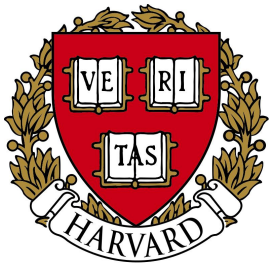

Biologically informed matrix factorization for joint inference of gene regulatory networks and transcription factor activities

Master's Thesis

Soel Micheletti

21st April 2023

Supervisor: Prof. Dr. J. Vogt
External supervisor: Prof. Dr. J. Quackenbush
Advisors: Dr. A. Marx, Dr. P. Mandros, Dr. J. Fischer
Department of Biostatistics, Harvard T.H. Chan School of Public Health
Department of Computer Science, ETH Zürich

A mia madre. A lei devo tutto.

Abstract

Since the initial sequencing of the human genome in 2001, the field of systems biology has strived to understand gene regulatory mechanisms, in particular the role of transcription factor binding. This is crucial to inform our understanding of diseases such as cancer: regulatory mechanisms are the primary alterations observed in tumors, and they are extensively investigated for treatment and drug design, as well as early detection. Despite the importance of discovering the underlying regulatory mechanisms, the task of inferring them from data while maintaining accuracy, explainability, scalability, and flexibility remains a significant open challenge.

This thesis proposes GIRAFFE, a scalable matrix factorization-based algorithm to jointly infer regulatory effects and transcription factor activities from gene expression data. GIRAFFE integrates prior knowledge about regulation to guide the optimization, yielding an interpretable model where regulatory weights are partial effects. Moreover, it can be customized to the requirements of the downstream application by adjusting for variables of interest, such as confounders, and adding sparsity constraints, which help to interpret the regulatory network.

We demonstrate the effectiveness of this approach with extensive experiments on synthetic, as well as real world data. Our algorithm outperforms state-of-the-art gene regulatory network inference methods in predicting interactions between transcription factors and target genes. Moreover, it is able to distinguish between activating and inhibitory effects, yielding plausible results in downstream applications such as gene set enrichment analysis.

Acknowledgements

First and foremost, I would like to deeply thank my advisor Alexander Marx, Professor John Quackenbush, and Professor Julia Vogt for making this exchange possible: the time I spent in Boston has been one of the most rewarding and fulfilling experiences of my life. Likewise, I am very grateful to Jonas and Panos for their invaluable guidance throughout this project, their wizardry in finding the best references and, of course, for our big arcade competitions. A huge thanks to everyone in the JQ lab for warmly welcoming me in the group, in particular to Viola for her precious advices and, most importantly, for always pointing me out the best free food spots.

I am grateful to my fellow students for sharing the challenges and fun of the ETH journey. To Enrico, for our continuous collaborations and (so far) vain attempts to bench 100kg. To Flavio, for introducing me to many cool ML ideas and telling me off whenever I was tempted to use nested for-loops. To Simone R., for being my continuous source of encouragement over the years.

I am grateful to all the people who are enriching my life. To Gabriele, for our lionhearted adventures in the Brooklyn's suburbs. To Sissy, for our *Bellevue* picnics and for always checking in on me. To Simone F., for the great times at *La Valascia*. To Jacopo, for always having the right word. To Sean and Mattia, for our Champions League nights. To Andrea, for our unhealthy addiction to Chickeria. To everyone else I am forgetting to mention.

I am grateful to the Swiss Study Foundation for providing me with precious networking and learning opportunities during my studies, and for their support during my stay in the US.

Last but not least, I want to thank Stephen Curry: the long coding nights would have not been the same without watching you hitting deep threes in every arena across the country. You are an incredible source of inspiration.

Contents

Contents **iv**

1 Introduction **1**
1.1 Research questions . . . 3
1.2 Contributions . . . 3
1.3 Structure of the thesis . . . 4

2 Background and motivation **5**
2.1 Gene regulation . . . 5
2.2 Biological networks . . . 6
2.3 Data . . . 8

3 Related Work **10**

4 Our Method **13**
4.1 Model . . . 13
4.2 Matrix factorization . . . 16
4.3 Controlling network sparsity and adjusting . . . 18
4.4 GIRAFFE in Python . . . 20

5 Experiments **21**
5.1 Simulation I . . . 23
5.1.1 Results . . . 24
5.2 Simulation II . . . 28
5.2.1 Results . . . 29
5.3 Validation of regulatory effects and TFA on yeast . . . 32
5.4 ChIP-seq validation on human data . . . 33
5.5 Oncogenes and tumor suppressor genes[1] . . . 35

[1]The authors thank Viola Fanfani for the thoughtful critiques and helpful review of this section.

5.6 Sex differences in lung adenocarcinoma 36

6 Discussion **41**

7 Concluding thoughts **45**

A Sparse optimization **46**
A.1 Proximal Adam 46
A.2 Tuning of the regularization parameter 48

B Supplementary materials **51**
B.1 Transcription factor expression is not a reliable surrogate for its activity 51
B.2 Addendum to Section 5.1 53
B.3 Addendum to Section 5.2 57
B.4 Addendum to Section 5.4 60
B.5 Addendum to Section 5.6 61

Bibliography **64**

Chapter 1

Introduction

Cancer is a leading cause of premature death, with a huge burden in every country of the world [Bray et al., 2018]. In 2020, there were an estimated 19.3 million new cancer patients, and almost 10 millions deaths worldwide [Ferlay et al., 2021].

After the initial sequencing of the human genome [US DOE Joint Genome Institute: Hawkins Trevor et al., 2001], data driven discovery has been a key tool to better understand factors driving the development and progression of cancer. Many studies have been conducted to understand how the genomic characteristic of a tumor, its mutational background, and the patterns of expressed genes can help alleviating the impact of cancer [Goldman et al., 2020]. Successful examples of translating this knowledge into therapeutics and diagnostics reinforce the potential of this approach, with the final goal of making personalized cancer medicine possible [Chin et al., 2011].

A major factor for this success has been the development of high-throughput sequencing technologies, enabling the collection of massive amount of data containing a wealth of information about the disease. Extracting this information from data leads to important insights for earlier detection and more effective treatment, both of which are crucial steps towards improving both the quality and quantity of cancer patients' lives [Paraskevi, 2012].

In this thesis we focus on data-driven approaches to infer gene regulation, a complex mechanism with many components. We focus on one of its key factors, where a particular type of regulators called transcription factors activate or inhibit the expression of their target genes. From a data science perspective, the gene regulatory machinery can be represented as a network, where nodes represent genes and transcription factors, while edges describe their relationships. Our goal is inferring the edges in the network, i.e. the (non-)existence of interactions between regulators and genes, and their enhancing or inhibitory nature.

Despite the plethora of available data, the task of inferring regulatory interactions remains challenging due to the high dimensionality [1] of the problem and the noise present in the data. While estimation methods yielding arguably accurate results exist, they suffer from one or multiple issues when it comes to interpretability, scalability to the human genome, and flexibility. For instance, state-of-the-art algorithms either do not distinguish between enhancing and inhibitory regulation, or do not scale beyond a few hundreds genes, therewith being inapplicable to the human genome. Moreover, not all existing methods put emphasis on interpretability, for example by incorporating complex non-linear relationships that optimize predictions, but that are incompatible with human reasoning, which is essential to ensure safety, ethics, and accountability of models supporting oncology decisions [Lu et al., 2023].

To overcome these issues, we propose GIRAFFE, a machine learning algorithm to jointly infer a gene regulatory network describing transcription factor-gene relationships, and a transcription factor availability matrix describing a sample-specific quantity of transcription factors available to regulate their targets. We aim for a model that is able to capture the relevant biological details while still being simple enough to ensure a reasonable interpretation. We, hence, use a matrix factorization that decomposes the observed gene expression as the product of transcription factor activity and regulatory network, leading to regulatory weights that can be interpreted as partial effects of a linear regression model. In particular, activating/inhibitory regulatory effects correspond to a positive/negative sign in the inferred regulatory network. Both the regulatory network and the transcription factor activity matrix can be estimated with off-the-shelf optimizers, helping GIRAFFE to efficiently scale to the entire human genome. To guide the optimization of our problem, we integrate prior knowledge about regulation. Including prior knowledge into the algorithm's behaviour has been shown to be beneficial both in theory and in practice [Greenfield et al., 2013, Wolpert and Macready, 1997], and it helps to overcome the problem of underdetermination, which is typical for biological applications of matrix factorization. Finally, GIRAFFE is flexible as it can be customized to the requirements of a concrete application: the user can optionally adjust for variables of interest and/or make the inferred regulatory matrix sparse.

To evaluate GIRAFFE's accuracy, we compare its performance to competing methods on both synthetic and real world datasets. To investigate its interpretability beyond the regulatory weights being partial effects, we check the consistency of the obtained sign, representing activating/inhibitory interactions, against well-defined ground-truths in simulated data and in biologi-

[1] A high-dimensional dataset is one where the number of features is much larger than the number of samples. In our case, we typically work with with 20 to 30 thousand genes (features), and only a few hundreds samples.

cal applications. Moreover, we assess the plausibility of obtained results for lung adenocarcinoma in graph differential analysis.

1.1 Research questions

We investigate if it is possible to improve the inference of gene regulatory networks with respect to

(i) Interpretability, for instance by distinguishing enhancing from inhibitory regulatory effects,

(ii) Scalability, by efficiently scale up to the human genome,

(iii) Flexibility, by adjusting for variables such as confounders,

(iv) Accuracy.

Importantly, we want to propose solutions that are not only algorithmically sound, but that also lead to valuable insights in a biological context.

1.2 Contributions

- We propose GIRAFFE (**G**ene-level **I**nference of **R**egulatory effects **A**s **F**actorizations of **F**unctions of **E**xpressions), a novel algorithm to estimate gene regulatory networks through a biologically informed matrix factorization.
- We design extensive experiments in silico, yeast, and human datasets to investigate GIRAFFE's performance w.r.t. the goals (i)-(iv) outlined in Section 1.1, demonstrating the superiority of our method over others.
- Both our code and processed data are publicly available under MIT license[2], together with extensive documentation to simplify the use for researchers without a computational biology background. We hope that this will foster future work in the field.
- We open source two additional libraries that might be interesting for the reader: `grn-thresholding`[3] to sparsify dense networks, and `grn-stability-selection`[4] to apply feature selection in gene regulatory networks while controlling for false discovery rates with high probability.

[2] https://github.com/soelmicheletti/giraffe
[3] https://github.com/soelmicheletti/grn-thresholding
[4] https://github.com/soelmicheletti/grn-stability-selection

1.3 Structure of the thesis

The thesis is organised as follows:

- In Chapter 2, we provide the necessary background required to understand the rest of the thesis. First, we present a primer on the relevant biological concepts. Then, we discuss graphs and why they are an appropriate abstract data type to model gene regulatory mechanisms. Finally, we describe the data we use and how they are collected.
- In Chapter 3, we give a succint overview of the lines of research most relevant to ours.
- In Chapter 4, we present the details of our algorithm from a conceptual, computational, and optimization perspective. First, we present our linear model formulation. Then, we show how it can be framed as a matrix factorization problem and optimized with established methods. Finally, we discuss extensions to obtain sparse solutions and adjust for variables of interest.
- In Chapter 5, we investigate our research questions in the context of GIRAFFE. We benchmark on synthetic data where the ground-truth is well-defined, and on human data with a gold standard. Then, we evaluate the plausibility of conclusions obtained on established downstream applications.
- In Chapter 6, we critically discuss GIRAFFE's performance and its contributions to combining accuracy, interpretability, scalability, and flexibility in gene regulatory networks inference.

Finally, we summarise the thesis and dicuss the avenues for future research that this work opens.

Chapter 2

Background and motivation

In this chapter we present relevant background to better understand the remainder of the thesis. We provide a biological primer on gene regulation; an introduction to biological networks and why they are appropriate to model gene regulation; and an overview of the data we use and how they are collected. Every section is self contained and independent from the others, such that the reader should feel free to skip and solely focus on the specific topics of interest.

2.1 Gene regulation

Each cell of the human body stores the whole genome in its nucleus. The genome is the complete collection of heritable genetic information about the organism, and can be modelled as long nucleotide sequences of DNA composed by sequences of symbols from the four letter alphabet $\{T, C, G, A\}$. Despite the fact that all cells share the same genetic material, the functions of different types of cells can differ significantly: compare, for instance, nerve and blood cells. The reason for this is that different cells express different genes[1], and the set of expressed genes ultimately determines their behaviour. The mechanism turning genes on and off is called *gene regulation*, and is essential not just to distinguish tissues from each other, but also because it can make the difference between health and disease [Ballestar and Esteller, 2008]. We now present a simplified model of gene regulation that will be useful to better understand the remainder of the thesis, and we refer the curious reader to Latchman [2007] and Ptashne and Gann [2002] for detailed biological explanations.

[1]For the purpose of this thesis, genes are stretches of DNA encoding some functionality: either protein, or other classes of functional RNAs. Even if genes are incredibly important, it is interesting to note how only a tiny fraction of DNA (approximately 3-5%) codes for proteins.

A version of the so called *central dogma of molecular biology* states that DNA makes RNA, and RNA makes proteins. Even if the statement is known to be wrong, a notable exception being the reverse transcription of viral RNA, this simple principle helps us explaining the flow of genetic information within a cell. During the process of transcription, an enzyme called RNA-polymerase slides along the DNA, opens the double strand, and produces mRNA, a single stranded "copy" of a gene. mRNA then serves as blueprint for a complex process that synthesize a protein as its final product. Particularly relevant for our purposes is a special kind of proteins called transcription factors (sometimes abbreviated TFs) that contribute, possibly by forming higher-order protein complexes together with other TFs, to increase or decrease the rate of transcription of a gene. More specifically, they bind to a motif in the promoter region of the target gene, thereby facilitating or preventing the RNA-polymerase to transcribe the gene into mRNA.

Note that the model described above is a useful simplification, but it is not the full story. For example, it can happen that a protein is not produced even if the gene associated with it is highly expressed in mRNA, or that gene expression is not affected by transcription factors only, but also by epigenetics, methylation, and environmental factors. As common in many related studies, we make simplifying assumptions and focus on the interactions between genes and transcription factors: genes produce TFs, TFs affect gene expression such that new - possibly different - TFs are produced. This interaction iteratively continues over time, determining the evolution of the cell's behaviour. This is particularly relevant to study cancer, as this process is altered during its development and progression.

2.2 Biological networks

Networks are a powerful abstract data type, particularly convenient for the depiction of relationships between entities. In general, a network or graph $G = (V, E)$ consists of a finite set of nodes V and a set of edges $E \subseteq V \times V$ connecting pairs of vertices. Hundreds of interesting computational problems are couched in terms of networks, with applications ranging from economics to social sciences [Cormen et al., 2022]. Similarly, many biological systems can be visualized using networks: nodes represent biologically relevant elements, and edges describe their relationships.

To gain an intuition on why networks are a suitable tool to model the underlying biology, let's consider gene expression data as an example. A gene expression dataset contains quantitative information about the expression of G genes across n samples. Typically, two distinct datasets are collected: the first one for healthy subjects, the second one for subjects with a particular type of cancer. A toy example is shown in Figure 2.1, where for each

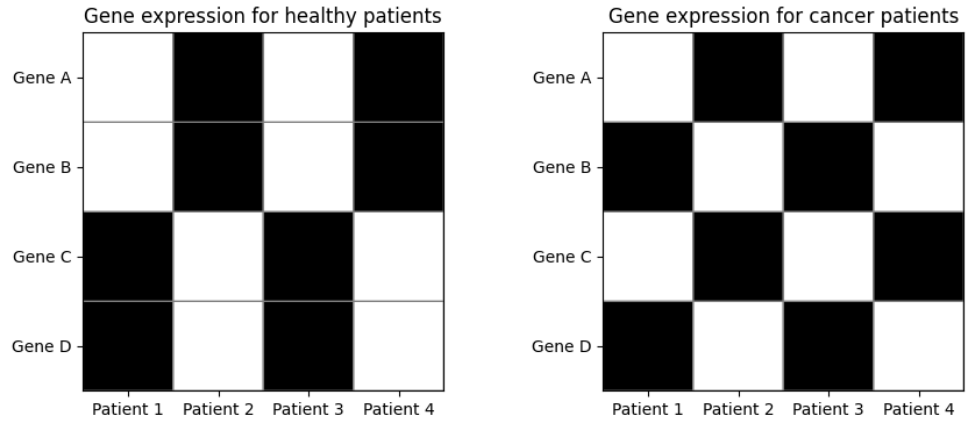

Figure 2.1: Two toy gene expression datasets (on the left for healthy patients, on the right for cancer patients). For the purpose of this example, a black entry means *highly expressed*, while a white entry means *lowly expressed*.

gene-patient pair a black value indicates high expression, while a white value indicates low expression. We see that there are differences between the datasets: for instance, *Gene A* has the same expression profile as *Gene B* on healthy patients, while on cancer patients is has the same expression profile as *Gene C*. However, a summary statistics computed directly from the gene expressions such as a gene-specific mean over the patients, is going to miss these differences: in both datasets, each gene is highly expressed in exactly two patients, and the mean expression for each dataset is the same for both phenotypes. These differences are better captured by a network. In this particular case, a gene co-expression network - where nodes represent genes, and edges a proper measure of co-expression between pairs of genes - can be very useful. In Figure 2.2 we show the corresponding co-expression networks for both datasets: the first one for healthy patients, the second one for cancer patients. We observe how, in contrast to simple statistics on gene expression, networks are able to capture structural differences in the data, with the potential of being more informative in biological applications such as cancer medicine.

In order to extract knowledge from biological networks, two fundamentals steps are necessary: *network inference* and *network analysis*. Networks inference refers to the task of recovering the unknown graph structure. Concretely, this means recovering the edges and their weights. Since this is the focus of this thesis, we do not discuss it further here. Network analysis, on the other hand, refers to the set of methods used to gain knowledge from the inferred networks. Classical approaches from the network analysis literature include hubs detection, modules identification, triad census, interpretation of hierarchical structures, centrality, and adherence to models such as core-

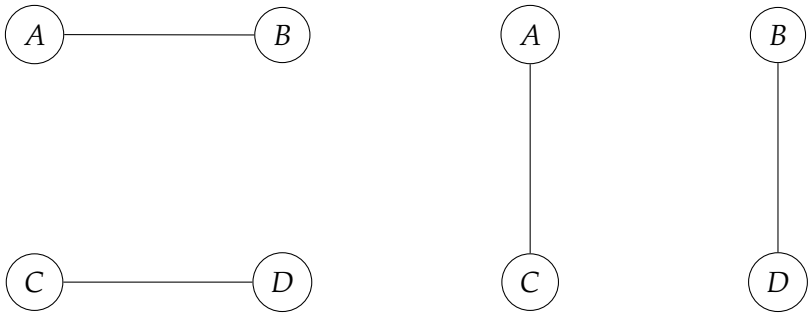

Figure 2.2: Network models for the gene expression datasets from Figure 2.1 (on the left the network for healthy patients, on the right the network for cancer patients). Edges represent co-expression between genes. For instance, since *Gene A* and *Gene B* shared the same expression profile on healthy patients, and hence are linked by an edge in the corresponding network.

periphery or threshold graphs [Brandes, 2005]. More modern comparative approaches, similar in spirit to the previous toy example, have been used to investigate evolutionary mechanisms [Crombach and Hogeweg, 2008], organ differentiation [Movahedi et al., 2011], and cancer [West et al., 2012, Gill et al., 2014, Li et al., 2016, Lopes-Ramos et al., 2020].

Network inference and analysis are mutually dependent: a great analysis is useless without a sufficiently good inference algorithm, and being able to recover the true structure is of limited utility if no valuable information can be extracted from it. Or, as a motivation for our work, higher-quality networks open up more opportunities to learn about the underlying biology.

2.3 Data

Our model integrates three types of data: gene expression, protein-protein interactions (sometimes abbreviated PPI), and a motif-based transcription factor-gene interaction prior. Moreover, we use ChIP-seq data as a gold standard in our validation experiments. In this section we aim to give a brief overview of these data and how they are collected.

Gene expression, measured individually for each sample across all genes and gathered in projects such as GTEx [Consortium et al., 2015] and TCGA [Gao et al., 2019], provides a quantitative profile about the information encoded in a gene that gets translated into functional products, such as mRNA. Originally measured with techniques based on nucleic acid hybridization such as microarrays [Schena et al., 1995] and SAGE [Velculescu et al., 1995], nowadays RNA-seq has settled as a standard, yielding higher throughput, better resolution, and lower noise [Wang et al., 2009]. In this work we use bulk RNA-seq, measuring the average across multiple cells within a tissue.

A PPI network is an undirected graph providing mechanistic insights into the interactions between transcription factors, for instance towards building higher-order protein complexes. Commonly used techniques, such as yeast two-hybrid [Van Criekinge and Beyaert, 1999] and affinity purification followed by mass spectrometry [Huttlin et al., 2021], apply markers in experimental setting that report detected interactions between pairs of proteins. Even it this does not necessarily yield perfectly accurate results, for instance because the experimental environment does not match the tissue under study, we incorporate known interactions from the String database [Szklarczyk et al., 2016] as prior knowledge.

Even if our algorithm is agnostic to the choice of the prior regulatory network, we often use a binary network with an edge between TF i and gene j if and only if the motif of TF i is detected in the promoter regions of gene j. These networks are built using the FIMO scanning tool [Grant et al., 2011] starting from sequences of promoter regions and a database of position weight matrices associating each TF to a set of motifs. While the presence of a transcription factor's motif in the promoter of a gene suggest that it is involved in its regulation, this is not necessarily true, as not all binding sites are active, and the binding of a single transcription factor might not be sufficient if the regulation happens cooperatively in a protein complex.

ChIP-seq is a more reliable technique. Instead of relying on the presence of a motif in the promoter region, it experimentally measures physical interactions between transcription factors and genes using chromatine immunoprecipitation [Park, 2009]. Whenever they are not used as validation, ChIP-seq data can be used as prior knowledge for our algorithm, as they provides more accurate information about the regulatory effects we aim to estimate.

Chapter 3

Related Work

Unveiling gene regulatory mechanisms is an essential task in systems biology, with the potential of informing our understanding of cellular processes and how they get altered by diseases such as cancer. Studied approaches to reverse engineer the structure of gene regulation include Boolean networks [Huang, 1999, Lim et al., 2016, Akutsu et al., 1999, Lähdesmäki et al., 2003], differential equation models [Sakamoto and Iba, 2001, De Hoon et al., 2002, Hossain et al., 2023], image-based methods [Puniyani and Xing, 2013, Wu et al., 2016, Yang et al., 2019], and deep learning architectures [Shrivastava et al., 2022, Kasabov, 2004, Chen et al., 2021, Kc et al., 2019]. In this chapter we aim to give a succinct overview of the lines of research most relevant to ours, and refer the interested reader to Mercatelli et al. [2020] for a comprehensive review.

Perhaps the most widely studied approach to understand gene regulatory mechanisms is to investigate the associations of genes through a gene co-expression network. It is defined as a network $C \in \mathbb{R}^{G \times G}$ representing undirected relationships between G genes of interest, and it's typically computed starting from a gene expression matrix measured from n samples. The interpretation of the entries in C heavily depends on the chosen measure of association between gene expressions. The simplest score that one may associate to a pair of vector-valued measurements is their correlation. WGCNA [Langfelder and Horvath, 2008], a widely established correlation-based method, is able to find modules of highly correlated genes. While being efficient to implement, correlation networks suffer from a major limitation: it is impossible to distinguish direct from indirect effects, and there is often a large number of false positives [Drakesmith et al., 2015]. Consider, for example, a situation where a transcription factor A regulates the expression of two otherwise independent genes B and C. In this case the correlation of B and C is non-zero, and hence the corresponding network will contain an edge between B and C. The relationship between B and

C, however, is only a consequence of their mutual relationship with the transcription factor A, and thus the inferred edge represents an indirect association. This phenomenon of two uncorrelated variables having a large correlation coefficient, sometimes summarised by the term *spurious correlation* [Aldrich, 1995], is often caused by confounding. A straight-forward approach to mitigate the issue of confounding is relying on Gaussian graphical models (GGMs), where edges represent partial correlations [Bühlmann and Van De Geer, 2011]. Intuitively, the partial correlation between two variables measures their correlation while accounting for the effects of the remaining variables in the data set. Assuming causal sufficiency, partial correlation is hence able to distinguish direct relationships between genes from those mediated by one or more transcription factors [Shutta et al., 2021]. While GGM are certainly a success story, their major drawback is assuming both normality of the data and linear relationships between genes. Information theoretic methods try to overcome this by using scores such as mutual information [Faith et al., 2007, Meyer et al., 2007]. While mutual information can accomomdate non-linear associations, it does not directly account for confounding and it can hence introduce spurious correlations. Mitigations based on the data processing inequality [Margolin et al., 2006], conditional mutual information [Liang and Wang, 2008, Aghdam et al., 2015, Zhang et al., 2012], or Markov blanket discovery [Liu et al., 2022, Ram and Chetty, 2009] have been explored. These methods might reduce the number of false positives, but their high computational burden prevents them to scale up to thousands of genes, therewith making them inapplicable to high-dimensional datasets such as the human genome [Sanguinetti et al., 2019]. More scalable approaches to incorporate non-linearity exploit (ensembles of) decision trees as feature selectors [Huynh-Thu et al., 2010, Petralia et al., 2015], support vector machines [Bruschi et al., 2022], and kernel based methods [Kontio et al., 2020, Iglesias-Martinez et al., 2021].

It is worth noting that regardless how they are computed, gene co-expression networks measure associations between genes, but do not necessarily imply causal relationships. In contrast, Gene Regulatory Networks (GRNs) have a higher potential when it comes to understand regulatory mechanisms. GRNs are bipartite graphs whose nodes are regulators and genes, and a directed regulator-gene edge indicates a direct relationship between them[1]. Thus, by definition, there is a causal relation from regulators to their target genes. Different approaches to construct GRNs have been proposed: MONSTER [Schlauch et al., 2017] is based on linear regression, TIGRESS [Haury et al., 2012] applies an ℓ_1-norm regularizer for feature selection, Wang et al. [2020] propose a solution based on Graph Neural Networks, and Patel and

[1]The term GRN has been used somehow ambiguously in the literature, sometimes even to refer to gene co-expression networks. For the scope of this thesis, we will stick to the bipartite definition.

Wang [2015] propose a semi-supervised learning framework generalizable to any classifier. All these algorithms consider transcription factors as regulators, and make the simplifying assumption that the gene expression of a transcription factor is a reliable surrogate for its activity. However, this is not true in general [Ma and Brent, 2021, Latchman, 1993]. Other methods avoid this issue by modeling regulation without explicitly considering transcription factor activity. PANDA [Glass et al., 2013] is an iterative algorithm inspired by message passing techniques that updates a protein-protein network, a co-expression network, and a regulation network until they reach an agreement. OTTER [Weighill et al., 2021] infers the GRN by solving a non-convex optimization problem, starting from the fundamental premise that the interactions between transcription factors, and the correlation between genes are noisy observations of the regulation matrix's projections. Both methods incorporate the same sources of prior knowledge - i.e. known protein-protein interactions and a motif-based prior - in similar ways, and they are the main inspiration for our work. We aim to improve their performance, as well as their interpretability, by additionally distinguishing enhancing form inhibitory regulation.

Recently, computational algorithms to infer both gene regulatory networks and transcription factor activity have emerged. In many cases, gene regulation and transcription factor activities are estimated using alternate optimization. Explored solutions include network component analysis [Fu et al., 2011], multi-task learning to combine heterogeneous datasets [Castro et al., 2019], bilinear models [Ma and Brent, 2021], and variational inference [Mahmood et al., 2022]. While these methods provide interesting insights on how to integrate prior knowledge, they have not been validated on human data. TIGER [Chen and Padi, 2022] and BITFAM [Gao et al., 2021] estimate regulation and transcription factor activity jointly using a Bayesian approach. The main goal of both methods is accurately estimating transcription factor activities, and they validate their results extensively in downstream applications. However, they have multiple common limitations. First, they both impose a Gaussian distribution for the posterior of the regulatory weights, which turns out to be a very strong assumption in many practical applications [Mar, 2019, Glass et al., 2013]. Second, they both neglect interactions between transcription factors. Incorporating known protein-protein interaction simplifies the inference process, and relying solely on expression and a regulatory prior potentially increases the risk of missing true regulatory interactions. Third, they are both computationally intense when applied to the human genome. Moreover, BITFAM does not distinguish between enhancing and inhibitory effects.

Chapter 4

Our Method

In this chapter we present GIRAFFE (**G**ene-level **I**nference of **R**egulatory effects **A**s **F**actorizations of **F**unctions of **E**xpressions), our algorithm to estimate GRNs. We aim to incorporate different sources of prior knowledge to jointly estimate a GRN and sample-specific transcription factor activities in an interpretable and flexible manner. As shown in Figure 4.1, we decompose the observed gene expression into latent factors represeting a GRN R, and a non-negative transcription factor activity matrix TFA. R contains information about intensity and nature of the relationships between TF and genes, while TFA describes the amount of proteins available to regulate their target genes. These latent factors are useful to inform our understanding of biology, but are expensive to measure with current technologies and are typically estimated computationally. In Section 4.1 we introduce the abstract model of the gene regulation machinery, in Section 4.2 we present our matrix factorization approach to identify a good such model, and in Section 4.3 we show two extensions to adjust for variables of interest and obtain sparse solutions.

4.1 Model

We model the gene regulation machinery as a graph $N = (V, E)$ consisting of a finite set of nodes $V = G \uplus TF$ and a set of edges $E \subseteq V \times V$. G is the set of genes, while TF is the set of transcription factors. We use $|G|$ and $|TF|$ to denote their cardinalities. There are two categories of edges: undirected edges connecting pairs of transcription factors, and directed edges connecting transcription factors (sources) to genes (targets). TF-TF edges represent interactions towards forming higher-order protein complexes, and are integrated as prior knowledge to learn the regulatory interactions between transcription factors and genes. An abstract representation of our model is depicted in Figure 4.2.

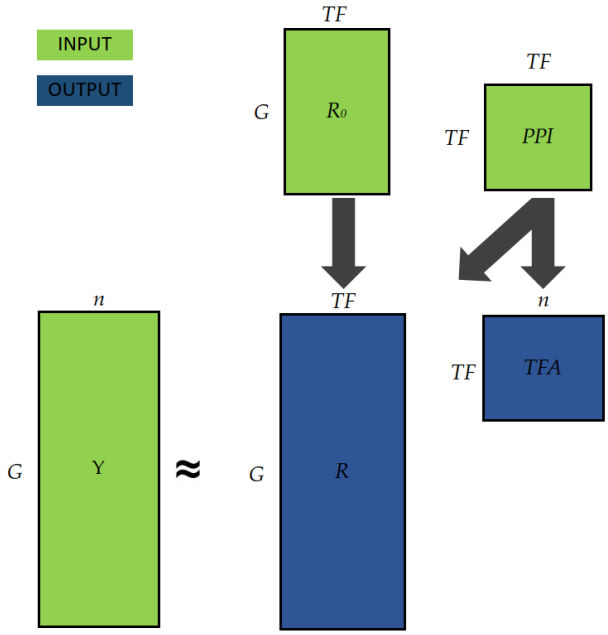

Figure 4.1: Schematic overview of GIRAFFE. Given gene expression, a prior R_0, and PPI, GIRAFFE computes a regulation matrix R and a transcription factor activity matrix TFA via a biologically informed matrix factorization.

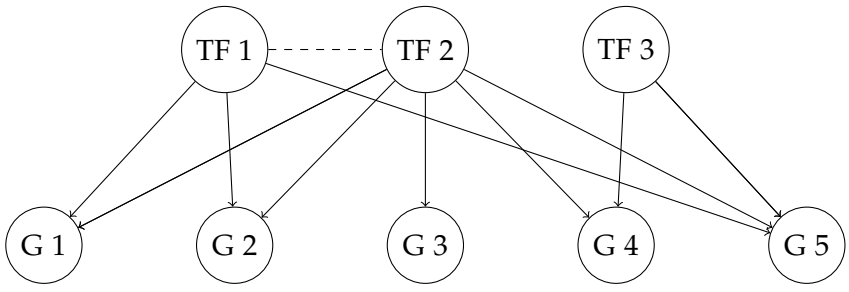

Figure 4.2: Abstract representation of the gene regulation machinery. The dashed edges between transcription factors represent protein protein interactions and are given as prior knowledge. We estimate the regulatory interactions as directed edges from transcription factors to genes.

We model the relationship between each gene and the transcription factor activities using a linear model:

$$Y_i = \sum_{k=1}^{|TF|} \beta_{i,k} \cdot TFA_k, \tag{4.1}$$

where $Y_i \in \mathbb{R}^n$ is the expression of gene i across n samples, and $TFA_k \in \mathbb{R}^n$ is the sample-specific activity of transcription factor k. Transcription factor activity is modeled for each sample because it incorporates the sample-specific environmental and epigenetic effects that influence transcription factor availability. Inferring the gene regulatory network is equivalent to inferring the coefficients $\beta_{i,k}$ for all $i \in G$ and $k \in TF$, which can be done by solving $|G|$ linear regressions "in parallel". Note that not only the regression coefficients must be inferred, but also the predictors TFA_k for $k \in [|TF|]$. The reason for this is that current technologies to measure transcription factor activity do not scale up to our setting , and using the mRNA expression of a transcription factor as a surrogate does not correctly model the complexity of the protein synthesis mechanism [Ma and Brent, 2021, Latchman, 1993][1]. In Figure 4.3 we show how the abstract representation of Figure 4.2 relates to our notation.

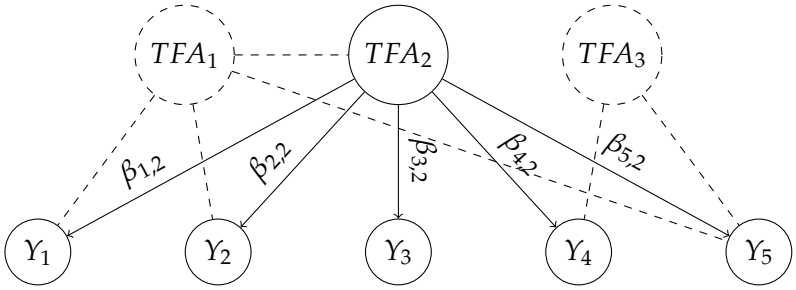

Figure 4.3: Link between the abstract representation of the model in Figure 4.2 and Equation 4.1. The expression of gene i, denoted $Y_i \in \mathbb{R}^n$, is modeled as a linear combination of the transcription factor activities of its regulators. $\beta_{i,k}$ denotes the contribution of TF k to the expression of gene i. The figure explicitly shows the notation for TF 2, which can be extended naturally for the dashed TFs.

From Equation 4.1, we observe that $\beta_{i,k}$ is the regression coefficient for the regulation between transcription factor k and gene i. For this reason, the weights computed by GIRAFFE can conveniently be interpreted as partial effects. We chose this design because it is able to correctly capture the emer-

[1]While this can be acknowledged as a known fact, in Appendix B.1 we present our own experiment to support this claim.

gent biological behavior while still being simple enough to ensure a reasonable interpretation. Importantly, it distinguishes activating regulation (positive regression coefficient) from inhibitory regulation (negative regression coefficient).

4.2 Matrix factorization

GIRAFFE jointly estimates a GRN and a transcription factor activity matrix through a biologically informed matrix factorization. More precisely, we decompose the gene expression matrix $Y \in \mathbb{R}^{|G| \times n}$ into the product of two matrices: a regulation matrix $R \in \mathbb{R}^{|G| \times |TF|}$, and a non-negative transcription factor activity matrix $TFA \in \mathbb{R}_+^{|TF| \times n}$. The setting is depicted in Figure 4.4.

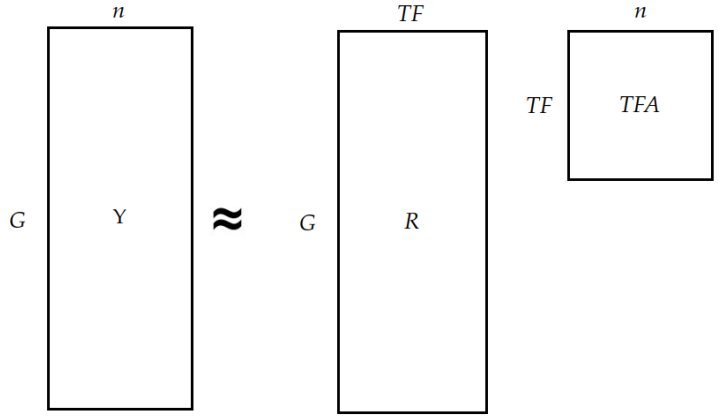

Figure 4.4: Visual representation of our matrix factorization. Given a gene expression matrix $Y \in \mathbb{R}^{|G| \times n}$, we decompose it as the matrix product between a regulation matrix $R \in \mathbb{R}^{|G| \times |TF|}$ and a non-negative transcription factor activity matrix $TFA \in \mathbb{R}_+^{|TF| \times n}$

In the example from Figure 4.3, the matrix formulation looks as follows:

$$\begin{bmatrix} | & | & & | \\ Y_1 & Y_2 & \dots & Y_5 \\ | & | & & | \end{bmatrix} \approx \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} \\ \beta_{4,1} & \beta_{4,2} & \beta_{4,3} \\ \beta_{5,1} & \beta_{5,2} & \beta_{5,3} \end{bmatrix} \cdot \begin{bmatrix} — & TFA_1 & — \\ — & TFA_2 & — \\ — & TFA_3 & — \end{bmatrix}.$$

In the example above, as well as in most practical settings, the number of transcription factors TF is larger than the number of samples n. This yield an underdetermined matrix factorization problem with an infinite number of solutions. For instance, one could set TFA to be a padded identity matrix,

and R to be a padded copy of the gene expression. This solution would exactly reconstruct the gene expression, but it would not correctly capture the underlying biology. As a consequence, we face a model selection problem and we rely on additional prior knowledge to make an informed decision.

GIRAFFE solves the problem by adding two sources of biological information to the matrix factorization: a prior for the regulatory networks $R_0 \in \mathbb{R}^{|G|\times|TF|}$, and a binary undirected graph $P \in \{0,1\}^{|TF|\times|TF|}$ representing known interactions between transcription factors. We are context agnostic for the prior, but in most cases we use a motif-based binary prior, where an edge connects a transcription factor to a gene if the sequence motif of the TF is present in the transcription factor binding site of the gene. More concretely, GIRAFFE solves the following optimization problem:

$$\begin{aligned}\underset{R\in\mathbb{R}^{|G|\times TF}, TFA\in\mathbb{R}_+^{|TF|\times n}}{\arg\min} \quad f(R,TFA) :=& \alpha||Y - R\cdot TFA||_F^2 \\ &+ \beta||R^T\cdot R - P||_F^2 \\ &+ \gamma||R\cdot R^T - C(Y)||_F^2 \\ &+ \delta||TFA\cdot TFA^T - P||_2 \\ &+ \lambda||R||_2^2,\end{aligned} \tag{4.2}$$

where $||\cdot||_F$ is the Frobenius norm, and $C(Y)$ is the correlation matrix of gene expression. $Y \approx R\cdot TFA$ minimizes the reconstruction error, $R^T\cdot R \approx P$ is an incentive for interacting proteins to target the same genes, $R\cdot R^T \approx C(Y)$ encourages correlated genes to have similar interactions with transcription factors, $TFA\cdot TFA^T \approx P$ yields more similar transcription factor activities for interacting proteins. The final term is used to regularize the solution by shrinking the regulation coefficients. R and TFA are estimated by minimizing $f(R,TFA)$. Since the problem is non-convex, we resort to gradient-based methods. Gradients can be found both analytically or numerically using a suitable tool. For all practical applications, we use the Adam optimizer [Kingma and Ba, 2014] implemented in Pytorch. The prior R_0 is used as initialization for R, while TFA is randomly initialized from $\mathcal{U}(0,1)$.

The objective function in 4.2 is a linear combination of different components, where the weight of each components $\alpha,\beta,\gamma,\delta$ should be picked according to two criteria: first, they shall moderate the influence of every component in a meaningful way, addressing the underlying trade-off; second, they shall satisfy all objects to a certain degree. The latter criterion is particularly challenging when the single components lie on different scales, e.g. when a single loss is of many order of magnitudes larger than the other ones. The

problem of balancing multiple loss functions that potentially lie on different scales is a well studied problem in the multi-task learning field, as many problems in engineering, natural sciences, and economics can be formulated analogously [Awad et al., 2015]. Most approaches for unsupervised learning problems can be classified into two broad categories: incorporating gradient statistics [Malkiel and Wolf, 2020, Chen et al., 2018], and loss rebalancing [Bischof and Kraus, 2021, Fernando and Tsokos, 2021, Liu et al., 2021]. Inspired by the insights of Lee and Kim [2020] in the context of monocular depth estimation, we suggest a rebalancing approach. More concretely, we set $\kappa = \frac{\sum_{\tau \in \{\alpha,\beta,\gamma,\delta\}} \tau \cdot \mathbb{I}[\tau \neq \kappa]}{\sum_{\tau \in \{\alpha,\beta,\gamma,\delta\}} \tau}$ for $\kappa \in \{\alpha, \beta, \gamma, \delta\}$. Finally, λ is left as a hyperparameter, with a default value of one. We show that this approach works well in a variety of scenarios, and hence we use it as default behaviour in our algorithm. In principle, however, GIRAFFE is agnostic to this choice. For this reason, we made our implementation fully customizable: the user can either provide a custom function to pick the weights at runtime, or even fixed scalar weights if desired.

4.3 Controlling network sparsity and adjusting

In this section, we present two additional features to fine tune GIRAFFE's predictions: sparsity, and adjusting for variables of interest.

A first observation is that optimizing GIRAFFE's objective as formulated in Equation 4.2 leads to a dense regulation matrix R. In particular, following our interpretation in terms of partial effects, all transcription factors are included in the set of parents of a target gene. This contrasts our understanding of the underlying biology, as we expect most transcription factors to target only a subset of genes [Wang et al., 2015]. While this is not an issue in many applications, where the identification of the regulators is done either implicitly [Lopes-Ramos et al., 2020, Van Dam et al., 2018] or in a post-processing step [Han and Zhu, 2008, Cassan et al., 2021], obtaining a sparse regulation matrix is desirable in applications where a correct identification of the regulators is essential. Moreover, sparse networks have advantages in terms of interpretability. To obtain a sparse regulatory network, we regularize $f(R, TFA)$ with the ℓ_1-norm by solving

$$g(R, TFA) := \underset{R \in \mathbb{R}^{|G| \times TF}, TFA \in \mathbb{R}_+^{|TF| \times n}}{\arg\min} f(R, TFA) + \lambda_1 ||R||_1, \tag{4.3}$$

which leads to a sparse regulation matrix because of the geometric properties of the ℓ_1-norm [Tibshirani, 1996]. Since use gradient-based methods, however, directly optimizing g using Adam does not automatically lead to

a sparse solution. Instead, we exploit the fact that the ℓ_1-norm is proximal friendly, and adapt ideas from proximal gradient descent [Chen et al., 2012] to our context [2]. See Appendix A.1 for a self-contained explanation of our solution. The tuning of the sparsity regularization parameter λ_1 is inherently challenging: on the one hand we want to select all true edges, on the other hand we do not want to include too many extra edges (false positives). We advocate a choice that is conservative enough to select all non-zero entries of R. This property is known as *variable screening*, and it has been shown to be achievable with high probability by the Lasso [Meinshausen and Bühlmann, 2006]. To control the number of false positives, Meinshausen and Bühlmann [2010] propose stability selection, a framework to select the most relevant edges. We provide the details in Appendix A.2.

The second additional feature allows the user to adjust for variables of interest. These include the patients' phenotypes that are believed to confound or bias the relationship between transcription factors and genes. To adjust for $Z \in \mathbb{R}^{k \times n}$, where k is the number of variables to be taken into account, we modify the objective in Equation 4.2 as follows.

$$\begin{aligned} \underset{R \in \mathbb{R}^{|G| \times TF}, TFA \in \mathbb{R}_{+}^{|TF| \times n}, \Theta \in \mathbb{R}^{|G| \times k}}{\arg\min} \quad f(R, TFA) := & \alpha ||Y - \begin{bmatrix} R & \Theta \end{bmatrix} \cdot \begin{bmatrix} TFA \\ Z \end{bmatrix} ||_F^2 \\ & + \beta ||R^T \cdot R - P||_F^2 \\ & + \gamma ||R \cdot R^T - C(Y)||_F^2 \\ & + \delta ||TFA \cdot TFA^T - P||_F^2 \\ & + \lambda || \begin{bmatrix} R & \Theta \end{bmatrix} ||_2^2. \end{aligned} \tag{4.4}$$

Note that we have to learn an additional parameter Θ that quantifies the partial effect of each variable on gene expression. Adjusting for variables can be naturally combined with Equation 4.3 to obtain sparse solutions.

[2] The proximal operator of a function is a mathematical operator that maps a point to its nearest point in the function's domain that minimizes the sum of the function and a weighted penalty term. When we say that the ℓ_1-norm is proximal friendly, we mean that its proximal operator can be computed efficiently.

4.4 GIRAFFE in Python

All described characteristics can be defined and customized using the following code snippet:

```
model = giraffe.Giraffe(
    expression,
    prior,
    ppi,
    adjusting = None,
    regularization = 0,
    iterations = 200,
    lr = 1e-5,
    lam = None,
    balance_fn = None,
    save_computation = False,
    seed = 42
)
```

Listing 4.1: Instance of a GIRAFFE's model in Python.

GIRAFFE is instantiated from an expression matrix Y, a prior R_0, and a PPI network P. The optional `adjusting` attribute is a real matrix of dimensionality $k \times n$, where k is the number of variables to be adjusted. The attribute `regularization` is the value of the λ_1 parameter in Equation 4.3, while `iterations` and `lr` define the number of Adam iterations and its learning rate. To customize the choice of the weights $\alpha, \beta, \gamma, \delta$ in Equation 4.2, `lam` and `balance_fn` can be used: `lam` is used to pick fixed scalar weights, while `balance_fn` is a user-defined function that is applied at every iteration to dynamically update the weights based on the loss values. When `save_computation` is set to `True`, the parameter γ in Equation 4.2 is set to zero. This speeds up computations on large datasets and is very useful for prototyping.

This wraps-up the presentation of GIRAFFE, our algorithm to jointly infer a GRN and a transcription factor activity matrix. We presented the model, our matrix factorization approach, and extensions to achieve sparsity and to adjust for variables of interest. In the remainder of the thesis, we evaluate its performance with extensive experiments in a variety of contexts.

Chapter 5

Experiments

In Chapter 4 we presented GIRAFFE, our algorithm to jointly infer a gene regulatory network and a transcription factor activity matrix. We now assess our method through extensive validation using both experimental and in silico data. We investigate the quality of the inferred transcription factor activity and regulatory matrix (in the latter case not only the correctness of the identified TF-gene interactions, but also their activating/inhibiting nature), the ability to recover sparse results, the convenience of adjusting for variables of interest, and the biological plausibility of results obtained by applying GIRAFFE in downstream applications. Table 5.1 provides an overview of our validation techniques, their goals, and the selected benchmarks.

	Quality of $\hat{R}$	**Enhancing vs inhibitory regulation**	**$T\hat{F}A$ validation**	**Sparsity**	**Adjusting**	**Downstream application**	**Benchmarks**
In silico experiment I, Section 5.1	AUROC with known ground-truth.	Accuracy with known ground-truth.	Gene expression reconstruction and identifiability.	AUROC between edge relevance and ground-truth.	AUROC with hidden variable, confounder, and causal sufficiency.		PANDA, OTTER, and prior.
In silico experiment II, Section 5.2	AUROC with known ground-truth.	AUROC with known ground-truth.	Relative error of gene expression reconstruction, absolute error with ground-truth.	AUROC between edge relevance and ground-truth.			PANDA, OTTER, and prior.
Yeast data, Section 5.3	AUROC with ChIP-seq data.		Ranks in interventional dataset.				PANDA, OTTER, and motif-based prior.
Human data, Section 5.4	AUROC with ChIP-seq data.				Adjusting for ischemic time.		PANDA, OTTER, TIGRESS, BITFAM, GENIE3, and motif-based prior.
Oncogenes and tumor suppressor genes, Section 5.5						Identification of important genes in cancer.	COSMIC.
Sex differences in lung adenocarcinoma, Section 5.6						Plausibility of findings in differential analysis.	Literature on sex differences in lung cancer, PANDA.

Table 5.1: Overview of used metrics to evaluate different aspects of GIRAFFE throughout our experiments.

5.1 Simulation I

To investigate the performance of GIRAFFE when our assumptions are met and ground-truth is well-defined, we create synthetic data. We consider $n = 50$ samples for $G = 500$ genes regulated by $TF = 100$ transcription factors, therewith being both time-efficient and realistic by matching the ordering of a typical human dataset. We need to generate the ground-truth regulation matrix $R \in \mathbb{R}^{G \times TF}$ and transcription factor activity $TFA \in \mathbb{R}^{TF \times n}$, as well as GIRAFFE's inputs PPI network $P \in \{0,1\}^{TF \times TF}$, gene expression $Y \in \mathbb{R}^{G \times n}$, and prior $R_0 \in \{0,1\}^{G \times TF}$. To generate Y, we need the ground-truth R and TFA, both of which are derived from P. The prior R_0 is obtained by corrupting R. Figure 5.1 shows our pipeline highlighting the relationships between the data.

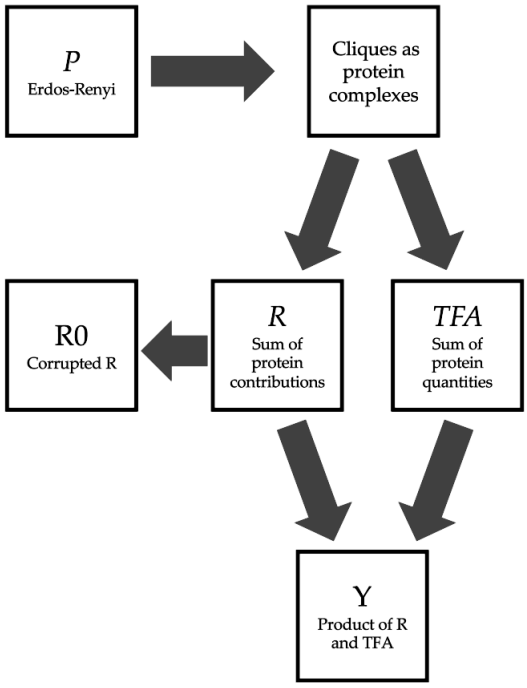

Figure 5.1: Generation pipeline for ground-truth and GIRAFFE's inputs.

The starting block of our pipeline is the PPI network P, that we generate using an Erdos-Rényi model with interaction probability estimated as the mean degree from the STRING database on human data [Szklarczyk et al., 2016]. This ensures that the generated PPI has a similar topology as in human datasets. To simulate the behaviour of transcription factors binding into higher-order protein complexes and jointly affecting the transcription rate of a gene, we compute cliques on the generated PPI, and assign a protein complex to each clique. If a protein has no neighbours in the PPI network,

we assign it to its own clique of size one.

Based on our matrix factorization model, we generate the gene expression $Y = R \cdot TFA$. To get R and TFA we rely on the protein complexes (or cliques) in P. The intuition is to first generate the regulatory interactions and activities for each protein complex, and then to extend them to the single transcription factors. More precisely, letting K be the set of cliques in P representing the protein complexes, we generate for each $k \in K$ a sample-specific activity $A_k \in \mathbb{R}^n$, and a regulatory vector $I_k \in \mathbb{R}^G$. The entries of A_k are i.i.d. samples from $\mathcal{U}(0,1)$, and I_k is a sparse vector, where the non-zero entries are drawn i.i.d. from $\mathcal{U}(-\frac{1}{2},\frac{1}{2})$. Sparsity is important because we assume that a protein complex regulates only a subset of the genes, and the vectors can have negative entries to incorporate potential inhibitory regulations. For each transcription factor $i \in [TF]$, we then generate the corresponding column/row in R and TFA by summing over the correct cliques as follows:

$$TFA_{i,:} = \sum_{k \in K} A_k^T \cdot \mathbb{I}\left[i \in k\right]$$
$$R_{:,i} = \sum_{k \in K} I_k \cdot \mathbb{I}\left[i \in k\right]$$

Finally, we compute the binary prior R_0 from the regulation matrix R in two steps: first, we map all non-zero entries of R to one, therewith getting a matrix describing which TF-gene pair have a regulatory relationship. Second, since motif-based TF binding scores are not a perfect proxy for regulation, we flip some entries of W.

5.1.1 Results

Now that we introduced the data generating process, we present our experiments and results.

Quality of inferred regulatory matrix To test the performance and robustness of GIRAFFE, we consider the Area Under the Receiver Operating Characteristics (AUROC) scores over a multitude of motif reliability settings, where the number of flipped entries is gradually increased. We average the results over $B = 50$ runs to show the stability of our conclusions, where the randomness originates from the data generating process.

AUROC	Method			
	GIRAFFE	OTTER	PANDA	Prior
Prior reliability: 99%	**0.997**±0.002	**0.997**±0.002	0.845±0.005	0.990±0.000
Prior reliability: 90%	**0.967**±0.001	0.960±0.001	0.845±0.001	0.900±0.000
Prior reliability: 80%	**0.930**±0.001	0.910±0.001	0.845±0.001	0.800±0.000
Prior reliability: 70%	**0.871**±0.001	0.839±0.001	0.844±0.001	0.700±0.000
Prior reliability: 60%	0.781±0.002	0.723±0.002	**0.843**±0.001	0.600±0.000
Prior reliability: 50%	**0.581**±0.004	0.466±0.005	0.500±0.030	0.500±0.000

Table 5.2: Comparison of AUROC score of the GRNs inferred by GIRAFFE, OTTER, PANDA, and the prior. The score are averaged over $B = 50$ runs, and we report the standard deviation.

From table 5.2 we clearly observe that GIRAFFE outperforms OTTER, PANDA, and the prior in terms of AUROC across most settings. The only exception is PANDA performing better (AUROC=0.843) when 40% of the prior matrix entries are flipped. Generally, the accuracy of the inferred GRN is larger for more accurate priors.

Distinguishing enhancing from inhibitory interactions To estimate the ability to distinguish between activating and inhibitory interactions, we compute the sign accuracy, defined as

$$\frac{1}{G \cdot TF} \sum_{i,j} \mathbb{I}\left[(\hat{R}_{i,j} \cdot R_{i,j}) = 1\right],$$

where $\hat{R}$ is the GRN inferred by GIRAFFE. Similarly as before, we repeat the experiment $B = 50$ times.

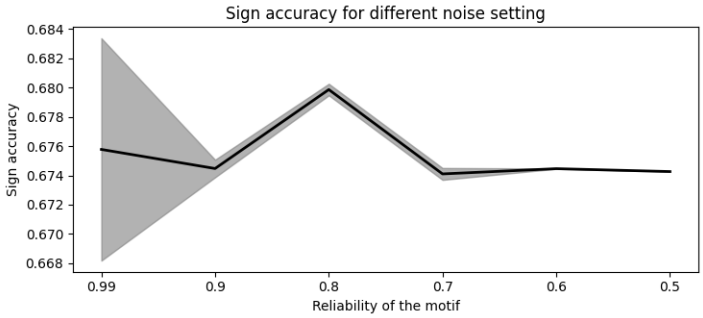

Figure 5.2: Sign accuracy of $\hat{R}$. The black line shows the mean over $B = 50$ runs, and the gray area is within one standard deviation from the mean.

Figure B.3 shows that, considering mean and standard deviation, the sign accuracy is consistently greater than 66%

Quality of inferred transcription factor activities To assess $T\hat{F}A$ we propose two metrics: the relative error of the gene expression reconstruction error,

$$\frac{\mathbb{E}\left[Y - \hat{Y}\right]}{\mathbb{E}\left[Y\right]}, \text{ with } \hat{Y} := \hat{R} \cdot T\hat{F}A,$$

and the relative value of the identifiability error

$$\frac{\mathbb{E}\left[TFA - T\hat{F}A\right]}{\mathbb{E}\left[TFA\right]}.$$

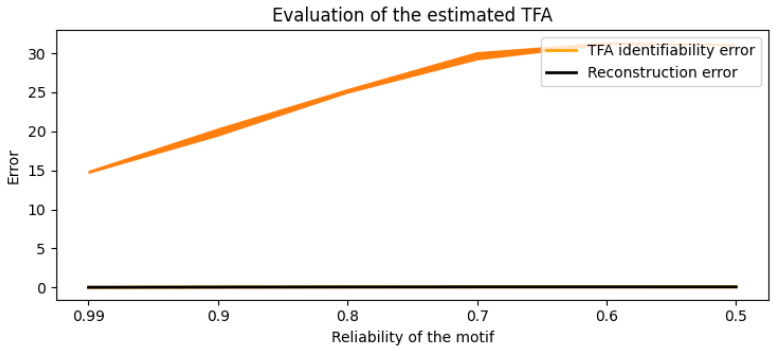

Figure 5.3: Gene expression reconstruction error and identifiability error for the estimated $T\hat{F}A$.

Figure 5.3 shows the trend of both metrics for different noise settings. We observe that the reconstruction error is much lower (around 1%) than the identifiability error (ranging from $15 - 30\%$). Moreover, the identifiability error drastically increases when the prior becomes less reliable.

Sparsity Figure 5.4 shows ROC curves for the accuracy of the estimated active set of R. The ROC curves are obtained by comparing the relevance assigned to each edge, evaluated as the number of times the edge was selected over $B = 1000$ runs with different values for the ℓ_1 regularization hyperparameter, with the ground-truth value of R, where non-zero edges are mapped to one. We observe that GIRAFFE's ability to recover the active set of edges heavily relies on the quality of the prior, ranging from an AUROC of 0.99 for an almost perfect prior, down to only slightly better than random for a prior where 50% of edges are corrupted.

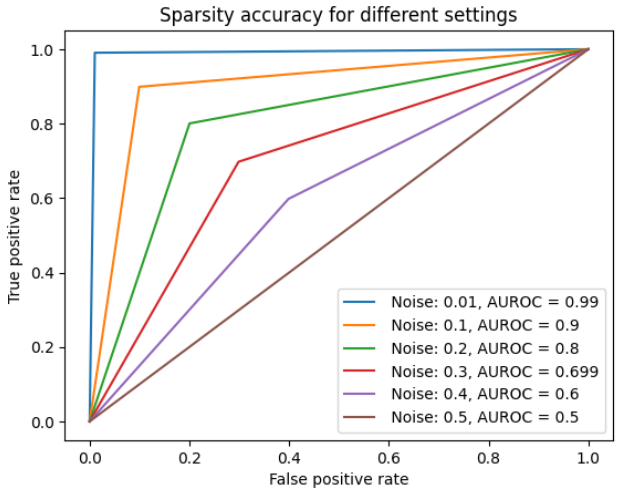

Figure 5.4: ROC curves for the accuracy of edges' relevance against the corresponding entry of R being non-zero. Edge relevance is computed by averaging how many times it was selected over $B = 1000$ runs with different values of λ_1 in Equation 4.3. $\hat{R}$ has been computed with a binary motif.

Adjusting Finally, we test GIRAFFE's feature to adjust for variables of interest. We hence modify the generating mechanism with a random Gaussian variable with mean zero and standard deviation one as follows: we add an additive biasing variable (Figure 5.5), a hidden confounder (Figure 5.6), and an hidden variable affecting the transcription factor activity only (Figure 5.7). For all settings, we compare GIRAFFE's vanilla version to both the correct model (without hidden variable) and GIRAFFE when adjusting for the hidden variable. Figures 5.9-5.11 show the corresponding ROC curves. We observe that adjusting for the hidden variable increases the score of GIRAFFE from 0.864 to 0.894 in case of a hidden variable, and from 0.873 to 0.890 for the confounding case. In both scenarios, the score gets closer to the score obtained with the correct model (0.906 in both cases). When we consider a generative model as in Figure 5.6, the scores do not change. Note that this last case corresponds to causal sufficiency, as we only perturbate the data generating process of the transcription factor activities.

Additional details and results to assess the robustness of our results can be found in Appendix B.2.

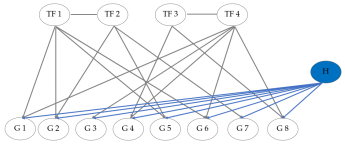

Figure 5.5: Hidden additive biasing variable.

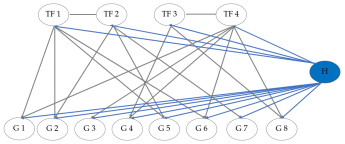

Figure 5.6: Hidden confounding variable

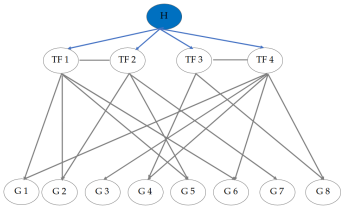

Figure 5.7: Hidden variable affecting transcription factor only.

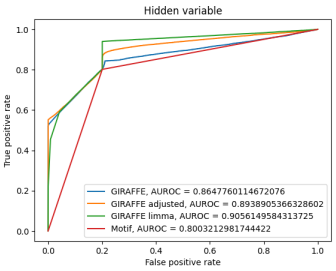

Figure 5.8: ROC curves for Figure 5.5.

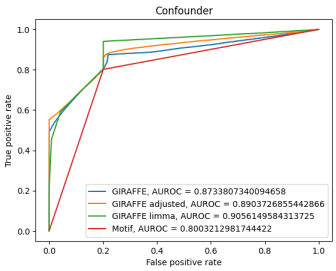

Figure 5.9: ROC curves for Figure 5.6

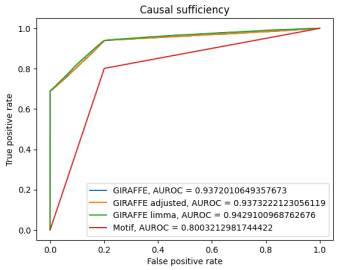

Figure 5.10: ROC curves for Figure 5.7.

5.2 Simulation II

To validate GIRAFFE's performance on datasets not explicitly designed to satisfy our modeling assumptions, we create a second synthetic dataset. Keeping the same dimensionality as in Section 5.1, we generate the matrices $R \in \mathbb{R}^{G \times TF}$, $P \in \mathbb{R}^{TF \times TF}$, $R_0 \in \mathbb{R}^{G \times TF}$, and $Y \in \mathbb{R}^{G \times n}$ for $n = 50$, $G = 500$, and $TF = 100$. As shown in Figure 5.11, we first generate the regulatory matrix R, from which we derive R_0, P, and Y.

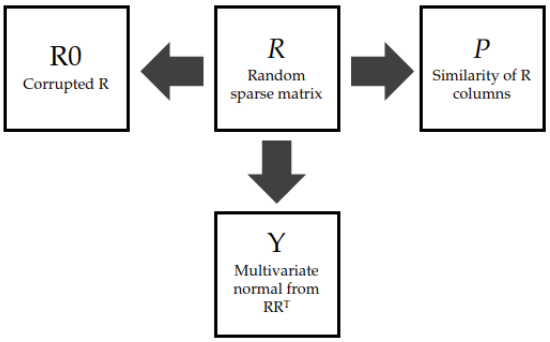

Figure 5.11: Generation pipeline for ground-truth and GIRAFFE's inputs.

Assuming that regulation is a noisy projection of the PPI and the gene expression matrix, we model R_0, P and Y from the randomly generated R. We do not model the TFA matrix. The ground-truth R is a sparse matrix, where we draw the non-zero entries i.i.d from $\mathcal{U}(-1,1)$. From R, the starting block of our pipeline, we generate P as a binary matrix, where the entry between a pair of TFs is set to one if and only if the cosine similarity between their regulation vectors is among the top 30% across the dataset (therewith maintaining a similar structure to a biological PPI); R_0 as a corrupted version of R, where we added zero-centered gaussian noise; and we sample Y from a multivariate normal distribution with mean zero and covariance matrix RR^T.

5.2.1 Results

Now that we introduced the data generating process, we present our experiments and results.

Quality of the inferred regulatory matrix In Table 5.3 we show AUROC scores as a mean to evaluate GIRAFFE's ability to recover R for different noise levels, and we compare it to PANDA, OTTER, and the prior R_0. All results are averaged over $B = 50$ runs, and we report mean and standard deviation of the results.

AUC	Method			
	GIRAFFE	OTTER	PANDA	Prior
$\sigma = 0.05$	0.929±0.002	**0.956**±0.001	0.690±0.002	**0.956**±0.001
$\sigma = 0.15$	**0.876**±0.002	0.870±0.003	0.648±0.003	0.869±0.003
$\sigma = 0.25$	**0.813**±0.002	0.788±0.003	0.611±0.003	0.788±0.003
$\sigma = 0.35$	**0.749**±0.003	0.713±0.003	0.579±0.004	0.712±0.003
$\sigma = 0.45$	**0.696**±0.002	0.658±0.003	0.557±0.004	0.658±0.003
$\sigma = 0.55$	**0.655**±0.003	0.618±0.004	0.543±0.005	0.618±0.004
$\sigma = 0.65$	**0.622**±0.005	0.591±0.005	0.532±0.004	0.590±0.005
$\sigma = 0.75$	**0.599**±0.004	0.571±0.004	0.525±0.004	0.570±0.004
$\sigma = 0.85$	**0.581**±0.005	0.559±0.004	0.521±0.005	0.558±0.004
$\sigma = 1$	**0.569**±0.006	0.547±0.004	0.516±0.004	0.547±0.004

Table 5.3: Comparison of AUC-ROC score of the GRNs inferred by GIRAFFE, OTTER, PANDA, and the prior. The score are averaged over $B = 50$ runs, and we report the standard deviation.

Similarly as in Section 5.1, the AUROC decreases when the prior becomes less reliable. Apart from the scenario with almost perfect prior knowledge, GIRAFFE outperforms the other methods.

Distinguishing enhancing from inhibitory interactions To estimate the ability to distinguish between enhancing and inhibitory interactions, we rely on the sign accuracy

$$\frac{1}{G \cdot TF} \sum_{i,j} \mathbb{I}\left[(\hat{R}_{i,j} \cdot R_{i,j}) = 1\right],$$

where $\hat{R}$ is the GRN inferred by GIRAFFE. Similarly as before, we repeat the experiment $B = 50$ times.

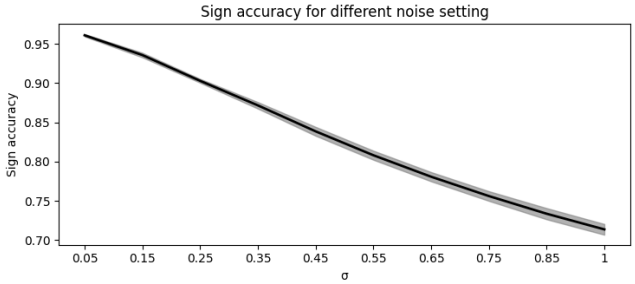

Figure 5.12: Sign accuracy of $\hat{R}$. The black line shows the mean over $B = 50$ runs, and the gray area is within one standard deviation from the mean.

Figure 5.12 shows sign accuracy of our GRNs as a proxy to evaluate GI-RAFFE's ability to distinguish enhancing from inhibitory regulation. It decreases proportionally to the reliability of the prior, and it is lower bounded by an accuracy of 0.72 with our noise settings.

Sparsity Similarly as in Section 5.1, Figure 5.13 shows the ROC curves for the accuracy of the estimated active set of R, obtained by comparing the edge relevance computed by bootstrapping with the ground-truth value of R, where non-zero edges are mapped to one. The results are much more robust than in Figure 5.4, indicating that shuffling the prior network affects the AUROC scores much more than corrupting it with additive noise.

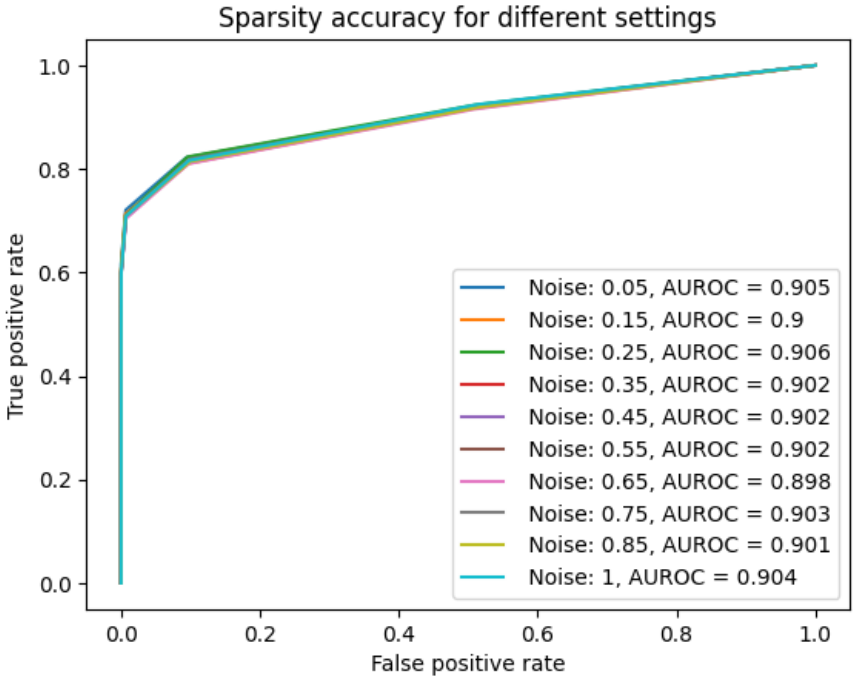

Figure 5.13: AUROC between the edge relevance and an indicator variable for the corresponding entry in the ground-truth being zero. Edge relevance is computed by averaging how many times it was selected over $B = 1000$ runs with different values of λ_1 in Equation 4.3.

Individual ROC curves for the performance estimation and additional results to assess the robustness of our results can be found in Appendix B.3.

5.3 Validation of regulatory effects and TFA on yeast

As a first validation on biological data, we use an experimental dataset for cell-cycle synchronized Baker's yeast cells (*Saccharomyces cerevisiae*). The gene expression values consist of two replicates of measurements for 3551 genes across 24 time points in the yeast cell-cycle time course [Pramila et al., 2006]. These data are combined with the presence of the motif in the promoter region of the genes for 105 transcription factors [Harbison et al., 2004], and we extract the interactions between those transcription factors from the String Database [Szklarczyk et al., 2016]. To assess GIRAFFE's ability to recover the regulation matrix R, we compute the AUROC between the estimated GRN $\hat{R}$ and a gold standard identified from ChIP-seq experiments [Harbison et al., 2004]. ChIP-seq exploits chromatin immunoprecipitation to determine *in vitro* which DNA fragments are enriched for a given protein. After applying a thresholding ($p < 0.001$), the procedure yields genomic fragments that are physically bound by a transcription factor [Park, 2009]. In Figure 5.14, we compare GIRAFFE's performance against three baselines, were the AUC considers only the subset of edges contained in our gold standard.

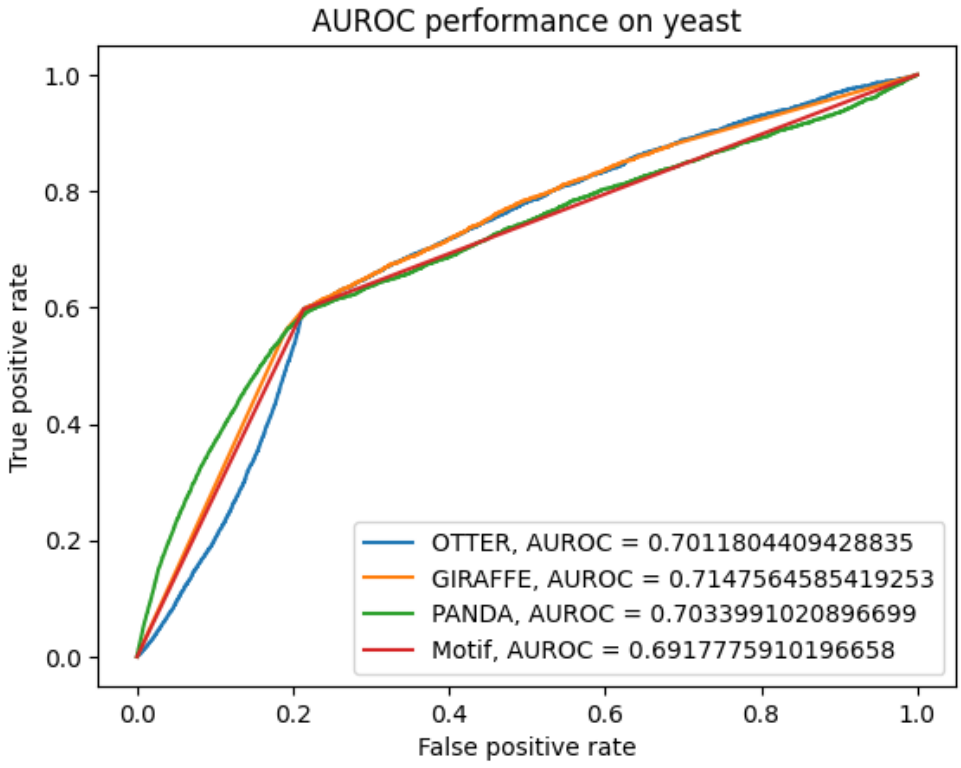

Figure 5.14: AUROC between estimated regulations and ChIP-seq data as a "gold standard".

We observe that all three methods improve over the prior motif (AUROC=0.692). OTTER (AUROC=0.701) and PANDA (AUROC=0.703) perform similarly, while GIRAFFE has a higher score (AUROC=0.715) and is the only method that consistently outperforms the motif's ROC curve.

To validate the inferred transcription factor activities, we consider a second dataset, where each sample collects the gene expression after a transcription factor has been knocked-out. GIRAFFE correctly identifies 31 out of 50 transcription factors as having the lowest transcription factor activity in the corresponding knock-out sample. The mean rank over all 50 interventions - where the lowest transcription factor activity is assigned rank 0 - is 8.17.

5.4 ChIP-seq validation on human data

In this section we validate GIRAFFE's performance on human data, using PANDA [Glass et al., 2013], OTTER [Weighill et al., 2021], GENIE3 [Huynh-Thu et al., 2010], TIGRESS [Haury et al., 2012], BITFAM [Gao et al., 2021], and a motif-based binary prior for TF-gene binding from CIS-BP [Weirauch et al., 2014] as benchmark. For all tissues, we download gene expression data from GRAND [Ben Guebila et al., 2022], and we integrate PPI downloaded from the STRING database [Szklarczyk et al., 2016]. As a gold standard, we use experimentally-defined TF-genes interactions using ChIP-seq data from hTFtarget [Zhang et al., 2020] and, in the case of kidney, from ReMap [Chèneby et al., 2018].

Due to computational requirements, we omit other algorithms from our benchmarks, such as ARACNE [Margolin et al., 2006], GLasso [Friedman et al., 2008], and TIGER [Chen and Padi, 2022]. Furthermore, in all our experiments, we have to reduce the default number of trees fitted by GENIE3 from 100 to 5, and run BITFAM on batches of genes to make them scale up to the human genome on a CPU machine with 16GB RAM . Table 5.4 reports the AUROC between the estimated regulation and the gold standard for five different tissues, and we observe that GIRAFFE has better scores in all of them. The corresponding plots are available in Figure B.9 in the appendix.

GRAND does not only collect gene expression data, but also metadata for each patient. We investigate the impact of ischemic time [1] on regulation in lung. Figure 5.15 shows that the two-dimensional embedding of gene expression is correlated with ischemic time, indicating that this is possibly a confounding or biasing variable in our regression model. Hence, we adjust for it. As a comparison, we preprocess the gene expression matrix using

[1] Sample-specific ischemic time is defined as the time from death or withdrawal of life-support until the time the sample is placed in a fixative solution or frozen [Consortium et al., 2015].

Method	Tissue					
	Breast	Colon	Kidney	Liver	Lung	Prostate
GIRAFFE	**0.697**	**0.651**	**0.565**	**0.715**	**0.730**	**0.780**
OTTER	0.668	0.562	0.547	0.673	0.474	0.511
PANDA	0.506	0.508	0.535	0.506	0.535	0.545
GENIE3	0.503	0.500	0.507	0.499	0.501	0.500
TIGRESS	0.512	0.505	0.502	0.502	0.501	503
BITFAM	0.521	0.524	0.521	0.517	0.519	0.517
Prior	0.513	0.515	0.533	0.529	0.505	0.562

Table 5.4: Comparison of AUC-ROC score with ChIP-seq gold standard for the GRNs inferred by GIRAFFE, OTTER, PANDA, prior, GENIE3, TIGRESS and BITFAM across five different tissues.

the `removeBatchEffect` function [2] from the `limma` package [Ritchie et al., 2015], a routine commonly used in bioinformatics to adjust for variables of interest. The AUROC using GIRAFFE's vanilla version is 0.730, which improves to 0.758 when adjusting for ischemic time, and to 0.744 when using the preprocessed gene expression.

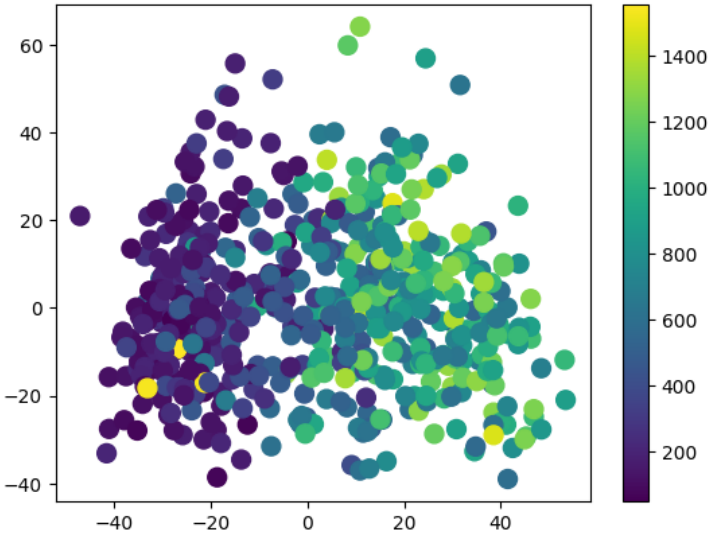

Figure 5.15: Two-dimensional projection of the gene expression dataset obtained via PCA. The colors show the ischemic time for each sample.

[2] To adjust for the effect of a variable h in a matrix X, `removeBatchEffect` first fits a linear regression on every column of X, using h and the other columns as predictors. The value of h multiplied by its corresponding estimated regression coefficient is then subtracted from the original value of the column, yielding the adjusted result.

5.5 Oncogenes and tumor suppressor genes[3]

Cancer driver genes, those that drive the transformation of malignant cells, can be also classified into oncogenes (OGs) or tumor suppressor genes (TSGs). Oncogenes, when their normal behavior is disrupted, drive changes in cells' behavior such that they start growing out of control, causing cells to become cancerous [Kontomanolis et al., 2020]. Tumor suppressor genes, on the other hand, are normally protecting the cell from damages and uncontrolled growth, and a loss of their function may lead to malignant transformation [Grandér, 1998]. In this section, we investigate whether GIRAFFE is able to correctly discriminate the role of OGs and TSGs in different human tissues. We expect TSGs to be more tightly regulated in healthy patients, and OGs being more tightly regulated in cancer patients [Lopes-Ramos et al., 2017].

First, we recover GRNs using GIRAFFE for the cancer population, using expression data provided in TCGA [Gao et al., 2019], and for the healthy population, using data provided in GTEx [Consortium et al., 2015]. In both cases, we employ a motif-based prior downloaded from CIS-BP [Weirauch et al., 2014] and PPI from the STRING database [Szklarczyk et al., 2016]. Then, we compute a regulation score for each gene, defined as the sum of its in-degree in the inferred GRN, and we compare the score difference between the healthy and cancer populations. More precisely, let $\beta^H_{i,k}$ be the GIRAFFE coefficient describing the relationship between TF k and gene i in the healthy network, and $\beta^C_{i,k}$ be the analogous quantity for a specific cancer population. We define the regulation score R^P_i for gene i in population $P \in \{H, C\}$ as

$$R^P_i := \sum_{k=1}^{|TF|} \beta^P_{i,k},$$

and we are interested in the difference $D_i := R^H_i - R^C_i$. Following our assumption, we expect a TSG gene i to have $D_i > 0$, and an OG $D_i < 0$. Importantly, this definition of the regulation score takes the activating or inhibitory nature of the predicted interactions into account. For instance, if a gene has mostly inhibitory interactions in the healthy network, D_i decreases, providing less evidence that gene i could be a TSG, and more that it could be a OG. We then compare our predictions with a curated list of tissue-specific oncogenes and tumor suppressor genes provided by COSMIC [Bamford et al., 2004]. First, we restrict our analysis to the drivers whose $|D_i|$ are in the top 10% of the regulation score distribution; we do not expect to detect different regulation for all drivers, but when the difference is relatively large,

[3]The authors thank Viola Fanfani for the thoughtful critiques and helpful review of this section.

we would like to get the correct sign ($D > 0$ for OG and $D < 0$ for tsg). GIRAFFE is able to correctly discriminate the difference in regulation for 6 TSGs in breast (86% accuracy), one OG in colon, 2 TSGs in skin, one OG in lung, one OG and one TSG in thyroid (all with 100% accuracy).

5.6 Sex differences in lung adenocarcinoma

In this section we apply GIRAFFE to *graph differential analysis*, an established downstream application to compare biological networks derived from different phenotypes. Examples of differences studied in the literature include healthy and cancer populations [Grechkin et al., 2016], different disease stages [Cuomo-Haymour et al., 2022, Lafaurie et al., 2023], epigenetic states [Del Real et al., 2017, Natanzon et al., 2018], and sex [Lopes-Ramos et al., 2020]. Figure 5.16 shows a typical graph differential analysis pipeline on two toy-size GRNs.

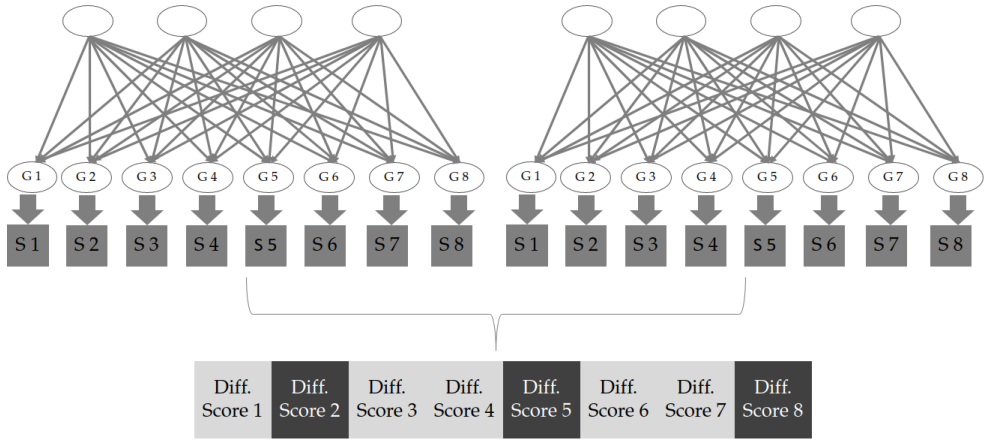

Figure 5.16: Visualization of a typical graph differential analysis pipeline. First, two networks are computed for the variables that we would like to analyze, e.g. disease states. Then, for each gene in a given network, a score is computed, for instance the sum of the incoming edges. The differential score for a given gene is then the difference between its corresponding score in both networks. The dark gray boxes in the differential score vectors represent three genes in a the same pathway.

First, the GRNs are computed using specific data for the conditions under study. In the case of GIRAFFE, each GRN has to be computed using expression data of individuals belonging to the corresponding population, and condition-specific prior knowledge should be used if available. Then, for each gene in each network, an aggregate score is computed. Options range from heuristics such as the sum of the incoming edges to more involved embeddings obtained via representation learning techniques [Grover and Leskovec, 2016, Perozzi et al., 2014]. These scores are then used to compute

a differential score for each gene, quantifying the similarity in the regulation mechanism with respect to the chosen aggregate score. Finally, statistical tests are performed to determine whether groups of genes belonging to a certain *biological pathway* [5] have large differential scores. Since the regulation of most genes is expected to be stable across different conditions [Singh et al., 2018], pathways exhibiting different regulatory mechanisms are likely to be responsible for phenotype-specific behaviours, and their identification can improve our understanding of the underlying biology, as well as our ability to treat, cure, and prevent diseases.

Previous research shows that multiple types of cancers are exhibited differently in males and females [Clocchiatti et al., 2016], and these differences are reflected in the gene regulatory networks for each sex [Kukurba et al., 2016, Sugathan and Waxman, 2013]. Inspired by Lopes-Ramos et al. [2020], we study sex differences in regulation for both healthy and lung adenocarcinoma (LUAD) cells. Our experimental setting is summarized in Figure 5.17.

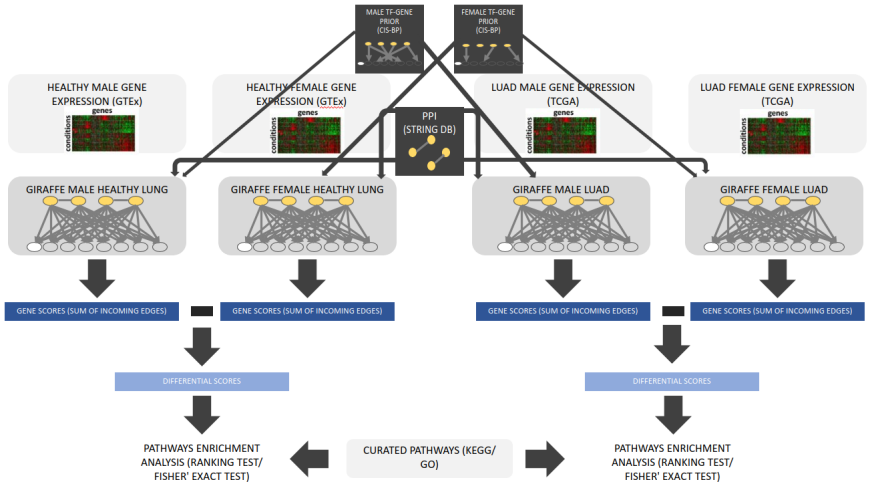

Figure 5.17: Visualization of our graph differential analysis pipeline. After recovering the GIRAFFE's networks using sex-specific prior knowledge (when possible), we computed the differential score for each gene as the difference between the sum of its in-degrees in both networks. Finally, we applied pathways enrichment analysis on the differential scores.

First we recover GIRAFFE networks for four different populations: healthy men (238 donors), healthy women (122 donors), men affected by LUAD (242 donors), and women affected by LUAD (280 donors). The LUAD data is provided in TCGA [Gao et al., 2019], and the normal lung data is provided

[5] The term biological pathway refers to a set of genes known to affect the organism's behaviour in a certain context. Examples include metabolic pathways, signaling pathways, and pathways determining the response to drugs.

in GTEx [Consortium et al., 2015]. We integrate a sex-specific binary prior for TF-gene binding from CIS-BP [Weirauch et al., 2014], and sex independent PPI data for 640 transcription from the STRING database [Szklarczyk et al., 2016]. We then run two separate differential analyses: the first one to study sex biases in healthy populations, and the second one to study sex biases in LUAD populations. In both cases, the pipeline to determine the enriched pathways is identical. Let $\beta^F_{i,k}$ be the GIRAFFE coefficient describing the relationship between TF k and gene i in the female network, and $\beta^M_{i,k}$ the analogous quantity for the male population. The differential scores for each gene i, defined as

$$D_i := \sum_{k=1}^{|TF|} \beta^M_{i,k} - \beta^F_{i,k},$$

are collected in a single vector that we use for gene enrichment analysis. In particular, we run the ranking tests implemented in FGSEA [Korotkevich et al., 2016] against KEGG's curated pathways [Kanehisa et al., 2017]. We use p-values corrected with FDR 0.05 using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995].

Figure 5.18 shows the pathways enriched for healthy patients, and Figure 5.19 for LUAD patients. In both figures, a negative normalized enrichment score (NES) indicates higher regulation in males, while a positive NES indicates stronger regulation in females. We observe that in normal lung, most enriched pathways have a higher score in men, indicating that their regulatory interactions tend to be stronger. This could be linked to the fact that women are more likely to get lung cancer than males [Patel, 2005]. Conversely, in LUAD, most enriched pathways have higher scores in women. This backs up the fact that women have greater survival rates regardless of stage, histology, or smoking status, even after adjusting for gender-specific life expectancy [North and Christiani, 2013]. Moreover, differentially targeted pathways include immune processes and drug metabolism-related pathways, potentially supporting previous work showing that men and women respond differently to drugs [Lopes-Ramos et al., 2018]. Finally, the most differentially expressed pathways in LUAD is MAPK signaling, a well known cancer pathway and a main target of therapies for lung cancer [Lu et al., 2011, Liang et al., 2012, Jain et al., 2021] that has recently been shown to participate in genetic differences between male and female in non-smoking lung adenocarcinoma patients [Xu et al., 2022].

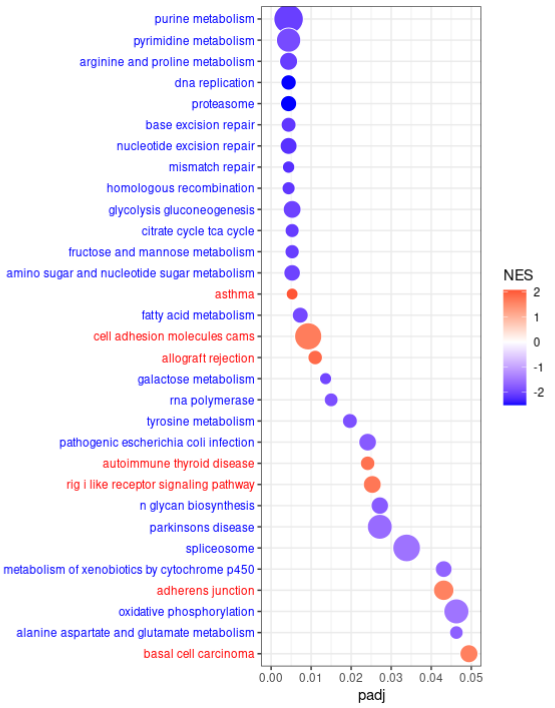

Figure 5.18: Pathways enriched for healthy lung using GIRAFFE.

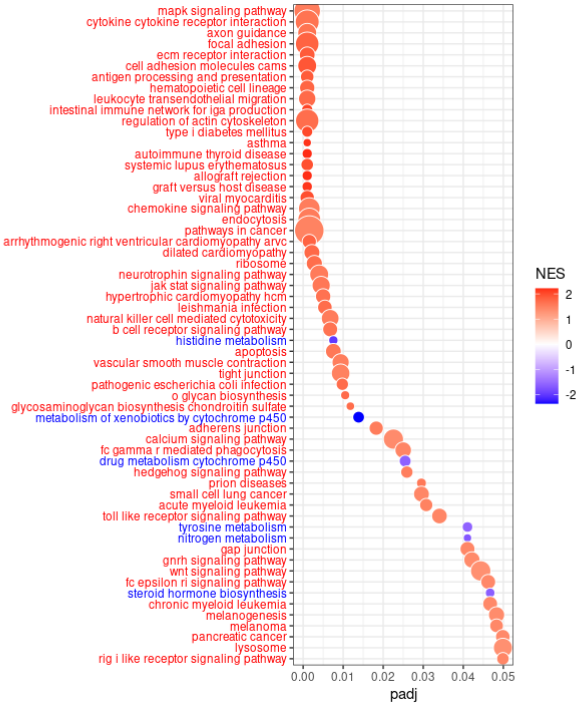

Figure 5.19: Pathways enriched for LUAD using GIRAFFE.

To further investigate the plausibility of our findings, we compare them with Lopes-Ramos et al. [2020], who conducted the same analysis using PANDA. Their results are reported in Figures B.10-B.11. In the healthy population, out of the 31 pathways enriched in GIRAFFFE and 27 pathways enriched in PANDA, 15 are shared: cell adhesion molecule cams, asthma, autoimmune tyroid disease, and allograft rejection with $NES > 0$; spliceosome, proteasome, DNA replication, pyrimidine metabolism, oxidative phosphorylation, metabolism of xenobiotics by cytochrome p450, RNA polymerase, fructose and mannose metabolism, base excision repair, mismatch repair, amino sugar, and nucleotide sugar metabolism with $NES < 0$. There are no enriched pathways with discordant NES sign. In the LUAD population, out of the 60 pathways are enriched in GIRAFFFE and 16 pathways enriched in PANDA, 10 are shared: intestinal immune network for iga production, asthmam, autoimmune tyroid disease, systemic lupus erythematosus, graft versus host disease, type I diabetes mellitus, leishmania infection, cytokine cytokine receptor interaction, antigen processing and presentation, all with $NES > 0$. Endocytosis and pathways in cancer are enriched with positive NES by GIRAFFE and negative NES by PANDA.

Chapter 6

Discussion

In this thesis we have developed GIRAFFE, a new algorithm for the joint inference of gene regulatory networks and transcription factor activities. Our approach is based on a factorization of the gene expression matrix, where we integrate biological knowledge, i.e. protein-protein interaction and a prior for the TF-gene binding, to guide the optimization. As backed-up by theoretical considerations [Wolpert and Macready, 1997], incorporating prior knowledge into the algorithm's behaviour is a good practice in many problems, and here it helps learning regulatory interactions that would be difficult to infer using any data type in isolation.

We modeled the problem as a graph whose node set is partitioned into genes and transcription factors. Undirected edges between pairs of transcription factors are extracted from the prior protein-protein interaction, while the directed edges from transcription factors (sources) to genes (targets) are inferred. In contrast to many other methods, that aim to understand regulation through gene co-expression networks, our structure better represents the underlying biology: an edge between TF i and gene j indicates that the former is, possibly by forming a higher-order protein complex, a regulator of the latter, and the edge weight quantifies the strength of the relationship.

GIRAFFE's model and loss function are inspired by OTTER and PANDA, which incorporate the same types of prior knowledge to infer edges from transcription factors to genes. In particular, the regularizers of our loss function - which encourage interacting proteins to target the same genes, and correlated genes to have similar interactions with TFs - resemble the message passing equations from PANDA and OTTER's objective. The advantage of GIRAFFE is framing the problem as a matrix factorization, yielding a model with a different interpretation. Instead of having larger weights indicating higher probability of regulation, GIRAFFE uses the inferred transcription factor activities to estimate partial effects, which allow to distinguish activating from inhibitory interactions.

In Chapter 5, we evaluated GIRAFFE in relation to the objectives outlined in Section 1.1. In particular, we investigated its ability to scale up to the size of the human genome, its accuracy, interpretability, flexibility, and the plausibility of the inferred networks in biological applications.

In terms of scalability, GIRAFFE is able to efficiently infer GRNs and transcription factor activities on human datasets, which are typically composed by 20 to 30 thousand genes. Using a single CPU machine with 16GB RAM, our algorithm took 150 seconds on average for the tissues in Table 5.4. Exploiting Pytorch's gradient-based methods comes with a significant computational advantage, and if necessary GIRAFFE can be further sped up with modern hardware such as GPUs.

We demonstrated the accuracy of the inferred regulatory interactions and transcription factor activities through extensive experiments in both synthetic and real world data. Considering the AUROC between the ground-truth and inferred regulatory network, GIRAFFE outperforms OTTER and PANDA in both our simulations, being the most robust method with respect to corruptions in the prior network. When we evaluated the AUROC using Ch-IP seq data as ground-truth for breast, colon, liver, lung, and prostate, GIRAFFE had the largest margin over the competing algorithms, which in some cases performed only slightly better than random. Since these tissues used the same database as a source for the ground-truth, to ensure independence, we performed additional experiments unsing different sources for ChIP-seq data on kidney and yeast. Even if with lower margin, GIRAFFE was still the best performing algorithm.

Note that, similarly as GIRAFFE, aso TIGRESS considers a regularized linear regression between TFs and genes. GENIE3 takes this a step further by accounting for non-linear interactions, which are known to increase the predictive power of the model. A reason why they perform worse than GIRAFFE could be that both of them consider gene expression as a proxy for TF activity, suggesting that incorporating TFA as a latent factor might better model the underlying mechanisms. Moreover, they do not incorporate prior knowledge, reinforcing the importance of integrating multiple data types to learn meaningful gene regulatory networks. Finally, we also studied the quality of the inferred TFA matrix: in the simulations in Section 5.1, we showed that GIRAFFE recovers transcription factor activities yielding a small reconstruction error, but without necessarily guaranteeing identifiability. Section 5.3 studied the inferred TFA from a more biological perspective, showing that it was able to correctly assign the lowest activity to the knocked-out transcription factor in 31 out of 50 cases, maintaining an average rank of 8.17.

The flexibility of GIRAFFE is reflected by the possibility to customize the model to the requirements of a downstream application. First, the linear re-

gression can be extended to incorporate further variables of interest. In Section 5.1, we showed that adjusting for confounders, i.e. variables affecting both the TF activity and gene expression in the data generating mechanism, and biasing variables, i.e. variables affecting gene expression only, is beneficial. As a proof-of-concept, in Section 5.4, we identified ischemic time as a potential batch effect in lung, and we showed that adjusting for it increases the AUROC from 0.730 to 0.758. This improvement is similar to the one we get when applying GIRAFFE with a gene expression matrix preprocessed with `limma`, an established package to correct for batch effects. The second option to customize GIRAFFE is promoting sparsity in the inferred regulatory matrix by exploiting the geometry of the ℓ_1-norm and the proximal operator. This can be useful for applications where a correct identification of the regulator is essential, e.g. to exploit genetic perturbations [Cai et al., 2013], and can be combined with the stability selection framework to control for the expected number of false positives.

To investigate GIRAFFE's behaviour in downstream applications, we combined it with gene set enrichment analysis to study sex differences in lung. We observed that GIRAFFE finds a greater number of enriched pathways. A reason for this could be related to the interpretation of the weights as partial effects, that assign a positive/negative sign to activating/inhibitory regulatory interactions. Let's consider, for instance, a pair of transcription factor-gene that is enhancing in men and inhibitory in women. PANDA and OTTER, if they are correctly confident that a regulatory relationship exists, would predict a large edge weight in both cases. In this way, however, the edge difference between the sexes could cancel out, therewith contributing to a pathway not being enriched. This suggests that integrating GIRAFFE in graph differential analysis pipelines could be beneficial to shed light on unknown cancer mechanisms, as well as sex biases in diseases.

There is still a number of limitations in GIRAFFE. First, our model does not take saturation effects into account. While it remains true that the probability of binding often depends linearly on the transcription factor activity [Halford and Marko, 2004], recent research suggests that increasing the activities beyond a certain threshold does not affect the expression of target genes [Koşar and Erbaş, 2022]. To address this issue one could explore the effect of incorporating (smooth) thresholding functions into GIRAFFE's objective function. Second, our model requires the introduction of a parameter for the number of iterations, whose value affects the configuration of the inferred regulatory network. For instance, when using a motif-based binary prior, the weights distribution can shift from a bimodal distribution with various peaks' distances to a unimodal distribution. Even if using a continuous motif, such as the Garcia-Alonso model [Garcia-Alonso et al., 2019], can mitigate the phenomenon, we aim to further investigate the role of early stopping in GIRAFFE and provide a more principled convergence criterion.

Our future work aims to expand the capabilities of GIRAFFE by integrating additional data types, such as genetic variants, mutations, chromatin accessibility, and SNPs, all of which can have an impact on regulatory interactions [Martin et al., 2019]. We believe that incorporating these data types has the potential to enhance our inference method. We could achieve this by integrating the epigenomic profile into the prior network in a preprocessing step, similarly as proposed for EGRET [Weighill et al., 2022], or by adapting GIRAFFE's optimization procedure. We are also excited to apply GIRAFFE to single-cell datasets and disease contexts, with the goal of generating novel testable hypotheses regarding the role of gene regulation in cancer.

Chapter 7

Concluding thoughts

In this thesis, we studied the problem of inferring gene regulatory mechanisms from data, in particular the role of transcription factor binding. As a crucial component to understand cancer, this has applications in drug discovery, early detection, and personalized treatment.

To tackle the challenge of combining accuracy, interpretability, scalability, and flexibility in a single model, we introduced GIRAFFE, a new algorithm to jointly infer regulatory effects and transcription factor activities. Assuming linear relationships, our model estimates partial effects considering transcription factor activities as covariates and gene expression as target variable. Hence, it is able to distinguish activating from inhibitory regulation based on the weight's sign. Framed as a biologically informed matrix factorization, our loss function is minimized with gradient-based methods, yielding an efficient algorithm that scales up to the size of the human genome. To further customize the model, we included the possibilities to adjust for variables of interest and promoting sparsity in the inferred regulatory network.

Our analysis demonstrated that GIRAFFE is able to accurately reverse engineer gene regulatory networks by outperforming state-of-the-art gene regulatory inference methods on both synthetic and real world datasets. By using in silico data and a yeast interventional dataset as proof-of-concepts, we also showed that it can infer reasonably meaningful transcription factor activities. Moreover, when applied to study sex differences in lung, GIRAFFE recovers knowledge that is consistent to both the literature and established methods, reinforcing its potential of leading to valuable insights in biological contexts.

A key advantage of GIRAFFE is that its regularized linear regression approach can be further generalized to integrate additional data types such as epigenomic profiling data, opening avenues for future research. We believe that this approach has tremendous potential to inform our understanding of cancer, contributing towards more effective and personalized healthcare.

Appendix A

Sparse optimization

In this appendix we show our approach to solve

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \lambda||\mathbf{x}||_1, \tag{A.1}$$

where f is a non-convex and differentiable objective function. Note that Problem 4.3 is an instance of Equation A.1. The goal is applying ℓ_1 regularization to the optimization process of $f(\mathbf{x})$ in order to obtain a sparse solution.

A.1 Proximal Adam

Adam [Kingma and Ba, 2014] is an adaptive gradient method to minimize a differentiable function $f(\mathbf{x})$. It iteratively adapts an initial guess $\mathbf{x_0}$ until convergence using the rule

$$\mathbf{x_{t+1}} = \mathbf{x_t} - \alpha \frac{\varphi(\mathbf{m_t})}{\psi(\mathbf{x_t})},$$

where $\alpha > 0$ is the learning rate, and $\mathbf{m}_t$ and $\mathbf{v_t}$ are estimates for the first and second order of $\nabla f(\mathbf{x})$. For our purposes, we can conveniently abstract from the exact formulation of the exponential decay functions φ and ψ. Note that we keep α constant to simplify the notation, but our conclusions can be naturally extended to incorporate an iteration-dependent learning rate α_t.

Similarly as in the proximal gradient method [Rockafellar, 1997], the vanilla version of Adam can be adapted to solve the composite optimization problem A.1 and get sparse solutions [Melchior et al., 2019].

Definition A.1 *The **proximal operator** of a convex function g at $\mathbf{x}$ is defined as*

$$prox_g(\mathbf{x}) := \arg\min_{\mathbf{y}}\{g(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||_2^2\}$$

Conveniently, the proximal operator for the ℓ_1-norm can be computed in closed form.

Lemma A.2 *If $g(\mathbf{x}) = \lambda||\mathbf{x}||_1$ for $\lambda \in \mathbb{R}_+$, then $prox_g(\mathbf{x})$ is given by*

$$prox_{\lambda||\cdot||_1}(x_i) = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda \\ 0 & \text{if } |x_i| \leq \lambda \\ x_i + \lambda & \text{if } x_i < -\lambda \end{cases}$$

*which is often referred to as **soft-thresholding operator**. In particular, the proximal operator for the ℓ_1-norm can be efficiently computed in $\mathcal{O}(p)$.*

Proof The ℓ_1-norm is separable, and thus we can consider each of its components separately. Since $g(x_i)$ is convex, it is sufficient to find y_i such that the optimality condition

$$0 \in \nabla(\lambda|y_i| + \frac{1}{2}(x_i - y_i)^2)$$

is satisfied. We apply a case distinction to the optimal solution y_i^*.

For $y_i^* > 0$, the optimality condition yields

$$\lambda + y_i^* - x_i = 0 \Longrightarrow y_i^* = x_i - \lambda.$$

and holds iff $x_i > \lambda$.

For $y_i^* < 0$, the optimality condition yields

$$-\lambda + y_i^* - x_i = 0 \Longrightarrow y_i^* = x_i + \lambda.$$

and holds iff $x_i < -\lambda$.

For $y_i^* = 0$ we rely on subdifferential theory. Since $\partial\lambda|x_i| \in [-\lambda, \lambda]$, we get

$$0 \in [-\lambda, \lambda] + y_i^* - x_i \Longrightarrow x_i \in [-\lambda, \lambda] \Longrightarrow |x_i| \leq \lambda,$$

which concludes the proof. □

After picking an initial guess $\mathbf{x_0} \in \mathbb{R}^p$, proximal Adam alternates between gradient update and proximal operator using

$$\mathbf{x_{t+1}} = prox_{\lambda||\cdot||_1}\left(\mathbf{x_t} - \alpha\frac{\varphi(\mathbf{m_t})}{\psi(\mathbf{x_t})}\right)$$

In other words, after computing the gradient step, proximal Adam shrinks to zero all entries of $\mathbf{x_t}$ that are less than λ in absolute value. While we couldn't formally derive optimality conditions for proximal Adam, we provide an intuition on why it has been shown to provide good performance in multiple applications including matrix factorization [Melchior et al., 2019]. The step direction of Adam can be interpreted as a bounded gradient [Kingma and Ba, 2014], and by introducing the approximation $\frac{\varphi(\mathbf{m_t})}{\psi(\mathbf{x_t})} \approx \nabla f(\mathbf{x_t})$, we get:

$$\begin{aligned}
& \mathbf{x}^* = prox_{\alpha\lambda||\cdot||_1}(\mathbf{x}^* - \alpha\nabla f(\mathbf{x}^*)) \\
\iff\ & 0 \in \alpha\partial(\lambda||\mathbf{x}^*||_1) + (\mathbf{x}^* - (\mathbf{x}^* - \alpha\nabla f(\mathbf{x}^*))) \\
\iff\ & 0 \in \partial(\lambda||\mathbf{x}^*||_1) + \nabla f(\mathbf{x}^*) \\
\iff\ & \mathbf{x}^* \text{ is a local optimum}
\end{aligned}$$

In other words, in case of convergence, proximal methods approach a stationary point. Moreover, under a specific tuning of the learning rate α, convergence can be guaranteed [Défossez et al., 2020]. However, despite of the lack of a proof for convergence, constant learning rates are common [Kingma and Ba, 2014].

A.2 Tuning of the regularization parameter

The main reason why we introduced ℓ_1 regularization in Equation A.1 is to obtain sparse solutions. The underlying assumption is that only some entries of R are are non-zero. We call this the *active set*

$$R^0 := \{(i,j) : R_{i,j} \neq 0\},$$

and our goal is estimating a gene regulatory networks $\hat{R}$ s.t. $\hat{R}^0 := \{(i,j) : \hat{R}_{i,j} \neq 0\} \approx R^0$. To achieve this result, the choice of the regularization hyperparameter λ is crucial: small values of λ lead to an $\hat{R}^0$ with larger size and potentially many false positives; on the other side of the spectrum large values of λ yield a very sparse estimate, with the risk of missing true edges.

In this Section we summarize stability selection, a framework proposed by Meinshausen and Bühlmann [2010] to estimate the active set. It allows to determine how "stable" the selection of a certain edge is, and how to achieve some type-I error control for the number of false positive edges.

The main idea is to assign a relevance score to each edge relying on a subsampling approach. Let I^* be a random subsample of $\{1, \dots, n\}$ of size $\lfloor \frac{n}{2} \rfloor$. After selecting the corresponding samples in the gene expression, we run GIRAFFE with a given λ to get an estimated active set $\hat{R}^0_\lambda(I^*)$. This procedure can be repeated B times to obtain different estimated active sets $\hat{R}^0_\lambda(I^{*1}), \dots, \hat{R}^0_\lambda(I^{*B})$. The relevance score for each edge is then computed as the overlap among the estimated active sets. More concretely, we define the relevance score for an edge (i, j) as

$$\hat{\Pi}_{(i,j)}(\lambda) := \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left[(i,j) \in \hat{R}^0_\lambda(I^{*b})\right].$$

Instead of computing the relevance score for a single value of λ, the stability framework suggest to consider a set of candidate values Λ, and then picking the stable active set $\hat{R}_{stable}$ via a cutoff π_{thr} as follows

$$\hat{R}_{stable} := \{(i,j) : \max_{\lambda \in \Lambda} \hat{\Pi}_{(i,j)}(\lambda) \geq \pi_{thr}\}.$$

In this way, the choice of the hyperparameter λ is relaxed (one can pick multiple candidate values, similarly as in CV) and the choice of the active set depends on the value of π_{thr} only. Note that a large value of π_{thr} is very conservative and potentially misses true edges, while a small value of π_{thr} can be too loose and incorporating many false positives. In other words, we pushed the burden from picking a suitable value for the regularization parameter λ to picking a suitable value of π_{thr}. Stability addresses this issue by providing a choice guaranteed to control for the expected number of false positives, i.e. type-I error control. Before we report their main theorem and how it can be used, we introduce some notation. Let $\hat{R}_\Lambda := \cup_{\lambda \in \Lambda} \hat{R}^0_\lambda$ be the set of selected variables for all variables of $\lambda \in \Lambda$, and let $q_\Lambda := \mathbb{E}\left[|\hat{R}^0_\lambda|\right]$ be the expected number of selected edges. Then the following theorem provides a principled way to select π_{thr}.

Theorem A.3 *Assuming both that the distribution of* $\mathbb{I}\left[(i,j) \in \hat{R}^0(\lambda)\right]$ *is exchangeable for all* $(i,j) \notin R^0$*, and that* $\hat{R}$ *is not worse than random guessing, then for* $\pi_{thr} \in (\frac{1}{2}, 1)$ *the expected number of false positives* V *is bounded by*

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q^2_\Lambda}{p},$$

where p is the size of our GRN.

In practice, we could run GIRAFFE for a set of regularization parameters and at keep only the top K edges at each iteration. This would ensure that $q_\Lambda \leq K$. Then, by picking

$$\pi_{thr} = \frac{1}{2} + \frac{K^2}{2pv_0},$$

we would know that the number of false positives is bounded by v_0. For instance, when inferring a GRN of size $p = 200 \cdot 10^3$, we could pick the top $K = 10^3$ edges and, for a guarantee of having at most 100 false positives we would have to pick $\pi_{thr} = 0.525$. In the original paper, the recommendation is picking a large value of K and letting the stability framework reduce the size of the estimated active set.

Appendix B

Supplementary materials

B.1 Transcription factor expression is not a reliable surrogate for its activity

In Chapter 4 we presented our model, where the relationship between each gene and the transcription factors is expressed as a linear regression (Equation 4.1). GIRAFFE does not only infer the coefficients of the linear regression (i.e. the GRN), but also its predictors (i.e. the transcription factor activities). The reason for this choice is that transcription factor activity is difficult to measure directly with current technologies and, due to the complex protein synthesis mechanism, using transcription factor expression as a surrogate is a poor idea. Here we present our experiments supporting this claim.

We consider a real gene expression dataset, with almost 20 thousand response variables Y_i measuring the expression of gene i across n samples, and $p = 481$ covariates X_j variables measuring the gene expression of transcription factor j across the same n samples. First, we fit Lasso regressions between the response variables Y_i and the predictors X_j, where we pick the regularization hyperparameters of the Lasso using 5-fold CV. To evaluate the quality of the predictors selected by the Lasso, we compare it with the motif. The underlying hypothesis is that transcription factors whose sequence motif is present in the promoter region of the target gene should have higher chances of being selected by the Lasso. However, when we average over all response variables, only 13.5% of the selected transcription factors have their sequence motif in the promoter region of the target. To show the robustness of our conclusion on a single gene selected at random, we apply a subsampling approach to assign a relevance score to each transcription factor. Figure B.1 shows the scores for transcription factors in the motif (black) and not in the motif (green). The motif transcription factors are not clustered on the right side of the plot (corresponding to high relevance pre-

dictors), supporting the claim that they are not more relevant than the other predictors.

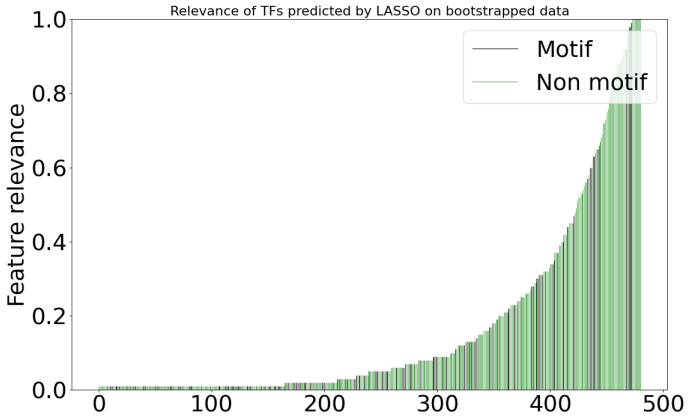

Figure B.1: Relevance scores for the transcription factors used as a predictor for the gene expression of a randomly selected target gene. The transcription factors whose sequence motif is present in the promoter region of the target gene are colored in black, while the other ones are depicted in green.

Finally, we run Ridge regressions for 500 genes (selecting the regularization hyperparameter with 5-fold CV) and we plot the weights for transcription factors whose motif is (not) in the promoter region of the target gene. Figure B.2 shows that both weights distributions have a similar shape, and that transcription factors in the motif tend to have lower weights. This contrast the biological intuition, as the motif is commonly used as a prior for regulation.

We conclude that we couldn't find any evidence that transcription factor mRNA expression can be reliably used as a proxy for its activity in a linear model such as ours. This observation, together with the conclusions from other studies [Ma and Brent, 2021, Latchman, 1993], is the main motivation why we opted to jointly infer transcription factor activities with the GRN.

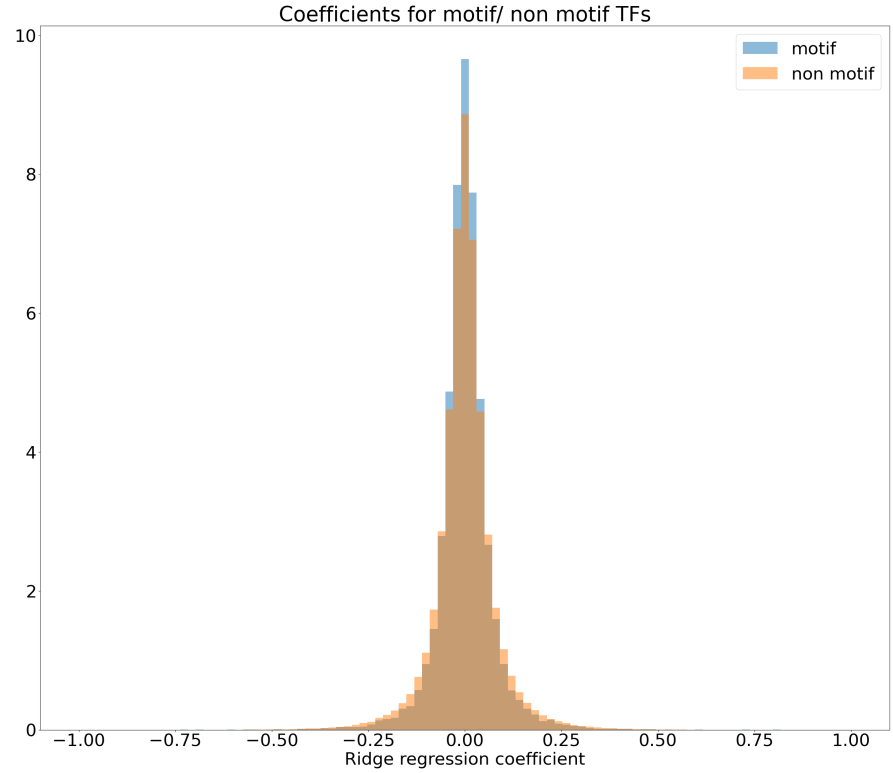

Figure B.2: Transcription factor coefficients distributions obtained with Ridge regression with regularization parameter computed with 5-fold CV. In orange the weights for transcription factors whose motif is not in the promoter region of the target, and in blue the weights for transcription factors whose motif is in the promoter region of the target.

B.2 Addendum to Section 5.1

In this section we show the robustness of the results presented in Section 5.1. In particular, we check the differences when we change the distribution of the non-zero entries of R (we test $\mathcal{U}(-a, a)$ for different values of a), and the value of the density parameter. Table B.2 shows the robustness of the results presented in Table 5.2 when 30% of the binary prior network are flipped for different combinations of sparsity/ width of the uniform distribution for R. Similarly as in the main experiment, all results are averaged over $B = 50$ runs. Similarly as in Figure 5.2, Figure B.3 shows the sign accuracy for different noise settings. We observe that the same trend holds for different distributions of the true regulation matrix R. Figure B.4 shows the ROC curves for the sparsity scores for different setting as in Figure 5.4. Finally, we report the individual ROC curves for Table 5.3.

Setting	Method			
	GIRAFFE	OTTER	PANDA	Prior
R density: 0.1	**0.890** $\pm$0.001	0.857$\pm$0.001	0.884$\pm$0.002	0.700$\pm$0.001
R density: 0.2	**0.852** $\pm$0.001	0.824$\pm$0.001	0.812$\pm$0.001	0.700$\pm$0.001
R density: 0.7	**0.759** $\pm$0.001	0.739$\pm$0.001	0.662$\pm$0.001	0.700$\pm$0.002
R density: 0.9	**0.734** $\pm$0.002	0.711$\pm$0.003	0.601$\pm$0.001	0.700$\pm$0.002
R i.i.d. from $\mathcal{U}(-1,1)$	**0.860** $\pm$0.001	0.839$\pm$0.004	0.844$\pm$0.001	0.700$\pm$0.001
R i.i.d. from $\mathcal{U}(-3,6)$	**0.865** $\pm$0.002	0.839$\pm$0.001	0.844$\pm$0.001	0.700$\pm$0.001
R i.i.d. from $\mathcal{U}(-5,10)$	**0.868** $\pm$0.001	0.840$\pm$0.001	0.844$\pm$0.001	0.700$\pm$0.001
R i.i.d. from $\mathcal{U}(-10,20)$	**0.870** $\pm$0.001	0.839$\pm$0.002	0.844$\pm$0.001	0.700$\pm$0.001

Table B.1: Comparison of AUC-ROC score of the GRNs inferred by GIRAFFE, OTTER, PANDA, and the prior. We flip 30% of the entries in the prior motif and investigate the robustness with respect to other parameters used in the simulation.

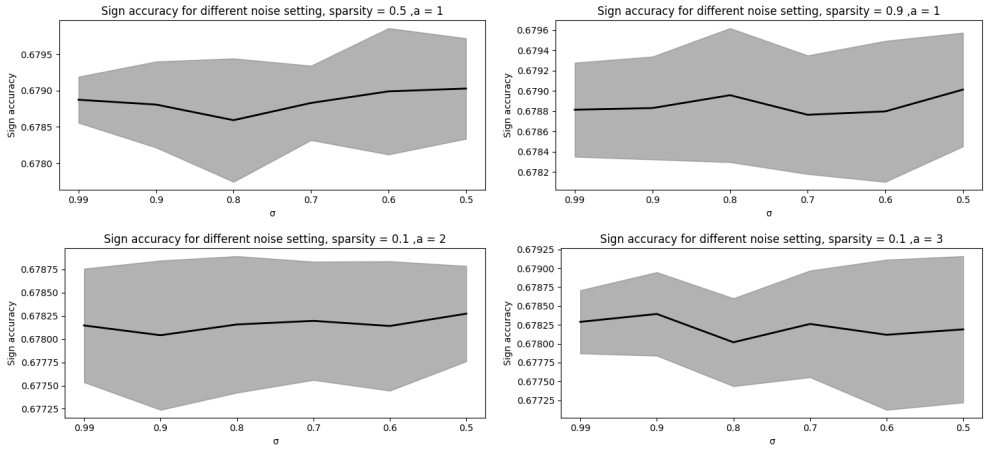

Figure B.3: Sign accuracy of $\hat{R}$ estimated by R for different distributions of R. The black line shows the mean over $B = 50$ runs, and the gray area is within one standard deviation from the mean.

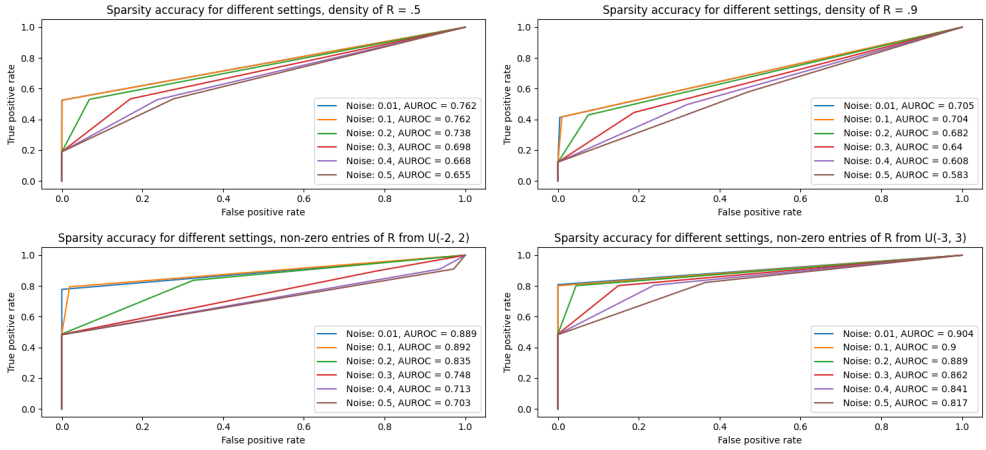

Figure B.4: AUROC for the accuracy of the estimated active set of R. We show the results for different densities of R and different width of the uniform distribution for non-zero entries.

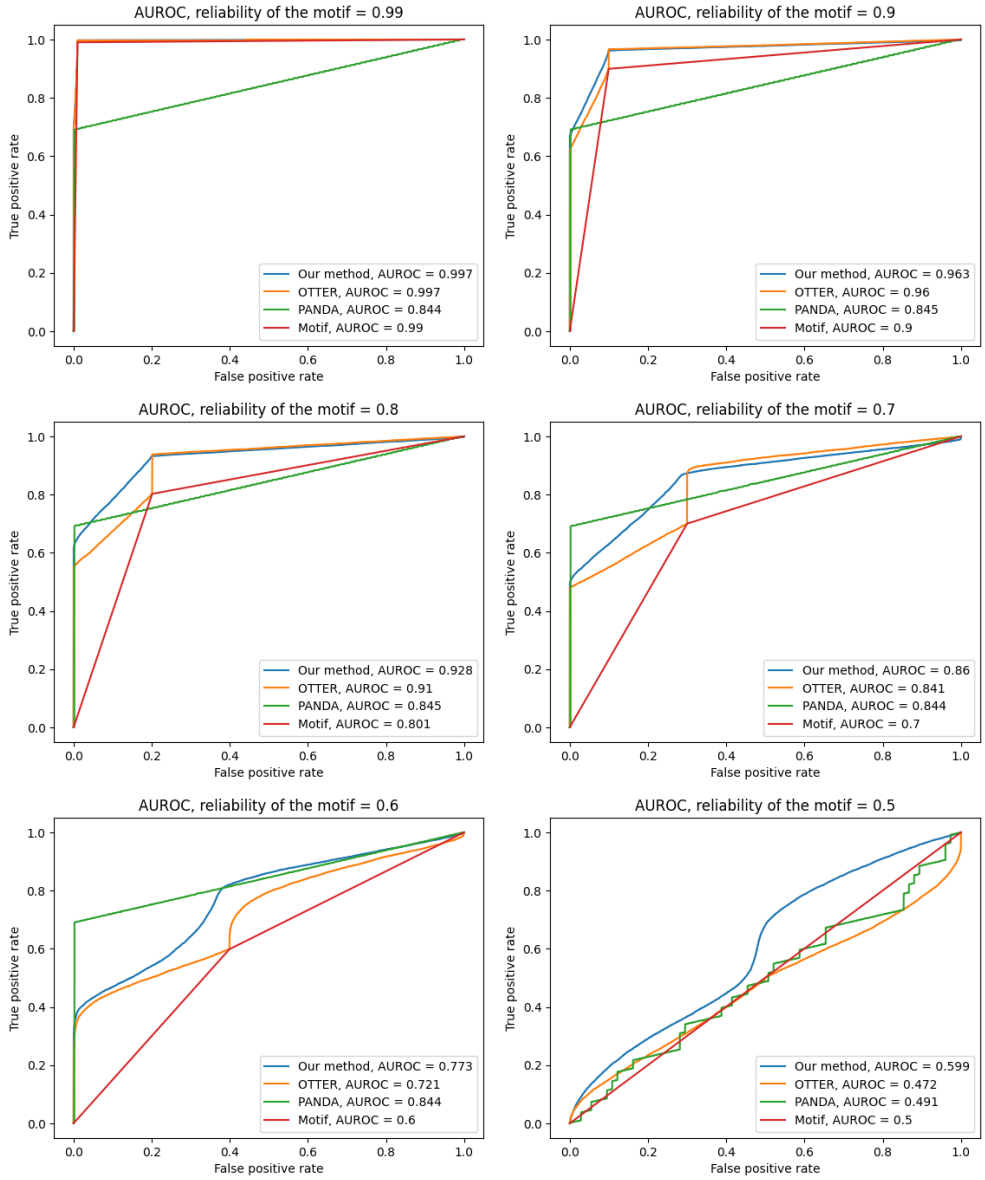

Figure B.5: ROC curves for the results in Table 5.2.

B.3 Addendum to Section 5.2

In this section we show the robustness of the results presented in Section 5.2. In particular, we check the differences when we change the distribution of the non-zero entries of R (we test $\mathcal{U}(-a, a)$ for different values of a), and the value of the density parameter. Table B.2 shows the robustness of the results presented in Table 5.3 for $\sigma = 0.5$ and different combinations of sparsity/ width of the uniform distribution for R. Similarly as in the main experiment, all results are averaged over $B = 50$ runs. Similarly as in Figure 5.12, Figure B.6 shows the sign accuracy for different noise settings. We observe that the same trend holds for different distributions of the true regulation matrix R. Figure B.7 shows the ROC curves for the sparsity scores for different setting as in Figure 5.13. Finally, we report the individual ROC curves for Table 5.3.

	Method			
Setting	GIRAFFE	OTTER	PANDA	Prior
R density: 0.3	**0.669** $\pm$0.002	0.645$\pm$0.002	0.553$\pm$0.002	0.646$\pm$0.003
R density: 0.5	**0.665** $\pm$0.002	0.645$\pm$0.002	0.551$\pm$0.003	0.650$\pm$0.002
R density: 0.7	**0.658** $\pm$0.002	0.648$\pm$0.001	0.544$\pm$0.001	0.649$\pm$0.001
R density: 0.9	**0.650** $\pm$0.003	0.646$\pm$0.001	0.542$\pm$0.005	0.648$\pm$0.002
R i.i.d. from $\mathcal{U}(-1, 1)$	**0.687** $\pm$0.001	0.647$\pm$0.004	0.552$\pm$0.004	0.647$\pm$0.004
R i.i.d. from $\mathcal{U}(-2, 4)$	**0.819** $\pm$0.001	0.796$\pm$0.002	0.614$\pm$0.002	0.795$\pm$0.003
R i.i.d. from $\mathcal{U}(-3, 6)$	**0.869** $\pm$0.002	0.862$\pm$0.001	0.646$\pm$0.001	0.861$\pm$0.001
R i.i.d. from $\mathcal{U}(-5, 10)$	0.908 $\pm$0.002	**0.916**$\pm$0.001	0.670$\pm$0.003	0.915$\pm$0.001
R i.i.d. from $\mathcal{U}(-10, 20)$	0.928 $\pm$0.002	0.954$\pm$0.002	0.691$\pm$0.003	**0.955**$\pm$0.002

Table B.2: Comparison of AUC-ROC score of the GRNs inferred by GIRAFFE, OTTER, PANDA, and the prior. We set $\sigma = 0.5$ and investigate the robustness with respect to other parameters used in the simulation.

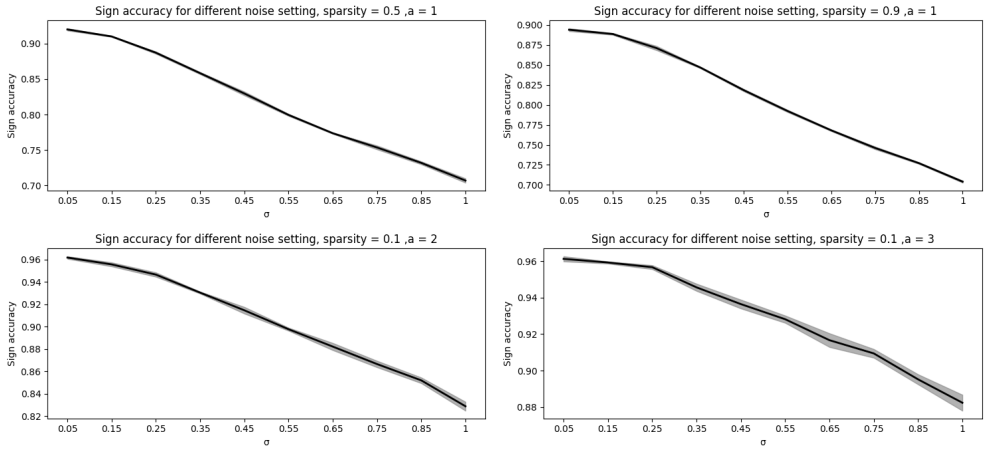

Figure B.6: Sign accuracy of $\hat{R}$ estimated by R for different distributions of R. The black line shows the mean over $B = 50$ runs, and the gray area is within one standard deviation from the mean.

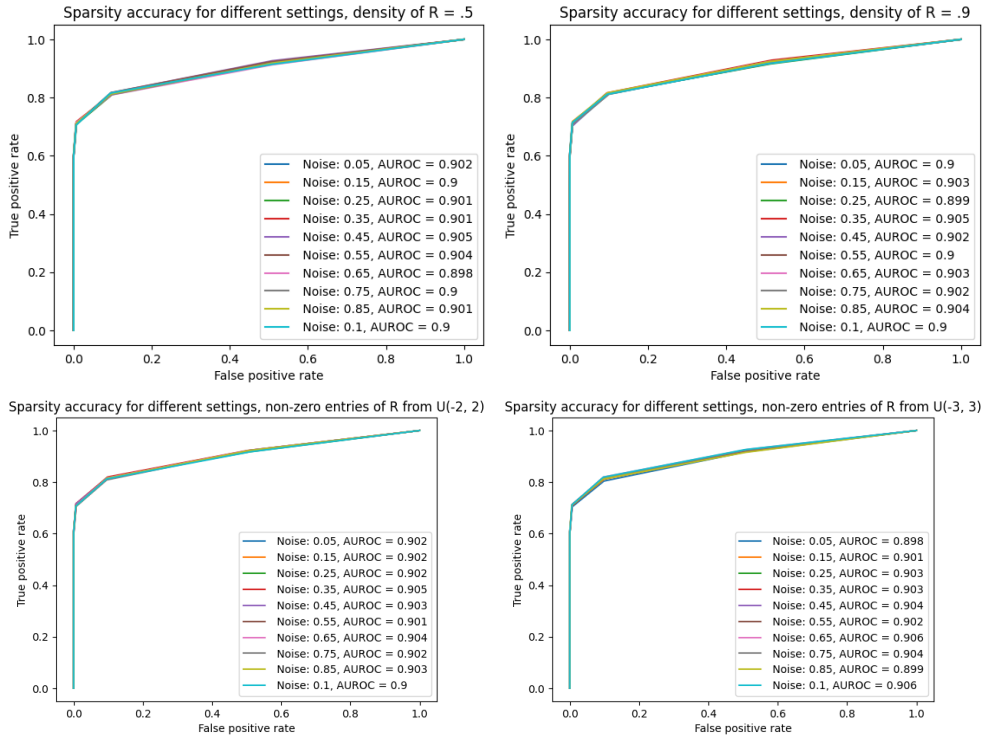

Figure B.7: AUROC for the accuracy of the estimated active set of R. We show the results for different densities of R and different width of the uniform distribution for non-zero entries.

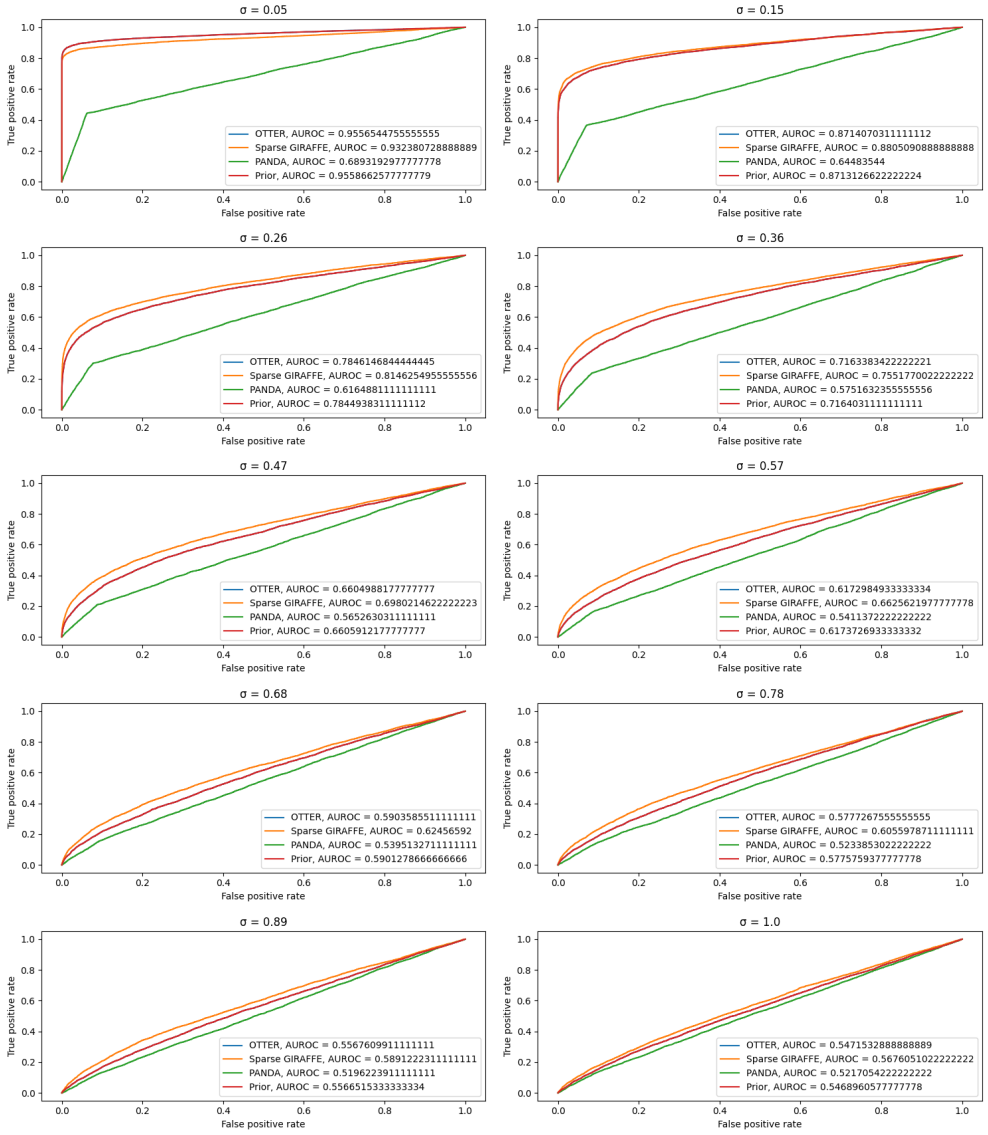

Figure B.8: ROC curves for the results in Table 5.3.

B.4 Addendum to Section 5.4

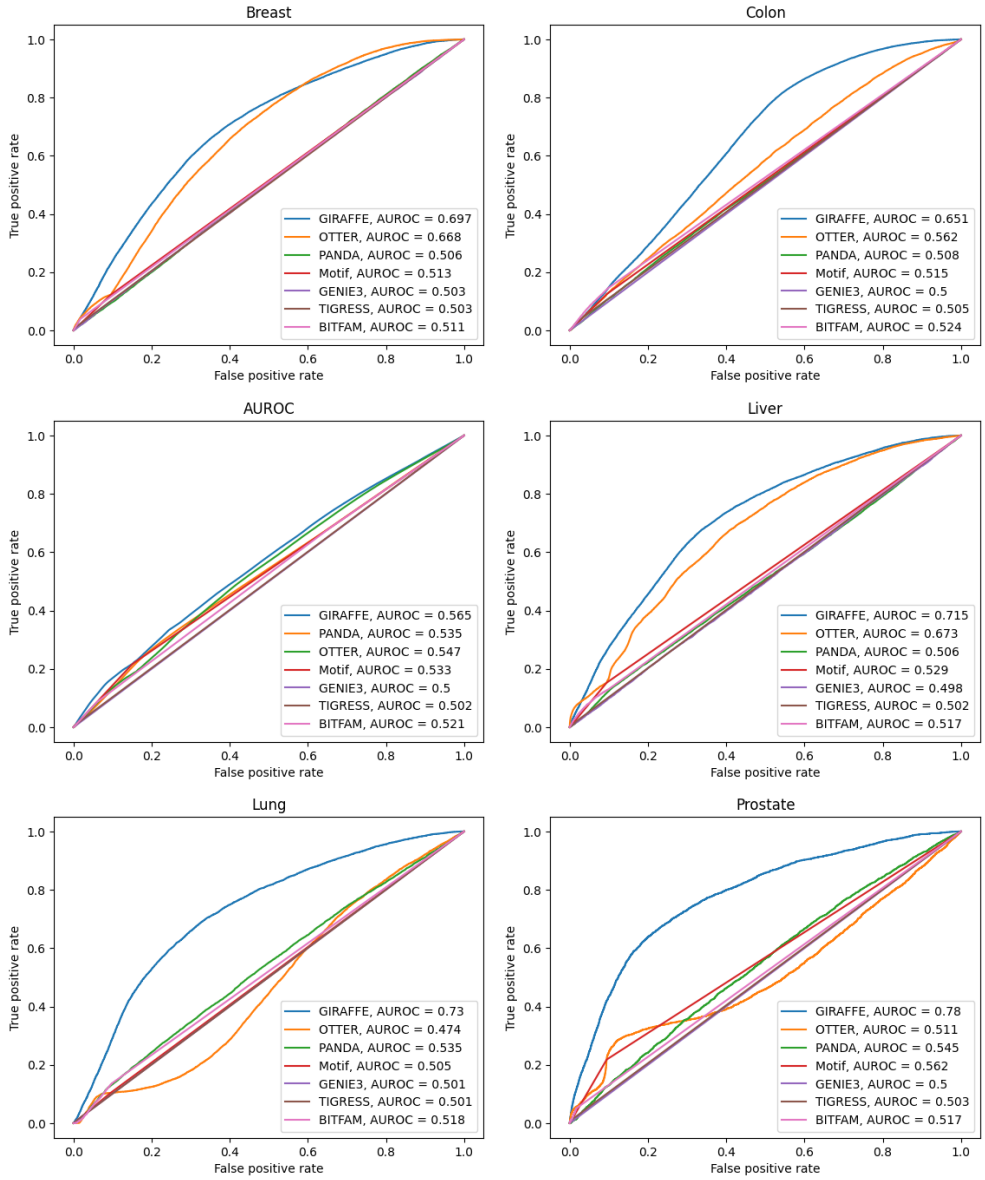

Figure B.9: ROC curves for the results in Table 5.4.

B.5 Addendum to Section 5.6

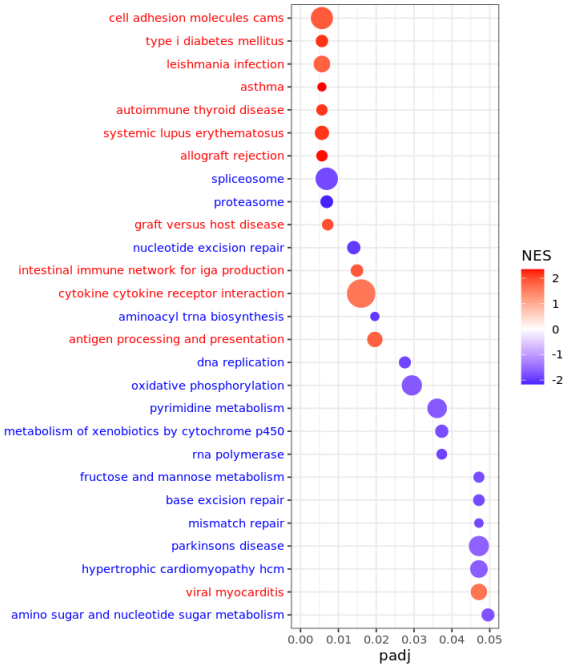

Figure B.10: Pathways enriched for healthy lung using PANDA.

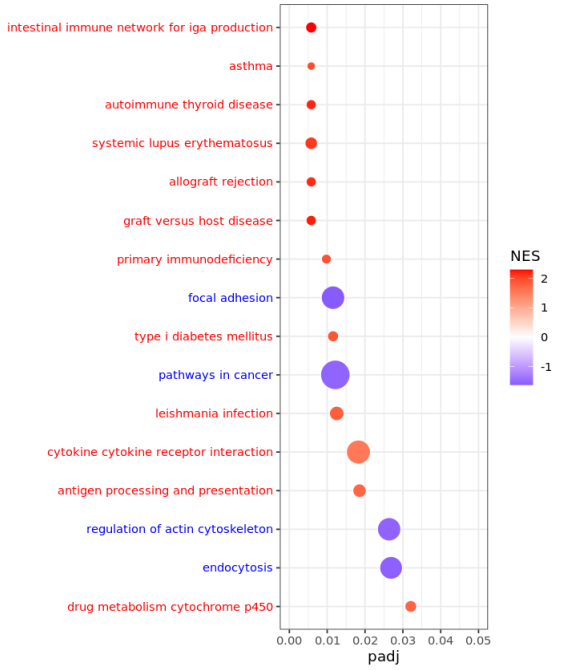

Figure B.11: Pathways enriched for LUAD using PANDA.

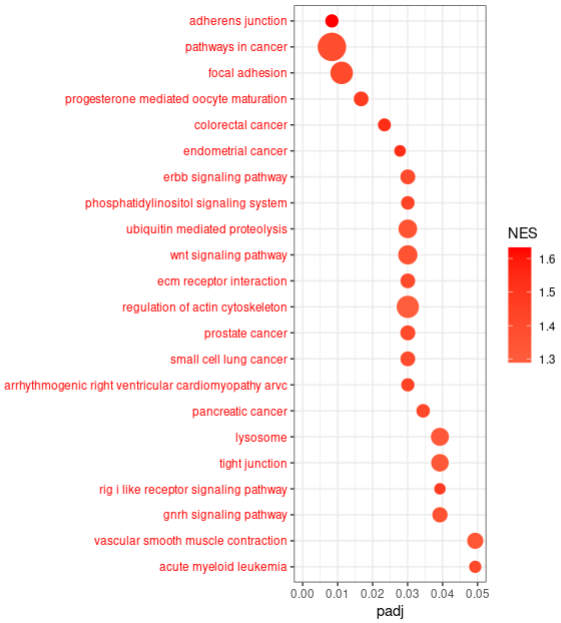

Figure B.12: Pathways enriched for healthy lung using OTTER.

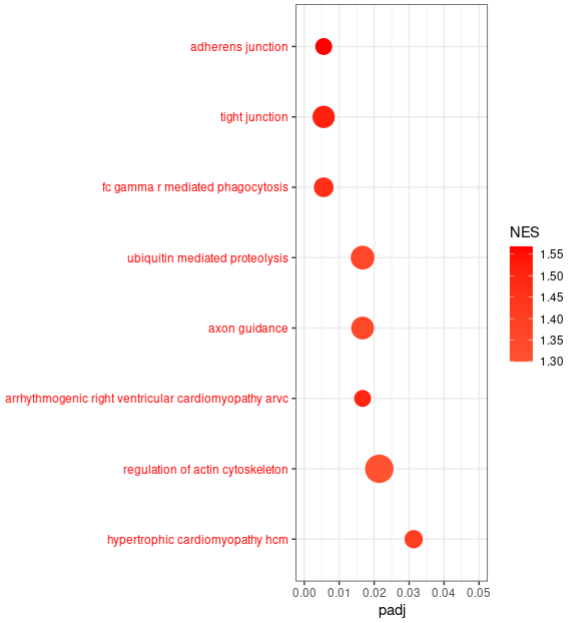

Figure B.13: Pathways enriched for LUAD using OTTER.

Bibliography

Rosa Aghdam, Mojtaba Ganjali, Xiujun Zhang, and Changiz Eslahchi. Cn: a consensus algorithm for inferring gene regulatory networks using the sorder algorithm and conditional mutual information test. *Molecular BioSystems*, 11(3):942–949, 2015.

Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Biocomputing'99*, pages 17–28. World Scientific, 1999.

John Aldrich. Correlations genuine and spurious in pearson and yule. *Statistical science*, pages 364–376, 1995.

Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. Multiobjective optimization. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 185–208, 2015.

Esteban Ballestar and Manel Esteller. Epigenetic gene regulation in cancer. *Advances in genetics*, 61:247–267, 2008.

Sally Bamford, Emily Dawson, Simon Forbes, Jody Clements, Roger Pettett, Ahmet Dogan, A Flanagan, Jon Teague, P Andrew Futreal, Michael R Stratton, et al. The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2):355–358, 2004.

Marouen Ben Guebila, Camila M Lopes-Ramos, Deborah Weighill, Abhijeet Rajendra Sonawane, Rebekka Burkholz, Behrouz Shamsaei, John Platig, Kimberly Glass, Marieke L Kuijjer, and John Quackenbush. Grand: a database of gene regulatory network models across human conditions. *Nucleic Acids Research*, 50(D1):D610–D621, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Rafael Bischof and Michael Kraus. Multi-objective loss balancing for physics-informed deep learning. *arXiv preprint arXiv:2110.09813*, 2021.

Ulrik Brandes. *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media, 2005.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

Maurizio Bruschi, Xhuliana Kajana, Andrea Petretto, Martina Bartolucci, Marco Pavanello, Gian Marco Ghiggeri, Isabella Panfoli, and Giovanni Candiano. Weighted gene co-expression network analysis and support vector machine learning in the proteomic profiling of cerebrospinal fluid from extraventricular drainage in child medulloblastoma. *Metabolites*, 12 (8):724, 2022.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Xiaodong Cai, Juan Andrés Bazerque, and Georgios B Giannakis. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology*, 9(5):e1003068, 2013.

Océane Cassan, Sophie Lèbre, and Antoine Martin. Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC genomics*, 22(1):387, 2021.

Dayanne M Castro, Nicholas R De Veaux, Emily R Miraldi, and Richard Bonneau. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS computational biology*, 15(1):e1006591, 2019.

Chen Chen and Megha Padi. Joint inference of transcription factor activity and context-specific regulatory networks. *bioRxiv*, pages 2022–12, 2022.

Jiaxing Chen, ChinWang Cheong, Liang Lan, Xin Zhou, Jiming Liu, Aiping Lyu, William K Cheung, and Lu Zhang. Deepdrim: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell rna-seq data. *Briefings in bioinformatics*, 22(6):bbab325, 2021.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, and Eric P Xing. Smoothing proximal gradient method for general structured sparse regression. 2012.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep

multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, and Benoit Ballester. Remap 2018: an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic acids research*, 46(D1):D267–D275, 2018.

Lynda Chin, Jannik N Andersen, and P Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, 17(3): 297–303, 2011.

Andrea Clocchiatti, Elisa Cora, Yosra Zhang, and G Paolo Dotto. Sexual dimorphism in cancer. *Nature Reviews Cancer*, 16(5):330–339, 2016.

GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348 (6235):648–660, 2015.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

Anton Crombach and Paulien Hogeweg. Evolution of evolvability in gene regulatory networks. *PLoS computational biology*, 4(7):e1000112, 2008.

Nagiua Cuomo-Haymour, Giorgio Bergamini, Giancarlo Russo, Luka Kulic, Irene Knuesel, Roland Martin, André Huss, Hayrettin Tumani, Markus Otto, and Christopher R Pryce. Differential expression of serum extracellular vesicle mirnas in multiple sclerosis: disease-stage specificity and relevance to pathophysiology. *International journal of molecular sciences*, 23 (3):1664, 2022.

Michiel JL De Hoon, Seiya Imoto, Kazuo Kobayashi, Naotake Ogasawara, and Satoru Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In *Biocomputing 2003*, pages 17–28. World Scientific, 2002.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Alvaro Del Real, Flor M Pérez-Campo, Agustín F Fernández, Carolina Sañudo, Carmen G Ibarbia, María I Pérez-Núñez, Wim Van Criekinge, Maarten Braspenning, María A Alonso, Mario F Fraga, et al. Differential analysis of genome-wide methylation and gene expression in mesenchymal stem cells of patients with fractures and osteoarthritis. *Epigenetics*, 12 (2):113–122, 2017.

Mark Drakesmith, Karen Caeyenberghs, A Dutt, G Lewis, AS David, and Derek K Jones. Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data. *Neuroimage*, 118: 313–333, 2015.

Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4): 778–789, 2021.

K Ruwani M Fernando and Chris P Tsokos. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (7):2940–2951, 2021.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Yao Fu, Laura R Jarboe, and Julie A Dickerson. Reconstructing genome-wide regulatory network of e. coli using transcriptome data and predicted transcription factor activities. *BMC bioinformatics*, 12:1–14, 2011.

Galen F Gao, Joel S Parker, Sheila M Reynolds, Tiago C Silva, Liang-Bo Wang, Wanding Zhou, Rehan Akbani, Matthew Bailey, Saianand Balu, Benjamin P Berman, et al. Before and after: comparison of legacy and harmonized tcga genomic data commons' data. *Cell systems*, 9(1):24–34, 2019.

Shang Gao, Yang Dai, and Jalees Rehman. A bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes. *Genome Research*, 31(7):1296–1311, 2021.

Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, 29(8):1363–1375, 2019.

Ryan Gill, Somnath Datta, and Susmita Datta. Differential network analysis in human cancer research. *Current pharmaceutical design*, 20(1):4–10, 2014.

Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832, 2013.

Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678, 2020.

Dan Grandér. How do mutated oncogenes and tumor suppressor genes cause cancer? *Medical oncology*, 15:20–26, 1998.

Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

Maxim Grechkin, Benjamin A Logsdon, Andrew J Gentles, and Su-In Lee. Identifying network perturbation in cancer. *PLoS computational biology*, 12 (5):e1004888, 2016.

Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, 2013.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

Stephen E Halford and John F Marko. How do site-specific dna-binding proteins find their targets? *Nucleic acids research*, 32(10):3040–3052, 2004.

Lide Han and Jun Zhu. Using matrix of thresholding partial correlation coefficients to infer regulatory network. *Biosystems*, 91(1):158–165, 2008.

Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1–17, 2012.

Intekhab Hossain, Viola Fanfani, John Quackenbush, and Rebekka Burkholz. Biologically informed neuralodes for genome-wide regulatory dynamics. *bioRxiv*, pages 2023–02, 2023.

Sui Huang. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of molecular medicine*, 77(6):469–480, 1999.

Edward L Huttlin, Raphael J Bruckner, Jose Navarrete-Perea, Joe R Cannon, Kurt Baltier, Fana Gebreab, Melanie P Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040, 2021.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

Luis F Iglesias-Martinez, Barbara De Kegel, and Walter Kolch. Kboost: a new method to infer gene regulatory networks from gene expression data. *Scientific Reports*, 11(1):15461, 2021.

Anisha S Jain, Ashwini Prasad, Sushma Pradeep, Chandan Dharmashekar, Raghu Ram Achar, Ekaterina Silina, Victor Stupin, Raghavendra G Amachawadi, Shashanka K Prasad, R Pruthvish, et al. Everything old is new again: Drug repurposing approach for non-small cell lung cancer targeting mapk signaling pathway. *Frontiers in Oncology*, 11:741326, 2021.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

Nikola Kasabov. Knowledge-based neural networks for gene expression data analysis, modelling and profile discovery. *Drug Discovery Today: BIOSILICO*, 2(6):253–261, 2004.

Kishan Kc, Rui Li, Feng Cui, Qi Yu, and Anne R Haake. Gne: a deep learning framework for gene network inference by aggregating biological information. *BMC systems biology*, 13(2):1–14, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Juho AJ Kontio, Marko J Rinta-Aho, and Mikko J Sillanpää. Estimating linear and nonlinear gene coexpression networks by semiparametric neighborhood selection. *Genetics*, 215(3):597–607, 2020.

Emmanuel N Kontomanolis, Antonios Koutras, Athanasios Syllaios, Dimitrios Schizas, Aikaterini Mastoraki, Nikolaos Garmpis, Michail Diakosavvas, Kyveli Angelou, Georgios Tsatsaris, Athanasios Pagkalos, et al. Role of oncogenes and tumor-suppressor genes in carcinogenesis: a review. *Anticancer research*, 40(11):6009–6015, 2020.

Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. *BioRxiv*, page 060012, 2016.

Zafer Koşar and Aykut Erbaş. Can the concentration of a transcription factor affect gene expression? *Frontiers in Soft Matter*, page 6, 2022.

Kimberly R Kukurba, Princy Parsana, Brunilda Balliu, Kevin S Smith, Zachary Zappala, David A Knowles, Marie-Julie Favé, Joe R Davis, Xin Li, Xiaowei Zhu, et al. Impact of the x chromosome and sex on regulatory variation. *Genome research*, 26(6):768–777, 2016.

Gloria Inés Lafaurie, Diana Marcela Castillo, Margarita Iniesta, Mariano Sanz, Luz Amparo Gómez, Yormaris Castillo, Roquelina Pianeta, Nathaly Andrea Delgadillo, Yineth Neuta, David Diaz-Báez, et al. Differential analysis of culturable and unculturable subgingival target microorganisms according to the stages of periodontitis. *Clinical Oral Investigations*, pages 1–15, 2023.

Harri Lähdesmäki, Ilya Shmulevich, and Olli Yli-Harja. On learning gene regulatory networks under the boolean network model. *Machine learning*, 52(1-2):147, 2003.

Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

David Latchman. *Gene regulation*. Taylor & Francis, 2007.

David S Latchman. Transcription factors: an overview. *International journal of experimental pathology*, 74(5):417, 1993.

Jae-Han Lee and Chang-Su Kim. Multi-loss rebalancing algorithm for monocular depth estimation. In *European Conference on Computer Vision*, pages 785–801. Springer, 2020.

Junyi Li, Yi-Xue Li, and Yuan-Yuan Li. Differential regulatory analysis based on coexpression network in cancer research. *BioMed research international*, 2016, 2016.

Hua Liang, Miaoning Gu, Chengxiang Yang, Hanbing Wang, Xianjie Wen, and Qiaoling Zhou. Sevoflurane inhibits invasion and migration of lung cancer cells by inactivating the p38 mapk signaling pathway. *Journal of anesthesia*, 26:381–392, 2012.

Kuo-Ching Liang and Xiaodong Wang. Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008:1–14, 2008.

Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. Btr: training asynchronous boolean models using single-cell expression data. *BMC bioinformatics*, 17(1):1–18, 2016.

Liyang Liu, Yi Li, Zhanghui Kuang, J Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. iclr, 2021.

Wei Liu, Yi Jiang, Li Peng, Xingen Sun, Wenqing Gan, Qi Zhao, and Huanrong Tang. Inferring gene regulatory networks using the improved markov blanket discovery algorithm. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–14, 2022.

Camila M Lopes-Ramos, Bruna P Barros, Fernanda C Koyama, Paola A Carpinetti, Julia Pezuk, Nayara TS Doimo, Angelita Habr-Gama, Rodrigo O Perez, and Raphael B Parmigiani. E2f1 somatic mutation within mirna target site impairs gene regulation in colorectal cancer. *PLoS One*, 12(7):e0181153, 2017.

Camila M Lopes-Ramos, Marieke L Kuijjer, Shuji Ogino, Charles S Fuchs, Dawn L DeMeo, Kimberly Glass, and John Quackenbush. Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer research*, 78(19):5538–5547, 2018.

Camila M Lopes-Ramos, Cho-Yi Chen, Marieke L Kuijjer, Joseph N Paulson, Abhijeet R Sonawane, Maud Fagny, John Platig, Kimberly Glass, John Quackenbush, and Dawn L DeMeo. Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell reports*, 31(12): 107795, 2020.

Sheng-Chieh Lu, Christine L Swisher, Caroline Chung, David Jaffray, and Chris Sidey-Gibbons. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Frontiers in Oncology*, 13:780, 2023.

Zhe Lu, Lei Ding, Heng Hong, John Hoggard, Qun Lu, and Yan-Hua Chen. Claudin-7 inhibits human lung cancer cell migration and invasion through erk/mapk signaling pathway. *Experimental cell research*, 317(13):1935–1946, 2011.

Cynthia Z Ma and Michael R Brent. Inferring tf activities and activity regulators from gene expression data with constraints from tf perturbation data. *Bioinformatics*, 37(9):1234–1245, 2021.

Omar Mahmood, Claudia Skok Gibbs, Richard Bonneau, and Kyunghyun Cho. Probabilistic matrix factorization for gene regulatory network inference. *bioRxiv*, pages 2022–09, 2022.

Itzik Malkiel and Lior Wolf. Mtadam: Automatic balancing of multiple training loss terms. *arXiv preprint arXiv:2006.14683*, 2020.

Jessica C Mar. The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond. *Biophysical reviews*, 11 (1):89–94, 2019.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

Vincentius Martin, Jingkang Zhao, Ariel Afek, Zachery Mielko, and Raluca Gordân. Qbic-pred: quantitative predictions of transcription factor bind-

ing changes due to sequence variants. *Nucleic acids research*, 47(W1):W127–W135, 2019.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Peter Melchior, Rémy Joseph, and Fred Moolekamp. Proximal adam: robust adaptive update scheme for constrained optimization. *arXiv preprint arXiv:1910.10094*, 2019.

Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.

Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007:1–9, 2007.

Sara Movahedi, Yves Van de Peer, and Klaas Vandepoele. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant physiology*, 156(3):1316–1330, 2011.

Yanina Natanzon, Ellen L Goode, and Julie M Cunningham. Epigenetics in ovarian cancer. In *Seminars in cancer biology*, volume 51, pages 160–169. Elsevier, 2018.

Crystal M North and David C Christiani. Women and lung cancer: what is new? In *Seminars in thoracic and cardiovascular surgery*, volume 25, pages 87–94. Elsevier, 2013.

Theofilou Paraskevi. Quality of life outcomes in patients with breast cancer. *Oncology reviews*, 6(1), 2012.

Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.

Jyoti D Patel. Lung cancer in women. *Journal of Clinical Oncology*, 23(14): 3212–3218, 2005.

Nihir Patel and Jason TL Wang. Semi-supervised prediction of gene regulatory networks using machine learning algorithms. *Journal of biosciences*, 40:731–740, 2015.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

Francesca Petralia, Pei Wang, Jialiang Yang, and Zhidong Tu. Integrative random forest for gene regulatory network inference. *Bioinformatics*, 31 (12):i197–i205, 2015.

Tata Pramila, Wei Wu, Shawna Miles, William Stafford Noble, and Linda L Breeden. The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle. *Genes & development*, 20(16):2266–2278, 2006.

Mark Ptashne and Alexander Gann. *Genes & signals*, volume 402. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY:, 2002.

Kriti Puniyani and Eric P Xing. Gini: From ish images to gene interaction networks. *PLoS computational biology*, 9(10):e1003227, 2013.

Ramesh Ram and Madhu Chetty. A markov-blanket-based model for gene regulatory network inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):353–367, 2009.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7): e47–e47, 2015.

R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Erina Sakamoto and Hitoshi Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 720–726. IEEE, 2001.

Guido Sanguinetti, Vân Anh Huynh-Thu, et al. Gene regulatory networks. *Springer, New York, NY*, 10:978–1, 2019.

Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

Daniel Schlauch, Kimberly Glass, Craig P Hersh, Edwin K Silverman, and John Quackenbush. Estimating drivers of cell state transitions using gene regulatory network models. *BMC systems biology*, 11(1):1–10, 2017.

Harsh Shrivastava, Xiuwei Zhang, Le Song, and Srinivas Aluru. Grnular: A deep learning framework for recovering single-cell gene regulatory networks. *Journal of Computational Biology*, 29(1):27–44, 2022.

Katherine H Shutta, Deborah Weighill, Rebekka Burkholz, Marouen Ben Guebila, Dawn L DeMeo, Helena U Zacharias, John Quackenbush, and Michael Altenbuchinger. Dragon: determining regulatory associations using graphical models on multi-omic networks. *arXiv preprint arXiv:2104.01690*, 2021.

Arun J Singh, Stephen A Ramsey, Theresa M Filtz, and Chrissa Kioussi. Differential gene regulatory networks in development and disease. *Cellular and Molecular Life Sciences*, 75:1013–1025, 2018.

Aarathi Sugathan and David J Waxman. Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. *Molecular and cellular biology*, 33(18): 3594–3610, 2013.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4 US DOE Joint Genome Institute: Hawkins Trevor, Branscomb Elbert, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921, 2001.

Wim Van Criekinge and Rudi Beyaert. Yeast two-hybrid: state of the art. *Biological procedures online*, 2:1–38, 1999.

Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional clas-

sification and gene–disease predictions. *Briefings in bioinformatics*, 19(4): 575–592, 2018.

Victor E Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.

Guohua Wang, Fang Wang, Qian Huang, Yu Li, Yunlong Liu, and Yadong Wang. Understanding transcription factor regulation by integrating gene expression and dnase i hypersensitive sites. *BioMed research international*, 2015, 2015.

Juexin Wang, Anjun Ma, Qin Ma, Dong Xu, and Trupti Joshi. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. *Computational and structural biotechnology journal*, 18:3335–3343, 2020.

Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

Deborah Weighill, Marouen Ben Guebila, Camila Lopes-Ramos, Kimberly Glass, John Quackenbush, John Platig, and Rebekka Burkholz. Gene regulatory network inference as relaxed graph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10263–10272, 2021.

Deborah Weighill, Marouen Ben Guebila, Kimberly Glass, John Quackenbush, and John Platig. Predicting genotype-specific gene regulatory networks. *Genome Research*, 32(3):524–533, 2022.

Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6): 1431–1443, 2014.

James West, Ginestra Bianconi, Simone Severini, and Andrew E Teschendorff. Differential network entropy reveals cancer system hallmarks. *Scientific reports*, 2(1):1–8, 2012.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Siqi Wu, Antony Joseph, Ann S Hammonds, Susan E Celniker, Bin Yu, and Erwin Frise. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences*, 113(16):4290–4295, 2016.

Linlin Xu, Lingchen Wang, and Minzhang Cheng. Identification of genes and pathways associated with sex in non-smoking lung cancer population. *Gene*, 831:146566, 2022.

Yang Yang, Qingwei Fang, and Hong-Bin Shen. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLoS computational biology*, 15(9):e1007324, 2019.

Qiong Zhang, Wei Liu, Hong-Mei Zhang, Gui-Yan Xie, Ya-Ru Miao, Mengxuan Xia, and An-Yuan Guo. htftarget: a comprehensive database for regulations of human transcription factors and their targets. *Genomics, proteomics & bioinformatics*, 18(2):120–128, 2020.

Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104, 2012.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

BIOLOGICALLY INFORMED MATRIX FACTORIZATION FOR JOINT INFERENCE OF GENE REGULATORY NETWORKS AND TRANSCRIPTION FACTOR ACTIVITIES

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
MICHELETTI	SOEL ANDREA

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Minusio, 21st April 2023	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.