several strategies have been proposed to link regulatory elements

in the DNA to their putative target genes (Fig. 3a). These methods either consider peaks of histone or chromatin accessibility data, or sites of hyper/hypo (that is, significantly increased or decreased) methylation as input. In window-based approaches, all candidate regulatory elements that lie within a predefined area around a gene are considered as regulatory elements⁵⁶. In nearest-gene approaches, each potential regulatory element is assigned to its closest gene in genomic distance⁵⁷. If sufficient data are available, correlation-based approaches can be used to assign candidate regulatory elements to their most likely target genes.

All of these strategies can be complemented using experimentally determined chromatin contacts. These inform about the presence of long-range regulatory interactions^{58,59}. The same kind of data can be used to determine topologically associated domains, which define the genomic borders of regulatory interactions⁶⁰. Finally, literature curated databases can be used to obtain previously reported and often experimentally validated regulatory elements per gene⁶¹.

Insights on tissue- and cell-type-specific regulators. Early studies have demonstrated that gene expression can be accurately predicted in silico from epigenomic data. Ouyang et al. demonstrated using principal component regression that TF ChIP-seq is able to explain 65% of the variation in gene expression in mouse embryonic stem cells⁵⁶. Subsequently, it was shown that chromatin accessibility data combined with predicted TF binding sites can be as good as or even more accurate than models relying on TF ChIP-seq data when it comes to predicting gene expression^{62,63}. Indeed, in a study of CD4⁺ T-cell differentiation, Costa et al. illustrated that regression coefficients inferred by a linear model using predicted TF binding sites allow for a biologically insightful interpretation⁶⁴.

In light of those promising results, scalable software solutions that could integrate and interpret the increasing amounts of epigenomic data produced, for instance within ENCODE² or IHEC¹⁵, have been developed. They are able to integrate epigenomic datasets on the cell-type/tissue level, making predictions across genes and thereby inferring general cell-type/tissue-specific regulatory information (Fig. 3b).

Mathematically, these algorithms build a feature matrix X composed of m features derived from epigenetic data, assessed for n different genes within one sample. To identify biological relevant signals, these feature matrices are usually used in linear models that exploit various regularization methods to fit a sparse vector of regression coefficients β predicting gene expression y . The vector of regression coefficients β can then be used for biological interpretation.

Some of those methods have been developed in the context of cancer biology, for instance RACER⁶⁵ and RABIT⁶⁶. Both methods are designed to address apparent confounders such as copy number variations, which frequently occur in cancer. RACER builds a linear regression model using lasso regularization of TF data, copy number variations, DNA methylation and microRNA (miRNA) expression signals as features to predict gene expression. RACER thus leverages sample-specific regulatory activities of TFs together with information about post-transcriptional regulation with miRNAs to generate gene-specific TF and miRNA interaction scores. RABIT follows a similar goal as RACER but uses the Frisch–Waugh–Lovell method of linear regression to generate a candidate set of TFs, while controlling for confounders such as copy number variations.

While RACER and RABIT are especially well suited to analyze cancer datasets, their dependence on TF ChIP-seq data limits their general applicability. The TEPIC framework uses putative TF binding sites predicted in accessible genomic loci to estimate gene expression and to infer tissue-specific regulators⁶⁷. TEPIC uses linear regression with elastic net regularization to provide a sparse

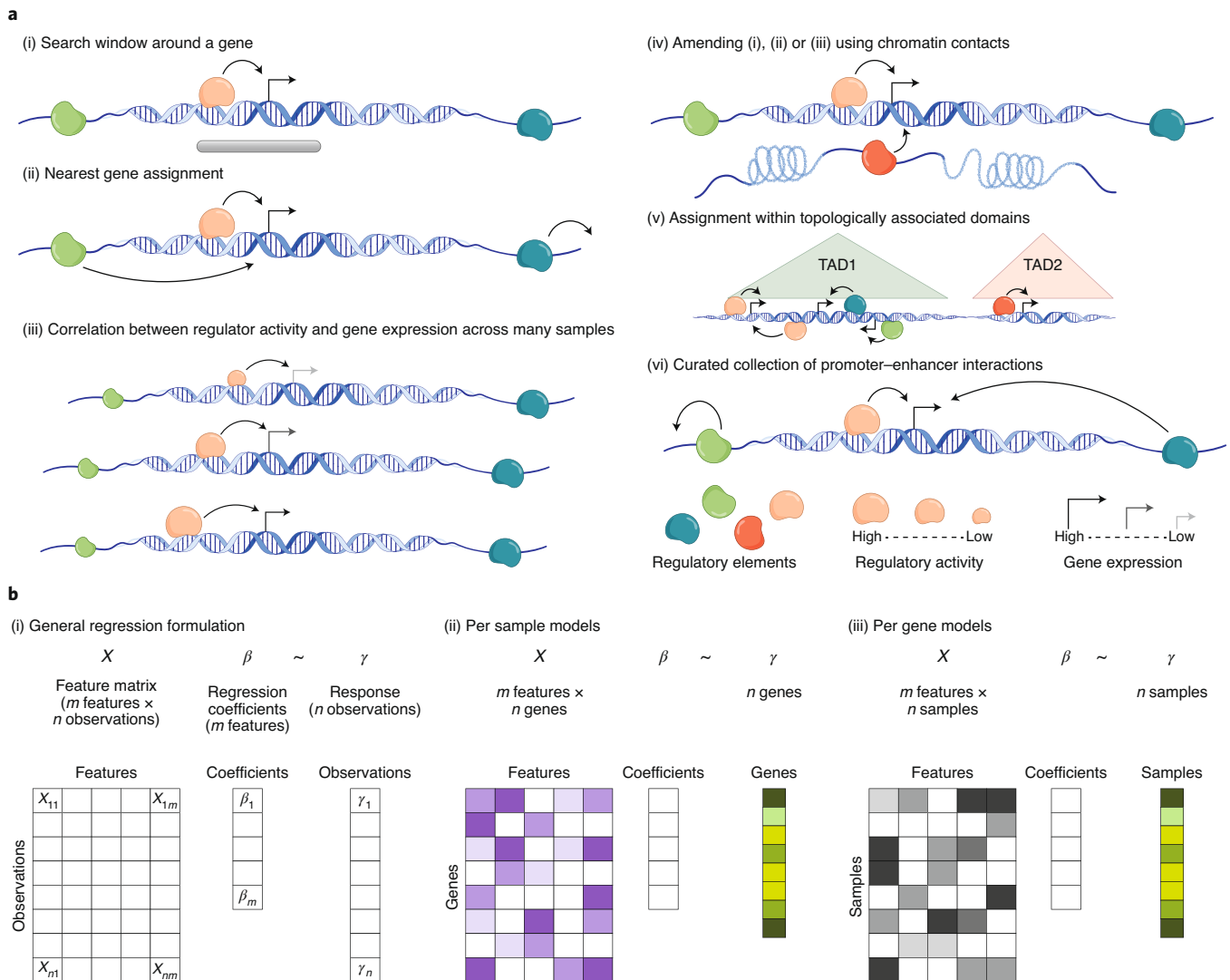


Fig. 3 | Feature generation and modeling options for gene expression prediction using epigenomics data. **a**, The many strategies to link regulatory elements (REMs) to putative target genes: (i) assignment using a search window assigned to a gene; (ii) assignment using genomic distance of a REM to a gene; (iii) correlation testing-based assignment between activity of regulators and gene expression; (iv) extension of methods (i–iii) using chromatin contacts as derived from Hi-C data; (v) restriction of (i–iv) to interactions occurring within topologically associated domains (TADs) and (vi) literature-based promoter–enhancer interactions. **b**, Set-ups of machine-learning models: (i) general set-up of a regression problem; (ii) per-sample model predicting gene expression across many genes within one sample; (iii) per-gene set-up predicting gene expression for a distinct gene across many samples.

solution that retains correlated features, which is important to model coregulation events. The latest version of TEPIC also supports the incorporation of chromatin contacts, defined for instance using Hi-C data. Although the addition of Hi-C data could improve model performance, it also complicates the optimization problem even further by increasing dimensionality of the feature matrix⁵⁸. Hi-C data are also used in TargetFinder⁵⁹ to define promoter–enhancer pairs across various cell lines. TargetFinder uses an ensemble of boosted decision trees to reconstruct the three-dimensional regulatory landscape, spanned by multiple epigenomic signatures to uncover interactions among TFs and other epigenetic modifications.

In addition to those aforementioned approaches, tools such as Efilter⁶⁸, which solves a linear regression problem considering genome-wide epigenetic data using the EDC2 criterion, and DeepChrome⁶⁹, a deep convolutional neural network approach that considers histone modification ChIP-seq data as input, further improved the accuracy of gene expression prediction within a sample. Unfortunately, these tools do not provide interpretable models.

Gene-specific regulatory information. The methods described thus far are trained across all genes within one sample to deduce general regulatory effects. However, with constantly increasing sample numbers in epigenomic data portals^{70,71}, per-gene models have become feasible to infer gene-specific regulatory information across cell-types or tissues (Fig. 3b). The inferred sets of regulatory gene interactions pose a valuable resource complementing experimentally determined associations as included, for instance, in the GeneCards database⁶¹. Similar to per-sample models, (regularized) linear and logistic regression models are employed by most methods. However, in per-gene models the feature matrix X is composed of m measurements of epigenetic data for one particular gene assessed for n different patients.

One of the first methods, JEME⁷², extracts all enhancers within 1 MB around a gene’s transcription start site (TSS) and generates a feature matrix using a measure of enhancer activity, for example, chromatin accessibility data. Next, a lasso linear regression model across many samples returns regression coefficients for each TSS

and enhancer. Sample-specific promoter-enhancer pairs are determined in a second step using random forests. Hence, the final output of JEME is cell-type-specific enhancer-target networks, which are available as a resource.

The FOCS approach⁷³ also aims at identifying promoter-enhancer pairs, but it omits the generation of sample-specific promoter-enhancer pairs as performed in JEME. Instead, a feature matrix consisting of the top ten closest candidate regulatory elements is constructed for each gene and used to predict gene-expression/promoter activity. The regulator selection is refined in an elastic net regression model.

Unlike JEME or FOCS, which rely on a universe of predefined regulatory elements (Fig. 3a), the STITCHIT approach⁷⁴ detects novel regulatory elements and links them to their target genes in a single optimization step that uses the minimum description length principle to find the optimal set of regulatory elements that explain gene-expression variance of an individual target gene. STITCHIT's regulator-gene associations are available via the EpiRegio web server⁷⁵.

The plethora of available methods (Supplementary Table 2) provides researchers with versatile options to integrate various epigenomic datasets in gene-expression prediction models to infer transcriptional regulators and regulatory elements at an unprecedented level of accuracy. With ongoing improvements in experimental protocols at single-cell resolution, even more fine-grained predictions can be expected in the future. In the next section, we discuss several of the already available methods that attempt to decipher regulatory landscapes on a single-cell level.

Gene regulation on single-cell level

In classical bulk sequencing, individual cell types have to be isolated a priori or deconvolution methods need to be applied to disentangle possible contributions of different cell types. Both approaches are not ideally suited for studying the heterogeneity of cellular communities and the complexity of cellular differentiation. Methods for studying the epigenetic state of individual cells are imperative for understanding community effects, for example, in the tumor microenvironment where the epigenetic state of immune cells plays a decisive role in immunotherapy treatment response. Single-cell epigenome and transcriptome sequencing technology have successfully addressed these issues and have already revolutionized our understanding of cellular differentiation and gene regulation. Existing computational methods for single cell epigenomics focus mostly on single-cell ATAC-seq (scATAC-seq) profiles of chromatin accessibility alone or in combination with scRNA-seq gene expression profiles. By contrast, single-cell epigenomic profiling of DNA methylation and ChIP-seq is still faced with technological challenges, but will likely become available for integration in the near future. Common applications are unsupervised machine learning methods for clustering, trajectory inference and gene regulatory network inference (Fig. 4).

Clustering and dimensionality reduction. Due to its low coverage, current single-cell sequencing protocols, in particular scATAC-seq, generate extremely sparse and noisy datasets. To account for this, a common analysis step following preprocessing is dimensionality reduction and/or clustering of the data. This allows for extracting meaningful mechanisms that can distinguish between cell types and differentiation stages across conditions or treatments. Moreover, this helps to account for confounders such as batch effects, experimental artefacts or the cell cycle⁷⁶. Unsupervised methods such as weighted k -medoids⁷⁷, matrix factorization⁷⁸, topic modeling⁷⁹, deep generative neural networks⁸⁰, weighted principal component analysis⁸¹ and k -nearest neighbor⁸² are often used for clustering cells with similar epigenetic patterns and for identifying key features specific to each cluster.

Epigenomic data integration. Integrating scATAC and scRNA-seq data further offers the opportunity to study gene regulation on a cell-type-specific level. As assays that capture both readouts from the same cells have only recently become available, several methods have been proposed to link these data types or to project them into a shared embedding space. An example of a linkage approach is SOMatic⁸³, which constructs independent low-dimensional representations using self-organizing maps for RNA-seq experiments and for ATAC-seq peaks and then links them using information about nearest genes. Another linkage approach is Conos⁸⁴, which applies nearest-neighbor or mutual nearest-neighbor mapping to different samples and different data modalities to retrieve conjoint clustering of cells.

Linkage of independent low-dimensional representations, though flexible, might be problematic due to the fact that scATAC-seq and scRNA-seq data may capture different aspects of biology and do thus not necessarily agree with respect to individual cell types⁸⁵. An alternative approach is to perform clustering of the cells such that scRNA-seq embedding benefits from scATAC-seq data, and vice versa. Many of the methods in this category are based on NMF. For instance, Duren et al.⁸⁶ proposed a method that utilizes a coupled clustering approach that systematically maps scATAC-seq peaks to genes for downstream analysis such as inferring gene regulatory networks at the single-cell level.

MATCHER⁸⁷ uses NMF to determine shared and dataset-specific metagenes across datasets and then uses the resulting factor space for a conjoint clustering of all cells. scAI⁸⁸ learns three sets of low-dimensional representations of high-dimensional data: the gene, locus, and cell-loading matrices describing the relative contributions of genes, loci, and cells in the inferred factors as well as the cell-cell similarity matrix used for aggregating sparse epigenomic data. MOFA+⁸⁹ is a generally applicable tool for integrating multi-omics data, which infers K latent factors with associated feature weight matrices (per data modality) that explain the major axes of variation across the datasets. Similarly, LIGER⁹⁰ uses integrative NMF⁹¹ to find shared and dataset-specific cell identities that can be further used as co-embedding.

Methods not based on NMF are MAESTRO⁹², UnionCom⁹³ and SCIM⁹⁴. MAESTRO computes a common embedding from two independent embeddings based on canonical correlation analysis and then maps the cells from two representations based on mutual nearest neighbors. UnionCom and SCIM are more generic approaches that allow integrating any single cell multi-omics data: they produce a single co-embedding space from independent samples and technologies. UnionCom is based on an adjusted version of generalized unsupervised manifold alignment and SCIM employs autoencoders for data compression and then bipartite graph matching for aggregation.

Trajectory inference. Studying cell differentiation is another challenging task in single-cell data analysis. Here, we leverage the fact that cells at all possible differentiation stages exist in parallel. This allows us to map a trajectory from undifferentiated stem cells to fully differentiated cells along with all intermediate states and possible branch points or metastable states where cell fate is decided⁷⁶. To quantify the degree of differentiation, the concept of a pseudotime is commonly used. STREAM⁹⁵ reconstructs complex trajectories along with pseudotime estimation from both single-cell transcriptomic and chromatin-accessibility data. There are also approaches such as APEC⁸² that cluster cells based on scATAC-seq, and integrate Monocle^{96,97} to reconstruct trajectories.

Regulatory network inference. As shown in Fig. 3a, the linkage between distal regulatory sequences and their target genes is not straightforward and remains an open issue also in single-cell epigenomics. Cicero⁹⁸ constructs putative *cis*-regulatory maps from

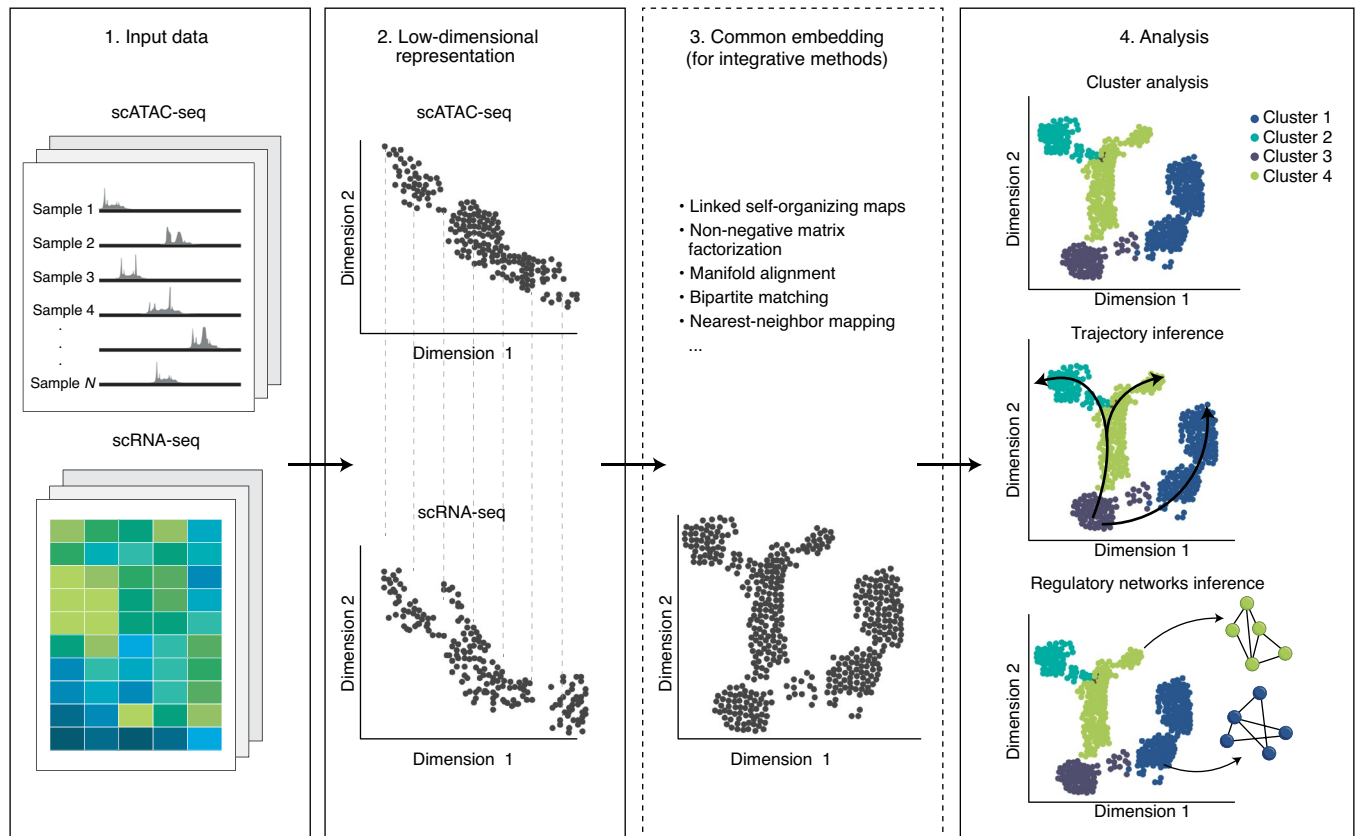


Fig. 4 | Workflow of single-cell epigenomics methods. Most approaches use scATAC-seq and/or scRNA-seq data as input for machine-learning methods to learn a low-dimensional representation and a common embedding space (for integrative methods). The analysis step can be grouped into three categories: (1) individual or integrative (scATAC-seq and scRNA-seq) clustering analysis; (2) trajectory inference (sc-qPCR, scRNA-seq or scATAC-seq data); (3) regulatory network inference, that is regulatory interactions between genes.

single-cell chromatin accessibility data and thus contributes to the understanding of eukaryotic gene regulation. In addition, Cicero can be used to identify target genes affected by genetic variants in regulatory regions that do not code for proteins and are often neglected in genome-wide association studies.

The large number of available methods (Supplementary Table 2) shows that the field of single-cell epigenomics evolves rapidly, leading to a more refined understanding of how epigenetic processes interact with one another to control gene expression. Ultimately, this has the potential to transform our knowledge about how the phenotype of the cell is maintained and how it is perturbed in disease.

Conclusions and future perspectives

The vast majority of the methods presented here utilize well-established, unsupervised machine-learning algorithms such as NMF, self-organizing maps, *k*-medoids or KNN. For supervised machine learning, regularized linear and logistic regression are a common choice. While they employ similar strategies, computational epigenomics methods differ largely in their unique combination of (1) the type of molecular data they leverage (for example, histone ChIP-seq, chromatin accessibility or DNA methylation, or combinations thereof) and how they have been processed (for example, ChIP-seq peak calling versus signal-based approaches); (2) the prior knowledge they consider (for example, TF-binding sites, known regulatory regions from public databases); and, for gene-based methods, (3) the way putative regulatory regions are linked to genes (for example, nearest-gene versus window-based approaches). New sequencing-based molecular profiling techniques are currently developed at a breathtaking rate and can be used to

improve computational inference. For instance, recent methods have begun to incorporate high-resolution 3D conformation data (for example, Hi-C) to refine the association of distal regulatory regions to their target genes.

While we have seen the first successful attempts at leveraging deep learning for analyzing these complex datasets (for example, DeepChrome³⁹), artificial neural networks and autoencoders do not currently show a dramatically improved performance compared to classical methods, and, at the same time, exhibit more limited interpretability. However, we expect that the exponential growth in the available epigenomic data and the constant technological improvement will work in favor of more advanced machine-learning methods, such as deep learning, and eventually give them an edge over classical methods. As gene regulation takes place on multiple epigenetic levels, we further expect that advances in multi-omics integration will lead to an improvement in method performance, in particular once multi-omics profiling techniques (for example, profiling of the methylome and transcriptome³⁹) become more widely used. Moreover, the field still lacks a robust technology for ChIP-seq to interrogate histone modifications and TF-binding on a single-cell level, currently leaving us ignorant to an important aspect of epigenetic regulation. Cell-type deconvolution analysis is currently based mostly on bulk samples and does not yet profit from single-cell-based signatures to the same extent as the field of transcriptomics, although Teschendorff et al. have recently proposed a promising approach to close this gap by inferring methylome profiles from single-cell transcriptomics data³⁶.

Beyond the consideration of the different levels of epigenome profiles, more ambitious integration efforts will be needed to

understand how gene regulation affects the phenotype and behavior of a cell. For instance, the interplay of epigenetic, transcriptional and post-transcriptional (for example, RNA epigenomics) gene regulation needs to be considered just as much as the effect of post-translational modifications on the level of proteins and interactions with the environment, including the microbiome¹⁰⁰. All of these factors require the development of new powerful data integration methods and multi-omics machine-learning strategies such as multi-view learning¹⁰¹.

Received: 15 October 2020; Accepted: 8 February 2021;
Published online: 15 March 2021

References

- Alberts, B. et al. *Molecular Biology of the Cell* 4th edn (Garland, 2002).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 3156 (2013).
- Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
- Stefansson, O. A. et al. A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Mol. Oncol.* **9**, 555–568 (2015).
- Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
- Yang, C., Zhang, Y., Xu, X. & Li, W. Molecular subtypes based on DNA methylation predict prognosis in colon adenocarcinoma patients. *Aging* **11**, 11880–11892 (2019).
- Koelsche, C. et al. Sarcoma classification by DNA methylation profiling. *Nat. Commun.* **12**, 498 (2021).
- Moran, S. et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol.* **17**, 1386–1395 (2016).
- Sheffield, N. C. et al. DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. *Nat. Med.* **23**, 386–395 (2017).
- Klughammer J. et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* **24**, 1611–1624 (2018).
- Huynh, J. L. et al. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. *Nat. Neurosci.* **17**, 121–130 (2014).
- Rakyan V. K. et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet.* **7**, e1002300 (2011).
- Pidsley, R. et al. Methyloic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. *Genome Biol.* **15**, 483 (2014).
- Stunnenberg, H. G. International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
- Harris, R. A. et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010).
- Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
- Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).
- Cazaly, E. et al. Making sense of the epigenome using data integration approaches. *Front. Pharmacol.* **10**, 126 (2019).
- Yong, W.-S., Hsu, F.-M. & Chen, P.-Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **9**, 26 (2016).
- Nakato, R. & Sakata, T. Methods for ChIP-seq analysis: a practical workflow and advanced applications. *Methods* <https://doi.org/10.1016/j.ymeth.2020.03.005> (2020).
- Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
- Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).
- McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- Finotello F. & Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).
- Sturm, G. et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).
- Sompairac N. et al. Independent component analysis for unraveling the complexity of cancer omics datasets. *Int. J. Mol. Sci.* **20**, 4414 (2019).
- Li, H. et al. DeconPeaker, a deconvolution model to identify cell types based on chromatin accessibility in ATAC-Seq data of mixture samples. *Front. Genet.* **11**, 392 (2020).
- Hübschmann D. et al. Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. Preprint at *bioRxiv* <https://doi.org/10.1101/199547> (2017).
- Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
- Rahmani, E. et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 443–445 (2016).
- Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods* **11**, 309–311 (2014).
- Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinf.* **13**, 86 (2012).
- Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinf.* **18**, 105 (2017).
- Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCOPE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).
- Arneson, D., Yang, X. & Wang, K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Commun. Biol.* **3**, 422 (2020).
- Chakravarthy, A. et al. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.* **9**, 3220 (2018).
- Kaushal, A. et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinf.* **18**, 216 (2017).
- Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
- Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* **7**, e41361 (2012).
- Scherer, M. et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecomPipeline, MeDeCom and FactorViz. *Nat. Protoc.* **15**, 3240–3263 (2020).
- Houseman E. A. et al. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinf.* **17**, 259 (2016).
- Onuchic, V. et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell Rep.* **17**, 2075–2086 (2016).
- Lutsik, P. et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 55 (2017).
- Sun, Z., Cunningham, J., Slager, S. & Kocher, J.-P. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* **7**, 813–828 (2015).
- Fortin, J.-P., Triche, T. J. Jr & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).
- Rahmani, E. et al. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome Biol.* **19**, 141 (2018).
- Li, Z. & Wu, H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.* **20**, 190 (2019).
- Rahmani, E. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat. Commun.* **10**, 1673 (2019).
- Thompson, M., Chen, Z. J., Rahmani, E. & Halperin, E. CONFINED: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets. *Genome Biol.* **20**, 138 (2019).
- Scherer M. et al. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.* **48**, e46 (2020).
- Scott, C. A. et al. Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.* **21**, 156 (2020).
- Vaquerez, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
- Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
- Ouyang, Z., Zhou, Q. & Wong, W. H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **106**, 21521–21526 (2009).
- González, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.* **47**, 1249–1259 (2015).

58. Schmidt, F., Kern, F. & Schulz, M. H. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenet. Chromatin*. **13**, 4 (2020).
59. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
60. Okonechnikov, K., Erkek, S., Korbel, J. O., Pfister, S. M. & Chavez, L. INTAD: chromosome conformation guided analysis of enhancer target genes. *BMC Bioinf.* **20**, 60 (2019).
61. Stelzer, G. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**, 1.30.1–1.30.33 (2016).
62. McLeay, R. C., Lesluyes, T., Cuellar Partida, G. & Bailey, T. L. Genome-wide in silico prediction of gene expression. *Bioinformatics* **28**, 2789–2796 (2012).
63. Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* **22**, 1711–1722 (2012).
64. Costa, I. G., Roider, H. G., do Rego, T. G., de Carvalho, F. & de, A. T. Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinf.* **12**, S29 (2011).
65. Li, Y., Liang, M. & Zhang, Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput. Biol.* **10**, e1003908 (2014).
66. Jiang, P., Freedman, M. L., Liu, J. S. & Liu, X. S. Inference of transcriptional regulation in cancers. *Proc. Natl Acad. Sci. USA* **112**, 7731–7736 (2015).
67. Schmidt, F. et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* **45**, 54–66 (2017).
68. Kumar, V. et al. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* **31**, 615–622 (2013).
69. Singh, R., Lanchantin, J., Robins, G. & Qi, Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639–i648 (2016).
70. Davis, C. A. et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
71. Bujold, D. et al. The International Human Epigenome Consortium Data Portal. *Cell Syst.* **3**, 496–499.e2 (2016).
72. Cao, Q. et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
73. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biol.* **19**, 56 (2018).
74. Schmidt F. et al. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. Preprint at *bioRxiv* <https://doi.org/10.1101/585125> (2019).
75. Baumgarten, N. et al. EpiRegio: analysis and retrieval of regulatory elements linked to genes. *Nucleic Acids Res.* **48**, W193–W199 (2020).
76. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
77. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9**, 2410 (2018).
78. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinf.* **19**, 253 (2018).
79. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
80. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
81. Urrutia, E., Chen, L., Zhou, H. & Jiang, Y. Destin: toolkit for single-cell analysis of chromatin accessibility. *Bioinformatics* **35**, 3818–3820 (2019).
82. Li, B. et al. APEC: an accession-based method for single-cell chromatin accessibility analysis. *Genome Biol.* **21**, 116 (2020).
83. Jansen, C. et al. Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS Comput. Biol.* **15**, e1006555 (2019).
84. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
85. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
86. Duren, Z. et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl Acad. Sci. USA* **115**, 7723–7728 (2018).
87. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138 (2017).
88. Jin, S., Zhang, L. & Nie, Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 25 (2020).
89. Argelaguet R. et al. MOFA: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
90. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).
91. Yang Z., Li S., Zha X., Sun J. & Wang Y. A source-type harmonic energy unbalance suppression method based on carrier frequency optimization for cascaded multilevel APF. In *2016 IEEE Energy Conversion Congress and Exposition (ECCE)* (2016).
92. Wang, C. et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 198 (2020).
93. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).
94. Stark S. G. et al. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).
95. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).
96. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
97. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
98. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
99. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
100. Miro-Blanch, J. & Yanes, O. Epigenetic regulation at the interplay between gut microbiota and host metabolism. *Front Genet.* **10**, 638 (2019).
101. Nguyen, N. D. & Wang, D. Multiview learning for understanding functional multiomics. *PLoS Comput. Biol.* **16**, e1007677 (2020).

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the eMed research and funding concept (grant Sys_CARE [01ZX1908A]) to M.L. and J.B.; VILLUM Young Investigator Grant 13154 and European Union's Horizon 2020 project RepoTrial 777111 to J.B.; the Bavarian State Ministry of Science and the Arts as part of the Bavarian Research Institute for Digital Transformation (bidit) to O.L.; BMBF project de.NBI-epi (031L0101D) to M.S. and J.W.; DZHK (German Centre for Cardiovascular Research, 81Z0200101), the Cardio-Pulmonary Institute (CPI) EXC 2026 to M.H.S. and the Agency for Science, Technology and Research, Singapore (Enabling Data Analytics Technologies for Next-Generation Pathology from 3D Transcriptomics – SERC Data Analytics, 1727600056) to F.S.

Author contributions

M.L. conceived this Review, supervised and contributed to the writing of the manuscript. M.S., F.S. and O.L. collected information about the tools and databases and wrote the corresponding parts of the manuscript. M.H.S., J.B. and J.W. provided critical feedback on the manuscript. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43588-021-00038-7>.

Correspondence should be addressed to M.L.

Peer review information *Nature Computational Science* thanks Lukas Chavez and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Fernando Chirigati was the primary editor on this Review and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

5. General Discussion and Outlook

The availability of massive amounts of biological data came with immense promise, and it is undoubtedly valuable for the improvement of our understanding of complex biological systems. However, unlike in many other fields where big data might be more easily interpretable, biological data represents a snippet of a system in a very particular moment and under certain assumptions. The data comes with biases, noise and it is, indeed, very big.

Despite these challenges, the field keeps progressing and slowly moving away from outdated disease definitions derived before rich molecular data availability. The International Classification of Diseases (ICD) [243] definitions are often based on the symptoms and simple biomarkers and therefore do not reflect mechanisms of diseases that might be patient-specific. For instance, conditions such as primary hypertension are usually treated with blood vessel-dilating drugs, allowing to normalize the elevated blood pressure. However, the real cause for high blood pressure remains unknown [5]. For some conditions, such as rheumatoid arthritis, we know that little, that even symptomatic treatment is often unsuccessful, leaving patients to deal with pain, destruction of joints, and resulting disability [244, 245]. Understanding of molecular mechanisms of diseases could allow personalized treatment, early diagnostics, or even prevention.

The purpose of this thesis is to study application of complex machine learning methods to diverse molecular data in order to understand the contribution of different data types to meaningful disease modules discovery. In particular, I focus on combining PPI networks with transcriptomics data [9, 10], but also epigenetic mechanisms are reviewed [11]. Critical assessment is performed with regards to limitations on the side of the data (using AMI test-suite [10]), on the side of the machine learning methods [11], and the produced results are systemized.

5.1. PPI networks as a prior knowledge source

PPI networks analysis has been extensively used for disease mechanisms identification in the last 10 years, but there is still no one right way to learn from PPI networks. As shown in the first publication (BiCoN), some data types (such as transcriptomics and PPI networks) can provide excellent results and be perfectly sound in theory, yet in practice, the results are not as biologically meaningful as expected due to protein's degree bias, algorithmic limitations and evaluation difficulties. This conclusion only became possible after exhaustive computational assessment (over 10 000 algorithm executions) to eliminate different explanations of poor performance such as bad quality of a particular network, extreme complexity of a particular condition, or inappropriate algorithmic approaches. We also assessed the performance

from the standpoint of the functional relevance of the retrieved gene sets and phenotype prediction accuracy. Both metrics confirmed that the ability of algorithmic frameworks to find meaningful subnetworks aided by transcriptomics data should be carefully reviewed.

While it is easy to suggest that we *just* need *better* PPI networks that are not affected by aforementioned biases, the issue can be also addressed from other perspectives. The correlation between a number of studies conducted on a protein and its degree is only one of the reasons that make PPI networks challenging for computational analysis. However, studying the most promising topics (proteins, in this case) is the fundamental nature of scientific cognition, and we can not expect scientists to pay equal attention to all proteins because scientific resources are limited. Instead, three possible solutions can break the acknowledged bias:

- **Advancement of PPI measurement.** Luck et al. recently proposed a novel "all-by-all" reference interactome map of human binary protein interactions (HuRi) [246]. To provide a uniform genome coverage, Luck et al. screened human proteome nine times with a panel of three yeast two-hybrid assays. This experimental protocol allows covering 90% of the protein-coding genome. Due to the "all-by-all" strategy, HuRi covers the proteome more uniformly and does not bias highly expressed proteins.
- **Advancement of computational methods to improve quality of PPI networks:** AlphaFold 2 has successfully demonstrated that modern computational algorithms have the capacity of solving large scale biological problems [47]. Identification of protein structures was mainly done experimentally, and therefore, many proteins did not have fully identified structures. AlphaFold 2 makes it possible to estimate all protein structures from a sequence. Moreover, AlphaFold 2 model allows to identify residue conservation and co-evolution directly from the sequence alignment which is useful for protein interaction prediction [247]. This research is still in its early stage, but it gives a hope that soon computational methods might be able to fill the knowledge gap in PPI networks, reducing human and technical bias.
- **Multi-omics approaches to diseases:** Independent omics layers should be used for confirmation of findings. As most diseases "leave a trace" on various molecular levels, the derived gene set can be usually confirmed using information about mutations, methylation, chromatin accessibility, copy number variation, or metabolomics assessment. This approach would naturally increase the cost of an experiment but also make the results reliable and reproducible.

5.2. Unsupervised learning approaches have a potential to overcome PPI biases

The key messages of BiCoN publication and the AMI testing suite look conflicting as one offers a novel network-constrained clustering method, and another suggests a limited value of PPI networks.

The AMI testing suite did not include BiCoN since only supervised methods fitted into benchmark criteria. In the case of unsupervised methods, clustering results do not have to correspond to the known phenotype. This makes BiCoN results not directly comparable to other Active Module Identification Methods (AMIMs) as supervised methods have an additional information source (patient groupings), and this gives them an advantage. Nevertheless, it was still possible to evaluate BiCoN with the same testing strategy. BiCoN performance was comparable to the best method in our benchmark study (DOMINO) in terms of gene set enrichment. However, BiCoN demonstrated a far better reaction to network permutations in task of identifying genes, predictive of a phenotype. These results can be explained by the fact that as an unsupervised method, BiCoN does not struggle with *overfitting* and, therefore, its results are more reliable. The results of the comparison are provided in the Appendix (Figure A.1).

To summarize, it would be wrong to suggest that PPI networks do not have any value for the disease module mining. However, it is easy to be misled by promising and sound strategies and thus lose a critical perception of algorithmic results. It is indeed hard to take into consideration all possible sources of bias, but with or without PPI networks, researchers are required to employ *in silico* control approaches to assess their results critically.

In silico control means that algorithmic results should be tested for different sources of biases. The AMI testing suite demonstrated the strategy for PPI value evaluation but did not cover all possible computational problems. For example, overfitting is a severe problem in bioinformatics as the data is usually highly dimensional and thus affected by a so-called *curse of dimensionality*. Cross-validation is a common approach to evaluate methods on previously unseen data that should be used for all supervised methods. Label permutation can help evaluate if a model overfitting and bootstrapping is extremely useful when the number of samples is not large enough.

On my side, to make the evaluation easier for PPI-based disease module mining methods, the code and all documentation for the AMI testing suite are made public and available for reuse.

5.3. Algorithmic roadblocks

The application of machine learning methods to the biomedical field is challenging. In classic computer science, the methods are usually developed to model a system that is mostly intuitively understood by a developer, such as a road system, book recommendation system, or text classification system. In the biomedical field, we apply machine learning to improve our understanding. However, given how much we yet do not know about complex organisms, it is very challenging to apply algorithms wisely. Several main limitations are discussed below:

Reproducibility crisis Reproducibility crisis is not specific to machine learning in biology, but affects the whole biomedical field [248]. Reproducibility issues can stem from poor models quality that leads to not reproducible results, as well as from poor software reporting

that leads to the publishing of unusable software [249]. In highly interdisciplinary fields like bioinformatics, the problem is particularly severe as researchers often combine roles of a software engineer, a machine learning engineer, and a biologist while lacking formal training in some of those disciplines. Therefore, sometimes not the best practices are used and several data limitations such as high dimensionality, low sample count and excessive noise are overlooked. This leads to non-reliable models that are unable to generalize to new data and therefore are useless for the community. The community can take full advantage of AI only if it was done wisely and its results are comprehensively reported while there is enough information about the data and the model itself that would allow other researchers to assess the experimental setup and reproduce it exactly.

During the conducted Ph.D. project, I have actively participated in the creation of a community reporting standard, "The AIME registry for artificial intelligence in biomedical research" (AIME) [250]. AIME registry is supposed to address the reproducibility issue by providing a registry that allows to quality-check an AI model and provides all necessary information for other researchers to reproduce the results.

While the issue with AI reporting is widely recognized in the biomedical community, many steps are required to overcome the problem. Plenty of effort has been invested in creating checklists for reporting, but those checklists are diverse and do not make AI reports accessible to the scientific community. Therefore, we created our reporting standard, assembled a team of more than 20 researchers from different institutions to provide their feedback, and created the first community-driven AI registry. AIME allows authors of new biomedical AIs to generate accessible, browsable, and citable reports that can be examined and reviewed by the scientific community <https://aime-registry.org>.

Interpretability Even when AI model is trained following all standards, there are still possibilities of mistakes due to "black box" nature of many AI models. Understanding the logic of a model, i.e. why the model makes exactly this prediction is essential for clinical usage. As models are trained on human collected data they can inherit its bias. An example of such bias can be a number of protein interactions (Publication 2). Other often mentioned biases are related to data collection and might result in underrepresentation certain ethnic groups or genders.

One of the motivations behind BiCoN was to provide users with an interpretable set of genes forming a disease mechanism, responsible for the patients stratification. Thus results can be more easily examined with respect to possible confounders. Classic clustering methods such as kmeans [251], DBSCAN [252] and other usually do not provide any interpretation of their results.

The above mentioned AIME registry also contributes to AI models interpretability, by making developers aware of possible limitations of their methods and suggesting approaches to interpret provided predictions.

Usability and trust Even reproducible and interpretable AI models are still often not used in clinical practice. Lack of communication between clinical practitioners and bioinformaticians,

difficult to comprehend models and complex interfaces (or an absence of thereof) are another roadblocks that do not make it easier for AI to enter into the clinical practice.

This issue can be addressed by embracing collaboration between bioinformaticians and clinicians and biologists. It is challenging to both sides to level with one another for a productive conversation, but yet it is the only way for AI to contribute to biomedical field.

During the course of my PhD, four collaborative projects with clinicians were successfully completed [109, 108, 253, 254]. Moreover, for BiCoN a web interface has been developed to accelerate its usage by biomedical researchers. To make the interface more accessible, several biologists have been surveyed to insure that the interface can be used by them in future.

Methods evaluation without gold standards In classic computer science, scientists often rely on gold standard data sets that are designed for specific domains, such MNIST for hand digits recognition [255], ImageNet for objects classification, or "Amazon reviews" for sentiment analysis. Bioinformaticians do not have the gold-standard diseases that can be used to benchmark their methods. While some methods can rely on human assistance (for instance in biomedical imaging analysis), genomic data can not be understood with a bare eye. Therefore, bioinformatics analysis often relies on different proxies that are often oversimplified and do not represent reality in all its complexity.

For the AMI testing suite, a particular challenge was to derive "biological meaningfulness" measures of the obtained gene modules. Evaluation of gene set's ability to predict a phenotype in question was a relatively easy choice, while the use of gene set over-representation analysis (with databases such as KEGG and DisGeNet) leaves more space for interpretation. Comparison of differentially expressed genes to KEGG pathways is a very conventional evaluation choice [256, 257, 258], but it also has several limitations mostly related to the fact that the current state of knowledge about disease-associated pathways is far from being complete [259].

5.4. Outlook

The field of biology made significant progress in the last 60 years due to the active use of algorithmic approaches for complex biological problems. The first bioinformatics software was created in the early 1960s - long before the success of next-generation sequencing (NGS). Since then, advancements in biology and computer science have allowed the application of highly complex algorithms to growing volumes of heterogeneous data. While many positive bioinformatics results are published every day, it is possible that sometimes, imperfect algorithms and incomplete databases will overlap and lead to inaccurate results.

A common example of such an issue is overfitting models trained on high-dimensional biological data [260]. Another relevant issue is the poor performance of published gene signatures on novel datasets [261]. However, while overfitting and irreproducibility are relatively well-known issues, more complex issues are hard to discover and evaluate precisely.

The main focus of this dissertation was to study PPIs network influence on disease module mining. This influence was particularly hard to fully understand because PPI networks

can also improve algorithms' performance. The benchmarks of BiCoN demonstrated that PPI networks increase the robustness of results to noise, batch effects, and other biases in transcriptomics data. Additionally, by focusing on only interacting proteins, the search space of an algorithm can be drastically reduced. This can compensate for the high dimensionality of the data (i.e., lifting the curse of dimensionality) and decrease the runtime of an algorithm.

Additionally, many of AMIMs were benchmarked using relatively small PPI networks (such as HPRD). Naturally, by restricting results to well-known proteins, the PPI network-based methods could outperform traditional differential expression methods by providing smaller gene sets consisting of more familiar genes.

It is possible that many other examples of such situations did not end up published but were just marked as unsuccessful experiments. Even though publication of the AMI testing suite [10] challenged the value of the previously conducted study (BiCoN [9]), discussion over negative results is absolutely necessary for the community as it allows to find the way out of the acknowledged problem.

I see the way out by continuously questioning existing results and exploring other methods to derive mechanistic definitions of patients' phenotypes. In particular, unsupervised methods should be prioritized because they can extract complex patterns without knowing to which group a patient initially belongs. This would allow the extraction of truly data-driven disease definitions which are not biased by pre-defined groupings. As we previously discussed, BiCoN outperforms other Active Module Identification Methods in terms of the ability to find predictive gene sets that actually take advantage of PPI interactions since BiCoN is an unsupervised method. This is indeed an inspiring result, once again demonstrating the power of unsupervised analysis.

Another critical point is that negative results are not as valuable if they do not help other academics to improve their research. Thus, a significant amount of time was invested in making the AMI testing suite usable by other researchers. Reproducible computational environments, documentation, and well-written programming code are crucial to ensure that other researchers can reuse, modify, and extend the testing approach. At this moment, several independent researchers are already working on an extension of the developed testing procedure to evaluate information gained from other molecular networks.

With the exponential development of AI for network analysis, we have a unique opportunity to analyze complex heterogeneous networks consisting of not only PPIs but also gene regulations, metabolic feedback, DNA interactions, binding, and other molecular interactions. The third publication reviewed several methods that combine single-cell ATAC-seq data with single-cell RNA-seq to extract regulatory networks for individual samples. While the typical number of samples in a single-cell study is still significantly lower than for bulk studies, this number is consistently increasing. Aggregation, analysis, and disease module mining in the single-cell sample-specific regulatory networks is an exciting challenge for the field.

Novel technologies are also developing at a fascinating speed. Just this year, the first single-cell metabolomics experimental method became available [262]. Given the role of metabolites in epigenetics, regulation of the immune system, inflammation control, and carcinogenesis [263], it is indeed thrilling to see how single-cell metabolomics will improve

our understanding of disease mechanisms.

To make complex AI algorithms ready for these exciting opportunities, the quality of the methods should be improved. AI must be transparent, interpretable and continuously challenged by the scientific community. At the moment, AI-based bioinformatics methods are used mainly by bioinformaticians and not by medical practitioners due to a lack of trust and understanding of modern AI methods. However, precision medicine of the future can not function this way, and therefore the changes are required on both sides.

Publication record

1. Arend L, Bernett J, Manz Q, Klug M, **Lazareva O**, Baumbach J, Bongiovanni D, List M. A systematic comparison of novel and existing differential analysis methods for CyTOF data. *Briefings in Bioinformatics*. 2021 Nov 30.
2. Bongiovanni, D., Klug, M., **Lazareva, O.**, Weidlich, S., Biasi, M., Ursu, S., Warth, S., Buske, C., Lukas, M., Spinner, C.D. and von Scheidt, M., 2021. SARS-CoV-2 infection is associated with a pro-thrombotic platelet phenotype. *Cell death & disease*, 12(1), pp.1-10.
3. Matschinske, J., Alcaraz, N., Benis, A., Golebiewski, M., Grimm, D.G., Heumos, L., Kacprowski, T., **Lazareva, O.**, List, M., Louadi, Z. and Pauling, J.K., 2021. The AIMe registry for artificial intelligence in biomedical research. *Nature methods*, 18(10), pp.1128-1131.
4. **Lazareva, O.***, Scherer, M*, Schmidt, F*, Walter, J., Baumbach, J., Schulz, M.H. and List, M., 2021. Machine learning for deciphering cell heterogeneity and gene regulation. *Nature Computational Science*, 1(3), pp.183-191.
5. **Lazareva, O.***, Klug, M.*, Kirmes, K., Rosenbaum, M., Lukas, M., Weidlich, S., Spinner, C.D., von Scheidt, M., Gosetti, R., Baumbach, J. and Ruland, J., 2021. Platelet Surface Protein Expression and Reactivity upon TRAP Stimulation after BNT162b2 Vaccination. *Thrombosis and Haemostasis*.
6. **Lazareva, O.**, Baumbach, J., List, M. and Blumenthal, D.B., 2021. On the limits of active module identification. *Briefings in Bioinformatics*.
7. **Lazareva, O.***, Lautizi, M.*, Fenn, A.*, List, M., Kacprowski, T. and Baumbach, J., 2021. Multi-Omics Analysis in a Network Context.
8. Sadegh, S., Skelton, J., Anastasi, E., Bernett, J., Blumenthal, D.B., Galindez, G., Salgado-Albarrán, M., **Lazareva, O.**, Flanagan, K., Cockell, S. and Nogales, C., 2021. Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nature Communications*, 12(1), pp.1-12.
9. **Lazareva, O.**, Canzar, S., Yuan, K., Baumbach, J., Blumenthal, D.B., Tieri, P., Kacprowski, T. and List, M., 2021. BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics*, 37(16), pp.2398-2404.

*shared first author

10. Klug M, Kirmes K, Han J, **Lazareva O**, Rosenbaum M, Viggiani G, von Scheidt M, Ruland J, Baumbach J, Condorelli G, Laugwitz KL. Mass cytometry of platelet-rich plasma: a new approach to analyze platelet surface expression and reactivity. *Platelets*. 2021 Dec 26:1-8.

A. Appendix

A.1. Assessment of BiCoN with the AMI testing suite

To assess the ability of BiCoN to learn from actual PPIs and not protein degrees, I evaluated its response to network perturbations and compared it to the leading AMIM DOMINO. Since PPI network choice did not affect methods performance, only one network was used in the comparison (HPRD). Three metrics were used for the assessment: mean mutual information wrt the phenotype, KEGG gene set enrichment p-values (negative and log-transformed), and overlap with DisGeNet disease-related gene sets (see Methods section for more details). I also used the most conservative network perturbation method: rewiring of a network while preserving its degree. This type of perturbation was the most challenging for all methods due to their inability to distinguish the actual PPI from its rewired version.

Figure A.1 demonstrates that BiCoN consistently outperforms DOMINO in all considered metrics. It also shows the decrease in scores in response to network rewiring even for the mean mutual information network that was the most challenging for all methods, including DOMINO.

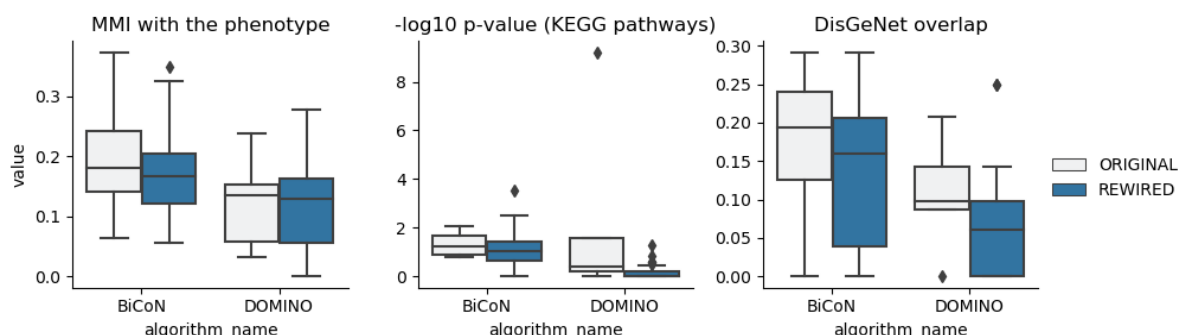


Figure A.1.: Comparison of BiCoN performance to the leading AMIM DOMINO from the AMI testing suite. Across all three metrics, BiCoN consistently performs better than DOMINO while demonstrating a decrease of the scores with respect to network perturbations. Unlike DOMINO, BiCoN also demonstrates a response to perturbations for the mean mutual information metric. This shows that BiCoN is the only method that finds disease modules predictive of a phenotype that actually relies on information from edges in PPI networks.

Acronyms

AI Artificial Intelligence.

AMIM Active Module Identification Methods.

ANN Artificial Neural Networks.

CNN Convolutional Neural Network.

DNA Deoxyribonucleic acid .

GNN Graph Neural Network.

miRNA Micro ribonucleic acid.

ML Machine Learning.

MLP Multilayer perceptron.

mRNA Messenger ribonucleic acid .

RF Random Forest.

RNA Ribonucleic acid .

RNN Recurrent neural network.

rRNA Ribosomal ribonucleic acid .

tRNA Transfer ribonucleic acid .

References

- [1] E. Björnson, J. Borén, and A. Mardinoglu. “Personalized cardiovascular disease prediction and treatment—a review of existing strategies and novel systems medicine tools”. In: *Frontiers in Physiology* 7 (2016), p. 2.
- [2] E. Lin and H.-Y. Lane. “Machine learning and systems genomics approaches for multi-omics data”. In: *Biomarker research* 5.1 (2017), pp. 1–6.
- [3] L. Kappler and R. Lehmann. “Mass-spectrometric multi-omics linked to function–state-of-the-art investigations of mitochondria in systems medicine”. In: *TrAC Trends in Analytical Chemistry* 119 (2019), p. 115635.
- [4] L. Hood and M. Flores. “A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory”. In: *New biotechnology* 29.6 (2012), pp. 613–624.
- [5] C. Nogales, Z. M. Mamdouh, M. List, C. Kiel, A. I. Casas, and H. H. Schmidt. “Network pharmacology: curing causal mechanisms instead of treating symptoms”. In: *Trends in Pharmaceutical Sciences* (2021). ISSN: 0165-6147. DOI: <https://doi.org/10.1016/j.tips.2021.11.004>.
- [6] Y. Silberberg, M. Kupiec, and R. Sharan. “GLADIATOR: a global approach for elucidating disease modules”. In: *Genome medicine* 9.1 (2017), pp. 1–14.
- [7] M. Kikuchi, S. Ogishima, T. Miyamoto, A. Miyashita, R. Kuwano, J. Nakaya, and H. Tanaka. “Identification of unstable network modules reveals disease modules associated with the progression of Alzheimer’s disease”. In: *PloS one* 8.11 (2013), e76162.
- [8] R. Batra, N. Alcaraz, K. Gitzhofer, J. Pauling, H. J. Ditzel, M. Hellmuth, J. Baumbach, and M. List. “On the performance of de novo pathway enrichment”. In: *NPJ systems biology and applications* 3.1 (2017), pp. 1–8.
- [9] O. Lazareva, S. Canzar, K. Yuan, J. Baumbach, D. B. Blumenthal, P. Tieri, T. Kacprowski, and M. List. “BiCoN: Network-constrained biclustering of patients and omics data”. In: *Bioinformatics* (Dec. 2020). btaa1076. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa1076](https://doi.org/10.1093/bioinformatics/btaa1076). URL: <https://doi.org/10.1093/bioinformatics/btaa1076>.
- [10] O. Lazareva, J. Baumbach, M. List, and D. B. Blumenthal. “On the limits of active module identification”. In: *Briefings in Bioinformatics* (2021).
- [11] M. Scherer, F. Schmidt, O. Lazareva, J. Walter, J. Baumbach, M. H. Schulz, and M. List. “Machine learning for deciphering cell heterogeneity and gene regulation”. In: *Nature Computational Science* 1.3 (2021), pp. 183–191.

- [12] L. Pray. "Discovery of DNA structure and function: Watson and Crick". In: *Nature Education* 1.1 (2008).
- [13] A. Klug. "Rosalind Franklin and the discovery of the structure of DNA". In: *Nature* 219.5156 (1968), pp. 808–810.
- [14] F. Crick. "Central dogma of molecular biology". In: *Nature* 227.5258 (1970), pp. 561–563.
- [15] H. Chial. "Rare genetic disorders: learning about genetic disease through gene mapping, SNPs, and microarray data". In: *Nature education* 1.1 (2008), p. 192.
- [16] D. N. Sheppard and L. S. Ostedgaard. "Understanding how cystic fibrosis mutations cause a loss of Cl⁻ channel function". In: *Molecular medicine today* 2.7 (1996), pp. 290–297.
- [17] J. Craig et al. "Complex diseases: Research and applications". In: *Nature Education* 1.1 (2008), p. 184.
- [18] J. C. Venter, H. O. Smith, and M. D. Adams. "The sequence of the human genome". In: *Clinical chemistry* 61.9 (2015), pp. 1207–1208.
- [19] F. Collins, E. Lander, J. Rogers, R. Waterston, and I. Conso. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945.
- [20] T. N. Williams. "Human red blood cell polymorphisms and malaria". In: *Current opinion in microbiology* 9.4 (2006), pp. 388–394.
- [21] C. Gonzaga-Jauregui, J. R. Lupski, and R. A. Gibbs. "Human genome sequencing in health and disease". In: *Annual review of medicine* 63 (2012), pp. 35–61.
- [22] M. L. Metzker. "Sequencing technologies—the next generation". In: *Nature reviews genetics* 11.1 (2010), pp. 31–46.
- [23] C. D. Bustamante, M. Francisco, and E. G. Burchard. "Genomics for the world". In: *Nature* 475.7355 (2011), pp. 163–165.
- [24] L. Bertram and R. E. Tanzi. "Genome-wide association studies in Alzheimer's disease". In: *Human molecular genetics* 18.R2 (2009), R137–R145.
- [25] S. L. Strickland, J. S. Reddy, M. Allen, A. N'songo, J. D. Burgess, M. M. Corda, T. Ballard, X. Wang, M. M. Carrasquillo, J. M. Biernacka, et al. "MAPT haplotype-stratified GWAS reveals differential association for AD risk variants". In: *Alzheimer's & Dementia* 16.7 (2020), pp. 983–1002.
- [26] A. Korte and A. Farlow. "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1 (2013), pp. 1–9.
- [27] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (2019), pp. 467–484.
- [28] A. M. Gross, S. S. Ajay, V. Rajan, C. Brown, K. Bluske, N. J. Burns, A. Chawla, A. J. Coffey, A. Malhotra, A. Scocchia, et al. "Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease". In: *Genetics in Medicine* 21.5 (2019), pp. 1121–1130.

-
- [29] X. Zhang, H. Chen, X. Zhao, J. Wang, and Y. Bai. *HER2 copy number variation detected by next-generation sequencing reveals HER2 expression heterogeneity in gastric cancer*. 2021.
- [30] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.
- [31] R. Bumgarner. "Overview of DNA microarrays: types, applications, and their future". In: *Current protocols in molecular biology* 101.1 (2013), pp. 22–1.
- [32] Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou. "Advanced applications of RNA sequencing and challenges". In: *Bioinformatics and biology insights* 9 (2015), BBI–S28991.
- [33] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari. "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences". In: *Briefings in bioinformatics* 19.2 (2018), pp. 286–302.
- [34] J. Shendure. "The beginning of the end for microarrays?" In: *Nature methods* 5.7 (2008), pp. 585–587.
- [35] U. Nagalakshmi, K. Waern, and M. Snyder. "RNA-Seq: a method for comprehensive transcriptome analysis". In: *Current protocols in molecular biology* 89.1 (2010), pp. 4–11.
- [36] S. Anders and W. Huber. "Differential expression analysis for sequence count data". In: *Nature Precedings* (2010), pp. 1–1.
- [37] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.
- [38] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. "A combined algorithm for genome-wide prediction of protein function". In: *Nature* 402.6757 (1999), pp. 83–86.
- [39] H. Keren, G. Lev-Maor, and G. Ast. "Alternative splicing and evolution: diversification, exon definition and function". In: *Nature Reviews Genetics* 11.5 (2010), pp. 345–355.
- [40] R. Piskol, G. Ramaswami, and J. B. Li. "Reliable identification of genomic variants from RNA-seq data". In: *The American Journal of Human Genetics* 93.4 (2013), pp. 641–651.
- [41] D. A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, and J. M. Akey. "A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data". In: *Genome research* 21.10 (2011), pp. 1728–1737.
- [42] L. Martens. "Proteomics databases and repositories". In: *Bioinformatics for comparative proteomics*. Springer, 2011, pp. 213–227.
- [43] G. N. Gowda and D. Djukovic. "Overview of mass spectrometry-based metabolomics: opportunities and challenges". In: *Mass Spectrometry in Metabolomics* (2014), pp. 3–12.
- [44] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool. "Proteomics: technologies and their applications". In: *Journal of chromatographic science* 55.2 (2017), pp. 182–196.
-

-
- [45] J. Cox and M. Mann. "Is proteomics the new genomics?" In: *Cell* 130.3 (2007), pp. 395–398.
- [46] D. Singh and V. Y. Soojin. "On the origin and evolution of SARS-CoV-2". In: *Experimental & Molecular Medicine* 53.4 (2021), pp. 537–547.
- [47] A. Team. "AlphaFold: A Solution to a 50-year-old Grand Challenge in Biology". In: (2021).
- [48] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, K. Tunyasuvunakool, O. Ronneberger, R. Bates, A. ŽilLOOKdek, A. Bridgland, et al. "High accuracy protein structure prediction using deep learning". In: *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)* 22 (2020), p. 24.
- [49] R. Zwanzig, A. Szabo, and B. Bagchi. "Levinthal's paradox". In: *Proceedings of the National Academy of Sciences* 89.1 (1992), pp. 20–22.
- [50] M. K. Higgins. "Can we AlphaFold our way out of the next pandemic?" In: *Journal of Molecular Biology* (2021), p. 167093.
- [51] A. Ali and A. Bagchi. "An overview of protein-protein interaction". In: *Current Chemical Biology* 9.1 (2015), pp. 53–65.
- [52] J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, and C. Orengo. "Transient protein-protein interactions: structural, functional, and network properties". In: *Structure* 18.10 (2010), pp. 1233–1243.
- [53] A. S. Worthington, H. Rivera Jr, J. W. Torpey, M. D. Alexander, and M. D. Burkart. "Mechanism-based protein cross-linking probes to investigate carrier protein-mediated biosynthesis". In: *ACS chemical biology* 1.11 (2006), pp. 687–691.
- [54] G. C. Koh, P. Porras, B. Aranda, H. Hermjakob, and S. E. Orchard. "Analyzing protein-protein interaction networks". In: *Journal of proteome research* 11.4 (2012), pp. 2014–2031.
- [55] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, et al. "The BioGRID interaction database: 2013 update". In: *Nucleic acids research* 41.D1 (2012), pp. D816–D823.
- [56] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, et al. "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets". In: *Nucleic acids research* 49.D1 (2021), pp. D605–D612.
- [57] M. Kotlyar, C. Pastrello, Z. Malik, and I. Jurisica. "IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species". In: *Nucleic acids research* 47.D1 (2019), pp. D581–D589.
- [58] D. O. Perkins, C. Jeffries, and P. Sullivan. "Expanding the 'central dogma': the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia". In: *Molecular Psychiatry* 10.1 (2005), pp. 69–78.
-

-
- [59] J. Van Vliet, N. Oates, and E. Whitelaw. “Epigenetic mechanisms in the context of complex diseases”. In: *Cellular and molecular life sciences* 64.12 (2007), pp. 1531–1538.
- [60] A. P. Feinberg and M. D. Fallin. “Epigenetics at the crossroads of genes and the environment”. In: *Jama* 314.11 (2015), pp. 1129–1130.
- [61] E. Cazaly, J. Saad, W. Wang, C. Heckman, M. Ollikainen, and J. Tang. “Making sense of the epigenome using data integration approaches”. In: *Frontiers in pharmacology* 10 (2019), p. 126.
- [62] L. D. Moore, T. Le, and G. Fan. “DNA methylation and its basic function”. In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.
- [63] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. “Chromatin accessibility and the regulatory epigenome”. In: *Nature Reviews Genetics* 20.4 (2019), pp. 207–220.
- [64] C. B. Clish. “Metabolomics: an emerging but powerful tool for precision medicine”. In: *Molecular Case Studies* 1.1 (2015), a000588.
- [65] S. Shaham-Niv, S. Rencus-Lazar, and E. Gazit. “Metabolite medicine offers a path beyond lists of metabolites”. In: *Communications Chemistry* 4.1 (2021), pp. 1–5.
- [66] S. Linari and G. Castaman. “Clinical manifestations and management of Gaucher disease”. In: *Clinical Cases in Mineral and Bone Metabolism* 12.2 (2015), p. 157.
- [67] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight. “Defining the human microbiome”. In: *Nutrition reviews* 70.suppl_1 (2012), S38–S44.
- [68] Y. Xia. “Correlation and association analyses in microbiome study integrating multi-omics in health and disease”. In: *Progress in Molecular Biology and Translational Science* 171 (2020), pp. 309–491.
- [69] J. M. Brenchley and D. C. Douek. “Microbial translocation across the GI tract”. In: *Annual review of immunology* 30 (2012), pp. 149–173.
- [70] Y. Chen, F. Yang, H. Lu, B. Wang, Y. Chen, D. Lei, Y. Wang, B. Zhu, and L. Li. “Characterization of fecal microbial communities in patients with liver cirrhosis”. In: *Hepatology* 54.2 (2011), pp. 562–572.
- [71] T. de Sablet, M. B. Piazuelo, C. L. Shaffer, B. G. Schneider, M. Asim, R. Chaturvedi, L. E. Bravo, L. A. Sicinski, A. G. Delgado, R. M. Mera, et al. “Phylogeographic origin of *Helicobacter pylori* is a determinant of gastric cancer risk”. In: *Gut* 60.9 (2011), pp. 1189–1195.
- [72] F. Sommer and F. Bäckhed. “The gut microbiota—masters of host development and physiology”. In: *Nature reviews microbiology* 11.4 (2013), pp. 227–238.
- [73] S. L. Russell, M. J. Gold, M. Hartmann, B. P. Willing, L. Thorson, M. Wlodarska, N. Gill, M.-R. Blanchet, W. W. Mohn, K. M. McNagny, et al. “Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma”. In: *EMBO reports* 13.5 (2012), pp. 440–447.
-

-
- [74] J. F. Cryan and T. G. Dinan. “Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour”. In: *Nature reviews neuroscience* 13.10 (2012), pp. 701–712.
- [75] X. Tang, Y. Huang, J. Lei, H. Luo, and X. Zhu. “The single-cell sequencing: new developments and medical applications”. In: *Cell & bioscience* 9.1 (2019), pp. 1–9.
- [76] B. Hwang, J. H. Lee, and D. Bang. “Single-cell RNA sequencing technologies and bioinformatics pipelines”. In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14.
- [77] Z. Chen, F. Gong, L. Wan, and L. Ma. “RobustClone: a robust PCA method for tumor clone and evolution inference from single-cell sequencing data”. In: *Bioinformatics* 36.11 (2020), pp. 3299–3306.
- [78] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature biotechnology* 37.1 (2019), pp. 38–44.
- [79] D. Kobak and P. Berens. “The art of using t-SNE for single-cell transcriptomics”. In: *Nature communications* 10.1 (2019), pp. 1–14.
- [80] C. Lin, S. Jain, H. Kim, and Z. Bar-Joseph. “Using neural networks for reducing the dimensions of single-cell RNA-Seq data”. In: *Nucleic acids research* 45.17 (2017), e156–e156.
- [81] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, et al. “Decoding the regulatory network of early blood development from single-cell gene expression measurements”. In: *Nature biotechnology* 33.3 (2015), pp. 269–276.
- [82] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, et al. “SCENIC: single-cell regulatory network inference and clustering”. In: *Nature methods* 14.11 (2017), pp. 1083–1086.
- [83] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. “Reversed graph embedding resolves complex single-cell trajectories”. In: *Nature methods* 14.10 (2017), pp. 979–982.
- [84] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. “Diffusion pseudotime robustly reconstructs lineage branching”. In: *Nature methods* 13.10 (2016), pp. 845–848.
- [85] Y. Hu, Q. An, K. Sheu, B. Trejo, S. Fan, and Y. Guo. “Single cell multi-omics technology: methodology and application”. In: *Frontiers in cell and developmental biology* 6 (2018), p. 28.
- [86] K. A. Knouse, J. Wu, and A. Amon. “Assessment of megabase-scale somatic copy number variation using single-cell sequencing”. In: *Genome research* 26.3 (2016), pp. 376–384.
- [87] M. Zamanighomi, Z. Lin, T. Daley, X. Chen, Z. Duren, A. Schep, W. J. Greenleaf, and W. H. Wong. “Unsupervised clustering and epigenetic classification of single cells”. In: *Nature communications* 9.1 (2018), pp. 1–8.
-

-
- [88] S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". In: *Nature methods* 11.8 (2014), pp. 817–820.
- [89] S. Darmanis, S. A. Sloan, D. Croote, M. Mignardi, S. Chernikova, P. Samghababi, Y. Zhang, N. Neff, M. Kowarsky, C. Caneda, et al. "Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma". In: *Cell reports* 21.5 (2017), pp. 1399–1410.
- [90] S. Bian, Y. Hou, X. Zhou, X. Li, J. Yong, Y. Wang, W. Wang, J. Yan, B. Hu, H. Guo, et al. "Single-cell multiomics sequencing and analyses of human colorectal cancer". In: *Science* 362.6418 (2018), pp. 1060–1063.
- [91] Q. H. Nguyen, N. Pervolarakis, K. Nee, and K. Kessenbrock. "Experimental considerations for single-cell RNA sequencing approaches". In: *Frontiers in cell and developmental biology* 6 (2018), p. 108.
- [92] A. Nguyen, W. H. Khoo, I. Moran, P. I. Croucher, and T. G. Phan. "Single cell RNA sequencing of rare immune cell populations". In: *Frontiers in immunology* 9 (2018), p. 1553.
- [93] Q. H. Nguyen, N. Pervolarakis, K. Blake, D. Ma, R. T. Davis, N. James, A. T. Phung, E. Willey, R. Kumar, E. Jabart, et al. "Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity". In: *Nature communications* 9.1 (2018), pp. 1–12.
- [94] S. V. Puram, I. Tirosh, A. S. Parikh, A. P. Patel, K. Yizhak, S. Gillespie, C. Rodman, C. L. Luo, E. A. Mroz, K. S. Emerick, et al. "Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer". In: *Cell* 171.7 (2017), pp. 1611–1624.
- [95] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, et al. "Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing". In: *Nature medicine* 24.7 (2018), pp. 978–985.
- [96] C. Zheng, L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, et al. "Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing". In: *Cell* 169.7 (2017), pp. 1342–1356.
- [97] J. H. Gross. *Mass spectrometry: a textbook*. Springer Science & Business Media, 2006.
- [98] L. Zhang and A. Vertes. "Single-cell mass spectrometry approaches to explore cellular heterogeneity". In: *Angewandte Chemie International Edition* 57.17 (2018), pp. 4466–4477.
- [99] T. M. Evers, M. Hochane, S. J. Tans, R. M. Heeren, S. Semrau, P. Nemes, and A. Mashaghi. *Deciphering metabolic heterogeneity by single-cell analysis*. 2019.
- [100] M. MartíLOOKnez-GarciLOOKa, F. Santos, M. Moreno-Paz, V. Parro, and J. Antón. "Unveiling viral–host interactions within the ‘microbial dark matter’". In: *Nature communications* 5.1 (2014), pp. 1–8.
-

-
- [101] N. Slavov. "Scaling up single-cell proteomics". In: *Molecular & Cellular Proteomics* (2021), p. 100179.
- [102] M. Binnewies, E. W. Roberts, K. Kersten, V. Chan, D. F. Fearon, M. Merad, L. M. Coussens, D. I. Gabrilovich, S. Ostrand-Rosenberg, C. C. Hedrick, et al. "Understanding the tumor immune microenvironment (TIME) for effective therapy". In: *Nature medicine* 24.5 (2018), pp. 541–550.
- [103] M. Crow, A. Paul, S. Ballouz, Z. Huang, and J. Gillis. *Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor*. *Nat Commun*. 2018.
- [104] Y. Zhu, M. Scheibinger, D. C. Ellwanger, J. F. Krey, D. Choi, R. T. Kelly, S. Heller, and P. G. Barr-Gillespie. "Single-cell proteomics reveals changes in expression during hair-cell development". In: *Elife* 8 (2019), e50777.
- [105] R. T. Kelly. "Single-cell proteomics: progress and prospects". In: *Molecular & Cellular Proteomics* 19.11 (2020), pp. 1739–1748.
- [106] J. M. Perkel. "Single-cell proteomics takes centre stage." In: *Nature* 597.7877 (2021), pp. 580–582.
- [107] R. Gadalla, B. Noamani, B. L. MacLeod, R. J. Dickson, M. Guo, W. Xu, S. Lukhele, H. J. Elsaesser, A. R. A. Razak, N. Hirano, et al. "Validation of CyTOF against flow cytometry for immunological studies and monitoring of human cancer clinical trials". In: *Frontiers in oncology* 9 (2019), p. 415.
- [108] D. Bongiovanni, M. Klug, O. Lazareva, S. Weidlich, M. Biasi, S. Ursu, S. Warth, C. Buske, M. Lukas, C. D. Spinner, et al. "SARS-CoV-2 infection is associated with a pro-thrombotic platelet phenotype". In: *Cell death & disease* 12.1 (2021), pp. 1–10.
- [109] M. E. Klug, O. Lazareva, K. Kirmes, M. Rosenbaum, M. Lukas, S. Weidlich, C. D. Spinner, M. von Scheidt, R. Gosetti, J. Baumbach, et al. "Platelet expression and reactivity after BNT162b2 vaccine administration". In: *medRxiv* (2021).
- [110] Y. Hasin, M. Seldin, and A. Lusic. "Multi-omics approaches to disease". In: *Genome biology* 18.1 (2017), pp. 1–15.
- [111] M. J. Goldman, J. Zhang, N. A. Fonseca, I. Cortés-Ciriano, Q. Xiang, B. Craft, E. Piñero-Yáñez, B. D. O'Connor, W. Bazant, E. Barrera, et al. "A user guide for the online exploration and visualization of PCAWG data". In: *Nature communications* 11.1 (2020), pp. 1–9.
- [112] T. C. Silva, A. Colaprico, C. Olsen, F. D'Angelo, G. Bontempi, M. Ceccarelli, and H. Noushmehr. "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages". In: *F1000Research* 5 (2016).
- [113] M. J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, et al. "Visualizing and interpreting cancer genomics data via the Xena platform". In: *Nature biotechnology* 38.6 (2020), pp. 675–678.
-

-
- [114] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, et al. "Next-generation characterization of the cancer cell line encyclopedia". In: *Nature* 569.7757 (2019), pp. 503–508.
- [115] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, et al. "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal". In: *Science signaling* 6.269 (2013), p11–p11.
- [116] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, et al. "Personal omics profiling reveals dynamic molecular and medical phenotypes". In: *Cell* 148.6 (2012), pp. 1293–1307.
- [117] M. R. Corces, J. D. Buenrostro, B. Wu, P. G. Greenside, S. M. Chan, J. L. Koenig, M. P. Snyder, J. K. Pritchard, A. Kundaje, W. J. Greenleaf, et al. "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution". In: *Nature genetics* 48.10 (2016), pp. 1193–1203.
- [118] F. C. Navarro, H. Mohsen, C. Yan, S. Li, M. Gu, W. Meyerson, and M. Gerstein. "Genomics and data science: an application within an umbrella". In: *Genome biology* 20.1 (2019), pp. 1–11.
- [119] M. Krassowski, V. Das, S. K. Sahu, and B. B. Misra. "State of the field in multi-omics research: From computational needs to data mining and sharing". In: *Frontiers in Genetics* 11 (2020).
- [120] G. T. Jung, K.-P. Kim, and K. Kim. "How to interpret and integrate multi-omics data at systems level". In: *Animal cells and systems* 24.1 (2020), pp. 1–7.
- [121] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika. "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and biology insights* 14 (2020), p. 1177932219899051.
- [122] Z.-Z. Tang, G. Chen, Q. Hong, S. Huang, H. M. Smith, R. D. Shah, M. Scholz, and J. F. Ferguson. "Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites". In: *Frontiers in Genetics* 10 (2019), p. 454.
- [123] M. Y. Lim, S. Hong, B.-M. Kim, Y. Ahn, H.-J. Kim, and Y.-D. Nam. "Changes in microbiome and metabolomic profiles of fecal samples stored with stabilizing solution at room temperature: a pilot study". In: *Scientific reports* 10.1 (2020), pp. 1–9.
- [124] D. Reiman, B. T. Layden, and Y. Dai. "MiMeNet: Exploring microbiome-metabolome relationships using neural networks". In: *PLoS Computational Biology* 17.5 (2021), e1009021.
- [125] M. Shaffer, K. Thurimella, K. Quinn, K. Doenges, X. Zhang, S. Bokatzian, N. Reisdorph, and C. A. Lozupone. "AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data". In: *BMC bioinformatics* 20.1 (2019), pp. 1–11.
- [126] N. J. Nilsson. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.
-

-
- [127] R. B. Altman, M. Buda, X. J. Chai, M. W. Carillo, R. O. Chen, and N. F. Abernethy. "RiboWeb: An ontology-based system for collaborative molecular biology". In: *IEEE Intelligent Systems and Their Applications* 14.5 (1999), pp. 68–76.
- [128] V. D'Argenio. "The high-throughput analyses era: are we ready for the data struggle?" In: *High-throughput* 7.1 (2018), p. 8.
- [129] R. M. Karp. "Heuristic algorithms in computational molecular biology". In: *Journal of Computer and System Sciences* 77.1 (2011), pp. 122–128.
- [130] S. Consoli and K. Darby-Dowman. *Combinatorial optimization and metaheuristics*. Tech. rep. Brunel University, 2006.
- [131] Blum and Roli. "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison". In: *CSURV: Computing Surveys* 35 (2003).
- [132] W. R. Pearson and D. J. Lipman. "Improved tools for biological sequence comparison". In: *Proceedings of the National Academy of Sciences* 85.8 (1988), pp. 2444–2448.
- [133] D. G. Higgins and P. M. Sharp. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". In: *Gene* 73.1 (1988), pp. 237–244.
- [134] S. F. Altschul and D. J. Lipman. "Protein database searches for multiple alignments." In: *Proceedings of the National Academy of Sciences* 87.14 (1990), pp. 5509–5513.
- [135] J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou. "Graemlin: general and robust alignment of multiple large interaction networks". In: *Genome research* 16.9 (2006), pp. 1169–1181.
- [136] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. "Conserved pathways within bacteria and yeast as revealed by global protein network alignment". In: *Proceedings of the National Academy of Sciences* 100.20 (2003), pp. 11394–11399.
- [137] C. Lund and M. Yannakakis. "On the hardness of approximating minimization problems". In: *Journal of the ACM (JACM)* 41.5 (1994), pp. 960–981.
- [138] I. Arel, D. C. Rose, and T. P. Karnowski. "Research frontier: deep machine learning—a new frontier in artificial intelligence research". In: *IEEE computational intelligence magazine* 5.4 (2010), pp. 13–18.
- [139] F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.
- [140] B. Widrow and M. A. Lehr. "30 years of adaptive neural networks: perceptron, mada-line, and backpropagation". In: *Proceedings of the IEEE* 78.9 (1990), pp. 1415–1442.
- [141] B. G. Buchanan. "A (very) brief history of artificial intelligence". In: *Ai Magazine* 26.4 (2005), pp. 53–53.
- [142] F. Bayes. "An essay towards solving a problem in the doctrine of chances". In: *Biometrika* 45.3-4 (1958), pp. 296–315.
-

-
- [143] M. Merriman. "On the history of the method of least squares". In: *The Analyst* 4.2 (1877), pp. 33–36.
- [144] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.
- [145] T. K. Ho. "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.
- [146] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.
- [147] R. E. Schapire. "The strength of weak learnability". In: *Machine learning* 5.2 (1990), pp. 197–227.
- [148] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.
- [149] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel. "Using machine learning algorithms for breast cancer risk prediction and diagnosis". In: *Procedia Computer Science* 83 (2016), pp. 1064–1069.
- [150] L. Pan, G. Liu, F. Lin, S. Zhong, H. Xia, X. Sun, and H. Liang. "Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia". In: *Scientific reports* 7.1 (2017), pp. 1–9.
- [151] L. Macyszyn, H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, L. Roccograndi, et al. "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques". In: *Neuro-oncology* 18.3 (2015), pp. 417–425.
- [152] F. Schmidt, F. Kern, P. Ebert, N. Baumgarten, and M. H. Schulz. "TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis". In: *Bioinformatics* 35.9 (2019), pp. 1608–1609.
- [153] C. M. Lynch, B. Abdollahi, J. D. Fuqua, R. Alexandra, J. A. Bartholomai, R. N. Balgeman, V. H. van Berkel, and H. B. Frieboes. "Prediction of lung cancer patient survival via supervised machine learning classification techniques". In: *International journal of medical informatics* 108 (2017), pp. 1–8.
- [154] K. P. F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. doi: 10.1080/14786440109462720.
- [155] L. McInnes, J. Healy, and J. Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).
- [156] L. Van der Maaten and G. Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).
- [157] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker. "Analysis of dimensionality reduction techniques on big data". In: *IEEE Access* 8 (2020), pp. 54776–54788.
-

-
- [158] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20.5 (2019), pp. 273–282.
- [159] S. R. Newcomer, J. F. Steiner, and E. A. Bayliss. "Identifying subgroups of complex patients with cluster analysis." In: *The American journal of managed care* 17.8 (2011), e324–32.
- [160] M. J. Lercher, A. O. Urrutia, and L. D. Hurst. "Clustering of housekeeping genes provides a unified model of gene order in the human genome". In: *Nature genetics* 31.2 (2002), pp. 180–183.
- [161] A. Bernstein and E. Burnaev. "Reinforcement learning in computer vision". In: *Tenth International Conference on Machine Vision (ICMV 2017)*. Vol. 10696. International Society for Optics and Photonics. 2018, 106961S.
- [162] S. Almahdi and S. Y. Yang. "An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown". In: *Expert Systems with Applications* 87 (2017), pp. 267–279.
- [163] A. L. Bazzan. "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control". In: *Autonomous Agents and Multi-Agent Systems* 18.3 (2009), pp. 342–375.
- [164] J. DiGiovanna, B. Mahmoudi, J. Fortes, J. C. Principe, and J. C. Sanchez. "Coadaptive brain-machine interface via reinforcement learning". In: *IEEE transactions on biomedical engineering* 56.1 (2008), pp. 54–64.
- [165] L. Wang, W. Zhang, X. He, and H. Zha. "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2447–2456.
- [166] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie. "Deep learning and its applications in biomedicine". In: *Genomics, proteomics & bioinformatics* 16.1 (2018), pp. 17–32.
- [167] S. Lawrence, C. L. Giles, and A. C. Tsoi. "Lessons in neural network training: Overfitting may be harder than expected". In: *AAAI/IAAI*. Citeseer. 1997, pp. 540–545.
- [168] S. Albawi, T. A. Mohammed, and S. Al-Zawi. "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)*. Ieee. 2017, pp. 1–6.
- [169] K. Jnawali, M. R. Arbabshirani, N. Rao, and A. A. Patel. "Deep 3D convolution neural network for CT brain hemorrhage classification". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics. 2018, p. 105751C.
- [170] J. Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.
-

-
- [171] L. Gondara. "Medical image denoising using convolutional denoising autoencoders". In: *2016 IEEE 16th international conference on data mining workshops (ICDMW)*. IEEE, 2016, pp. 241–246.
- [172] S. Sharma, I. Umar, L. Ospina, D. Wong, and H. R. Tizhoosh. "Stacked autoencoders for medical image search". In: *International Symposium on Visual Computing*. Springer, 2016, pp. 45–54.
- [173] A. Sherstinsky. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306.
- [174] X. Zhou, Y. Li, and W. Liang. "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020).
- [175] A. N. Jagannatha and H. Yu. "Bidirectional RNN for medical event detection in electronic health records". In: *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2016. NIH Public Access, 2016, p. 473.
- [176] F. Liu, Y. Miao, Y. Liu, and T. Hou. "RNN-VirSeeker: a deep learning method for identification of short viral sequences from metagenomes". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020).
- [177] Q. Liu, L. Fang, G. Yu, D. Wang, C.-L. Xiao, and K. Wang. "Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data". In: *Nature communications* 10.1 (2019), pp. 1–11.
- [178] X. Pan and H.-B. Shen. "Inferring disease-associated microRNAs using semi-supervised multi-label graph convolutional networks". In: *Isience* 20 (2019), pp. 265–277.
- [179] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, and D. Xu. "scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses". In: *Nature communications* 12.1 (2021), pp. 1–11.
- [180] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang. "Graph neural networks and their current applications in bioinformatics". In: *Frontiers in Genetics* 12 (2021).
- [181] V. Gligorijević, P. D. Renfrew, T. Kosciolk, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, et al. "Structure-based protein function prediction using graph convolutional networks". In: *Nature communications* 12.1 (2021), pp. 1–14.
- [182] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In: *Bioinformatics* 37.15 (2021), pp. 2112–2120.
- [183] Y. Bathaee. "The artificial intelligence black box and the failure of intent and causation". In: *Harv. JL & Tech.* 31 (2017), p. 889.
- [184] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard. "Navigating the pitfalls of applying machine learning in genomics". In: *Nature Reviews Genetics* (2021), pp. 1–13.
-

-
- [185] C. Yan, J. Ma, H. Luo, G. Zhang, and J. Luo. "A novel feature selection method for high-dimensional biomedical data based on an improved binary clonal flower pollination algorithm". In: *Human heredity* 84.1 (2019), pp. 34–46.
- [186] D. Gatherer. "So what do we really mean when we say that systems biology is holistic?" In: *BMC systems biology* 4.1 (2010), pp. 1–12.
- [187] F. Mazzocchi. "Complexity and the reductionism–holism debate in systems biology". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 4.5 (2012), pp. 413–427.
- [188] A.-L. Barabási. *Network medicine—from obesity to the "diseasome"*. 2007.
- [189] N. A. Christakis and J. H. Fowler. "The spread of obesity in a large social network over 32 years". In: *New England journal of medicine* 357.4 (2007), pp. 370–379.
- [190] K.-I. Goh and I.-G. Choi. "Exploring the human diseasome: the human disease network". In: *Briefings in functional genomics* 11.6 (2012), pp. 533–542.
- [191] J. Gorky and J. Schwaber. "Conceptualization of a parasympathetic endocrine system". In: *Frontiers in neuroscience* 13 (2019), p. 1008.
- [192] W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, and D. de Ridder. "Topology of molecular interaction networks". In: *BMC systems biology* 7.1 (2013), pp. 1–15.
- [193] E. M. Blackwood and J. T. Kadonaga. "Going the distance: a current view of enhancer action". In: *Science* 281.5373 (1998), pp. 60–63.
- [194] D. Wang and X.-D. Fu. "DNA interaction networks: an information highway for regulated gene expression in the 3-dimensional space of the nucleus". In: *Cell research* 19.12 (2009), pp. 1316–1319.
- [195] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou. "An overview of SNP interactions in genome-wide association studies". In: *Briefings in functional genomics* 14.2 (2015), pp. 143–155.
- [196] X. Zhu, Z. Duren, and W. H. Wong. "Modeling regulatory network topology improves genome-wide analyses of complex human traits". In: *Nature communications* 12.1 (2021), pp. 1–15.
- [197] K. Titeca, I. Lemmens, J. Tavernier, and S. Eyckerman. "Discovering cellular protein-protein interactions: Technological strategies and opportunities". In: *Mass spectrometry reviews* 38.1 (2019), pp. 79–111.
- [198] M. Rezaei-Tavirani, S. Rezaei-Tavirani, V. Mansouri, M. Rostami-Nejad, and M. Rezaei-Tavirani. "Protein-protein interaction network analysis for a biomarker panel related to human esophageal adenocarcinoma". In: *Asian Pacific journal of cancer prevention: APJCP* 18.12 (2017), p. 3357.
- [199] J. I. Castrillo, P. Pir, and S. G. Oliver. "Yeast Systems Biology: towards a systems understanding of regulation of eukaryotic networks in complex diseases and biotechnology". In: *Handbook of Systems Biology*. Elsevier, 2013, pp. 343–365.
-

- [200] E. H. Wong, J. C. Fox, M. Y. Ng, and C.-M. Lee. "Toward personalized medicine in the neuropsychiatric field". In: *International review of neurobiology* 101 (2011), pp. 329–349.
- [201] K. Strimbu and J. A. Tavel. "What are biomarkers?" In: *Current Opinion in HIV and AIDS* 5.6 (2010), p. 463.
- [202] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, et al. "Molecular portraits of human breast tumours". In: *nature* 406.6797 (2000), pp. 747–752.
- [203] T. Sorlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, et al. "Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein P Lonning, Borresen-Dale AL Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". In: *Proc Natl Acad Sci USA* 98 (2001), pp. 10869–74.
- [204] E. Collisson, J. Campbell, A. Brooks, A. Berger, W. Lee, J. Chmielecki, D. Beer, L. Cope, C. Creighton, L. Danilova, et al. "Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network". In: *Nature* 511.7511 (2014), pp. 543–550.
- [205] M. Van Vliet, F. Reyal, H. Horlings, and M. Vijver. "van de, Reinders MJ, Wessels LF: Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability". In: *BMC Genomics* 9 (2008), p. 375.
- [206] D. Venet, J. E. Dumont, and V. Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In: *PLoS computational biology* 7.10 (2011), e1002240.
- [207] W. Cui, H. Xue, L. Wei, J. Jin, X. Tian, and Q. Wang. "High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis". In: *Human genomics* 15.1 (2021), pp. 1–9.
- [208] O. Lazareva, M. Lautizi, A. Fenn, M. List, T. Kacprowski, and J. Baumbach. "Multi-Omics Analysis in a Network Context". In: *Reference Module in Biomedical Sciences*. Elsevier, 2020. ISBN: 978-0-12-801238-3. DOI: <https://doi.org/10.1016/B978-0-12-801238-3.11647-2>.
- [209] A. Sharma, A. Halu, J. L. Decano, M. Padi, Y.-Y. Liu, R. B. Prasad, J. Fadista, M. Santolini, J. Menche, S. T. Weiss, et al. "Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates Type 2 Diabetes associated genes". In: *NPJ systems biology and applications* 4.1 (2018), pp. 1–11.
- [210] A. Halu, S. Liu, S. H. Baek, B. D. Hobbs, G. M. Hunninghake, M. H. Cho, E. K. Silverman, and A. Sharma. "Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis". In: *Human molecular genetics* 28.14 (2019), pp. 2352–2364.

-
- [211] A. Sharma, J. Menche, C. C. Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, et al. "A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma". In: *Human molecular genetics* 24.11 (2015), pp. 3005–3020.
- [212] Z. Ding, W. Guo, and J. Gu. "ClustEx2: gene module identification using density-based network hierarchical clustering". In: *2018 Chinese Automation Congress (CAC)*. IEEE, 2018, pp. 2407–2412.
- [213] S. D. Ghiassian, J. Menche, and A.-L. Barabási. "A DIseAse MOdule Detection (DI-AMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome". In: *PLoS computational biology* 11.4 (2015), e1004120.
- [214] H. Levi, R. Elkon, and R. Shamir. "DOMINO: a network-based active module identification algorithm with reduced rate of false calls". In: *Molecular Systems Biology* 17.1 (2021), e9593.
- [215] N. Alcaraz, H. Küçük, J. Weile, A. Wipat, and J. Baumbach. "KeyPathwayMiner: detecting case-specific biological pathways using expression data". In: *Internet Mathematics* 7.4 (2011), pp. 299–313.
- [216] M. A. Reyna, M. D. Leiserson, and B. J. Raphael. "Hierarchical HotNet: identifying hierarchies of altered subnetworks". In: *Bioinformatics* 34.17 (2018), pp. i972–i980.
- [217] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.
- [218] M. Kanehisa. "Toward understanding the origin and evolution of cellular organisms". In: *Protein Science* 28.11 (2019), pp. 1947–1951.
- [219] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe. "KEGG: integrating viruses and cellular organisms". In: *Nucleic acids research* 49.D1 (2021), pp. D545–D551.
- [220] S. J. Larsen, H. H. Schmidt, and J. Baumbach. "De Novo and supervised Endophenotyping using network-guided ensemble learning". In: *Systems Medicine* 3.1 (2020), pp. 8–21.
- [221] A. Liaw and M. Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: <https://CRAN.R-project.org/doc/Rnews/>.
- [222] M. Dorigo, M. Birattari, and T. Stutzle. "Ant colony optimization". In: *IEEE computational intelligence magazine* 1.4 (2006), pp. 28–39.
- [223] E. Aarts, E. H. Aarts, and J. K. Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 2003.
- [224] E.-G. Talbi. *Metaheuristics: from design to implementation*. Vol. 74. John Wiley & Sons, 2009.
-

-
- [225] H. Ma, E. E. Schadt, L. M. Kaplan, and H. Zhao. “COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method”. In: *Bioinformatics* 27.9 (2011), pp. 1290–1298.
- [226] R. Breitling, A. Amtmann, and P. Herzyk. “Graph-based iterative Group Analysis enhances microarray interpretation”. In: *BMC bioinformatics* 5.1 (2004), pp. 1–10.
- [227] Ş. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. “Gene expression network analysis and applications to immunology”. In: *Bioinformatics* 23.7 (2007), pp. 850–858.
- [228] N. Alcaraz, J. Pauling, R. Batra, E. Barbosa, A. Junge, A. G. Christensen, V. Azevedo, H. J. Ditzel, and J. Baumbach. “KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape”. In: *BMC systems biology* 8.1 (2014), pp. 1–6.
- [229] M. List, N. Alcaraz, M. Dissing-Hansen, H. J. Ditzel, J. Mollenhauer, and J. Baumbach. “KeyPathwayMinerWeb: online multi-omics network enrichment”. In: *Nucleic Acids Research* 44.W1 (2016), W98–W104.
- [230] G. Barel and R. Herwig. “NetCore: a network propagation approach using node coreness”. In: *Nucleic acids research* 48.17 (2020), e98–e98.
- [231] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.
- [232] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O’Donnell, G. Leung, R. McAdam, et al. “The BioGRID interaction database: 2019 update”. In: *Nucleic acids research* 47.D1 (2019), pp. D529–D541.
- [233] D. Alonso-Lopez, M. A. Gutiérrez, K. P. Lopes, C. Prieto, R. SantamarıLOOKa, and J. De Las Rivas. “APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks”. In: *Nucleic acids research* 44.W1 (2016), W529–W535.
- [234] D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas. “APID database: redefining protein–protein interaction experimental evidences and binary interactomes”. In: *Database* 2019 (2019).
- [235] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al. “Human protein reference database—2009 update”. In: *Nucleic acids research* 37.suppl_1 (2009), pp. D767–D772.
- [236] J. Piñero, J. M. RamiLOOKrez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong. “The DisGeNET knowledge platform for disease genomics: 2019 update”. In: *Nucleic acids research* 48.D1 (2020), pp. D845–D855.
- [237] B. Bao, C. Zheng, B. Yang, Y. Jin, K. Hou, Z. Li, X. Zheng, S. Yu, X. Zhang, Y. Fan, et al. “Identification of subtype-specific three-gene signature for prognostic prediction in diffuse type gastric cancer”. In: *Frontiers in oncology* 9 (2019), p. 1243.
-

-
- [238] T. Bonome, D. A. Levine, J. Shih, M. Randonovich, C. A. Pise-Masison, F. Bogomolnii, L. Ozbun, J. Brady, J. C. Barrett, J. Boyd, et al. "A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer". In: *Cancer research* 68.13 (2008), pp. 5478–5486.
- [239] F. Cardoso, L. J. van't Veer, J. Bogaerts, L. Slaets, G. Viale, S. Delaloge, J.-Y. Pierga, E. Brain, S. Causeret, M. DeLorenzi, et al. "70-gene signature as an aid to treatment decisions in early-stage breast cancer". In: *New England Journal of Medicine* 375.8 (2016), pp. 717–729.
- [240] D. Pratt, J. Chen, D. Welker, R. Rivas, R. Pillich, V. Rynkov, K. Ono, C. Miello, L. Hicks, S. Szalma, et al. "NDEx, the network data exchange". In: *Cell systems* 1.4 (2015), pp. 302–305.
- [241] D. Pratt, J. Chen, R. Pillich, V. Rynkov, A. Gary, B. Demchak, and T. Ideker. "NDEx 2.0: a clearinghouse for research on cancer pathways". In: *Cancer research* 77.21 (2017), e58–e61.
- [242] R. T. Pillich, J. Chen, V. Rynkov, D. Welker, and D. Pratt. "NDEx: a community resource for sharing and publishing of biological networks". In: *Protein Bioinformatics*. Springer, 2017, pp. 271–301.
- [243] U. D. of Health and H. Service. *The international classification of diseases*. World Health Organization, 1980.
- [244] J. S. Smolen, D. Aletaha, M. Koeller, M. H. Weisman, and P. Emery. "New therapies for treatment of rheumatoid arthritis". In: *The Lancet* 370.9602 (2007), pp. 1861–1874.
- [245] Z. Wang, H.-w. Bao, and Y. Ji. "A systematic review and meta-analysis of rituximab combined with methotrexate versus methotrexate alone in the treatment of rheumatoid arthritis". In: *Medicine* 99.8 (2020).
- [246] K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, et al. "A reference map of the human binary protein interactome". In: *Nature* 580.7803 (2020), pp. 402–408.
- [247] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. W. Senior, T. Green, A. ŽilLOOKdek, R. Bates, S. Blackwell, J. Yim, et al. "Protein complex prediction with AlphaFold-Multimer". In: *bioRxiv* (2021).
- [248] F. Prinz, T. Schlange, and K. Asadullah. "Believe it or not: how much can we rely on published data on potential drug targets?" In: *Nature reviews Drug discovery* 10.9 (2011), pp. 712–712.
- [249] M. List, P. Ebert, and F. Albrecht. *Ten simple rules for developing usable software in computational biology*. 2017.
- [250] J. Matschinske, N. Alcaraz, A. Benis, M. Golebiewski, D. G. Grimm, L. Heumos, T. Kacprowski, O. Lazareva, M. List, Z. Louadi, et al. "The AIME registry for artificial intelligence in biomedical research". In: *Nature methods* 18.10 (2021), pp. 1128–1131.
-

-
- [251] S. Lloyd. "Least squares quantization in pcm. Information Theory". In: *IEEE Transactions* ().
- [252] M. Hahsler, M. Piekenbrock, and D. Doran. "dbscan: Fast Density-Based Clustering with R". In: *Journal of Statistical Software* 91.1 (2019), pp. 1–30. DOI: 10.18637/jss.v091.i01.
- [253] L. Arend, J. Bernett, Q. Manz, M. Klug, O. Lazareva, J. Baumbach, D. Bongiovanni, and M. List. "A systematic comparison of novel and existing differential analysis methods for CyTOF data". In: *Briefings in Bioinformatics* (2021).
- [254] M. Klug, K. Kirmes, J. Han, O. Lazareva, M. Rosenbaum, G. Viggiani, M. von Scheidt, J. Ruland, J. Baumbach, G. Condorelli, et al. "Mass cytometry of platelet-rich plasma: a new approach to analyze platelet surface expression and reactivity". In: *Platelets* (2021), pp. 1–8.
- [255] L. Deng. "The mnist database of handwritten digit images for machine learning research [best of the web]". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.
- [256] T. Peng, L. Ma, X. Feng, J. Tao, M. Nan, Y. Liu, J. Li, L. Shen, X. Wu, R. Yu, et al. "Genomic and transcriptomic analyses reveal adaptation mechanisms of an *Acidithiobacillus ferrivorans* strain YL15 to alpine acid mine drainage". In: *PloS one* 12.5 (2017), e0178008.
- [257] Y. Shen, X. Wang, Y. Jin, J. Lu, G. Qiu, and X. Wen. "Differentially expressed genes and interacting pathways in bladder cancer revealed by bioinformatic analysis". In: *Molecular medicine reports* 10.4 (2014), pp. 1746–1752.
- [258] S. Udhaya Kumar, D. Thirumal Kumar, R. Bithia, S. Sankar, R. Magesh, M. Sidenna, C. George Priya Doss, and H. Zayed. "Analysis of differentially expressed genes and molecular pathways in familial hypercholesterolemia involved in atherosclerosis: a systematic and bioinformatics approach". In: *Frontiers in Genetics* 11 (2020), p. 734.
- [259] Y. Liu and M. R. Chance. "Pathway analyses and understanding disease associations". In: *Current genetic medicine reports* 1.4 (2013), pp. 230–238.
- [260] H. Han. "Analyzing support vector machine overfitting on microarray data". In: *International Conference on Intelligent Computing*. Springer. 2014, pp. 148–156.
- [261] J.-z. Xu and C.-w. Wong. "Hunting for robust gene signature from cancer profiling data: sources of variability, different interpretations, and recent methodological developments". In: *Cancer letters* 296.1 (2010), pp. 9–16.
- [262] L. Rappez, M. Stadler, S. Triana, R. M. Gathungu, K. Ovchinnikova, P. Phapale, M. Heikenwalder, and T. Alexandrov. "SpaceM reveals metabolic states of single cells". In: *Nature Methods* 18.7 (2021), pp. 799–805.
- [263] D. German Cancer Research Center (Deutsches Krebsforschungszentrum. "SpaceM reveals metabolic states of single cells". In: *ScienceDaily* (2021). URL: www.sciencedaily.com/releases/2021/07/210706101952.html.
-