UNIVERSITY OF PADOVA

DEPARTMENT OF SURGICAL ONCOLOGICAL
AND GASTROENTEROLOGICAL SCIENCES

CURRICULUM IN BIOINFORMATICS & COMPUTATIONAL BIOLOGY

CYCLE XXXVI

ALLSTAR: A Novel Bioinformatic Algorithm to Infer Causal Rules between Somatic Mutations and Cancer Phenotypes

---

PhD Candidate:
Antonio Collesei

Coordinator & Supervisor:
Prof. Stefano Indraccolo

Co-supervisor:
Prof. Fabio Vandin

Thesis written with the financial contribution of
Veneto Institute of Oncology IOV-IRCCS

February 28, 2024

UNIVERSITY OF PADOVA

Abstract

Department of Surgical Oncological
and Gastroenterological Sciences

Doctor of Philosophy

***ALLSTAR*: A Novel Bioinformatic Algorithm to Infer Causal Rules between Somatic Mutations and Cancer Phenotypes**

by Antonio Collesei

Motivation: Recent advances in DNA sequencing technologies have allowed the detailed characterization of genomes in large cohorts of tumors, highlighting their extreme heterogeneity, with no two tumors sharing the same complement of somatic mutations. Such heterogeneity hinders our ability to identify somatic mutations important for the disease, including mutations that determine clinically relevant phenotypes (e.g., cancer subtypes). Several tools have been developed to identify somatic mutations related to cancer phenotypes. However, such tools identify correlations between somatic mutations and cancer phenotypes, with no guarantee of highlighting causal relations.
Results: This thesis is centered around *ALLSTAR*, a novel tool I developed as a result of a joint collaboration between the Veneto Institute of Oncology and the Department of Information Engineering at the University of Padova. The tool is able to infer reliable causal relations between combinations of somatic mutations and cancer phenotypes. *ALLSTAR* ranks causal rules based on the highest impact in terms of average effect on the phenotype. Since proving that the underlying computational problem is NP-hard, I developed a branch-and-bound approach, employing protein-protein interaction networks and novel bounds for pruning the search space, while properly correcting for multiple hypothesis testing. The extensive experimental evaluation on synthetic data shows that *ALLSTAR* is able to identify reliable causal relations in large cancer cohorts. Moreover, the reliable causal rules identified in cancer data show that my approach is able to retrieve several somatic mutations known to be relevant for cancer phenotypes, as well as novel biologically meaningful relations.
Availability and Implementation: Code, data, and scripts to reproduce the experiments are available at https://github.com/VandinLab/ALLSTAR.

Acknowledgements

Even if it comes first in the order of the thesis, this chapter is always the last one person wants to write, because it is the hardest of all. In my case, it is no exception. It is the hardest piece to write because it represents the awareness that a significant period of your life has come to an end, and you feel responsible to give credit to every single person that made it so unique. I will try not to make it a boring list, more like a short story which can fit everyone in. Well, let's jump right into it.

After my Master's graduation, I honestly was not into pursuing a research career. I looked for industry jobs, and landed a position in a biomedical company, based in Padova, but with lots of international commercial contacts. Practically, my role was more customer-related than research-oriented. I had wonderful colleagues and I often had the chance to travel, but I felt that there was no learning curve. I wanted something more suited for my thirst for knowledge, and I must thank my parents Cinzia and Italo, who made me realize this. They did more: they also sponsored an additional course to strengthen my data analysis skills in omics data. I enjoyed it, and obviously wanted to apply my new expertise. For this reason, I preliminarily scouted research positions in industry, but the ones I liked the most required a Ph.D.: so the idea of starting this journey, after having been far from the academic world for a couple of years, began to pervade my mind.
It was only during the COVID-19 pandemic that I put on all my efforts to find a doctoral program that could tick all my boxes. Despite the tragedy that was unfolding upon each one of us, I tried to get the most out of it, and managed to obtain a Ph.D. position at the Veneto Institute of Oncology, in collaboration with the Department of Oncology (DiSCOG) and Information Engineering (DEI) at the University of Padova. And I would like to express my gratitude to my brother Eugenio, that, among all things, halped me get through the isolation period with game nights, chats, laughs, and fights.

I am extremely grateful to Prof. Paola Zanovello, Prof. Stefano Indraccolo, and Prof. Giuseppe Opocher, who had the vision to trust an engineer for this new position, despite them being oriented to biomedical research. And of course my biggest thank goes to Prof. Fabio Vandin, who welcomed me into his research group, mentored me during the development of the *ALLSTAR* project with his intuition, and *teased* me in the right way to push me beyond my limits. He invited me to stimulating journal clubs with brilliant people: Leonardo, Andrea, Davide, Diego, they are all able to raise the bar, when it comes to scientific discussion; I especially grew a stronger bond with Dario Simionato (co-author of *ALLSTAR*, fellow pro basketball player) and Ilie Sarpe (founder of the BBQ competition), with both of whom I shared awesome moments during, and, especially, after office hours.

Talking about office, during this Ph.D., I had a truly multi-disciplinary experience by being given a seat also at the core of the genetic data production lab of the Veneto Institute of Oncology. At the Hereditary Tumor facility, I met Francesca, Silvia, Claudia, Veronica, and Fly, who tried to make me feel at home despite our backgrounds were apparently divided by an incurable

rift. And since I got a stable research position at the Institute, I can learn more out of this weird, but somehow successful, dynamic.

Since I consider myself a social animal, where would I be without my friends? I managed to surround myself with a selected, stimulating and diverse group of people, each one of whom has been a source of inspiration and support. I will not name them one by one, but I am sure they know which collective they fit in. I thank the Canederli for the drinks at Gottino, the Hidden Embers for the sleepless nights, the Trapattoni for the couch sessions, and my teammates for the joys and tears (just like a PhD).

Last but not least, words cannot express my gratitude to Laura, who has had multiple roles throughout this journey. Co-author, flatmate, number 1 fan, she also happens to be my girlfriend, and she is very good at it. I won't say she is perfect, because she did not invite me as collaborator to her published paper on Nature, but she is special. We share an unsatisfiable thirst for knowledge, and we both put all we have on the table when it comes to our passions. She managed to apply this same attitude troughout my doctoral path as well, supporting me with curiosity and critical mind. I could not have asked for more.

In these past three years, I traveled, shared meals and thoughts with people from every continent, presented my research in front of hundreds of listeners, won an award, published a few papers, and even got elected as student representative. Luckily, I see this thesis not as a landing point, but as a trampoline. There will be more to come: paraphrasing an old saying, *my most beautiful goal is the next one*.

For the moment, I hope you enjoy this piece of writing, as much as I had fun developing *ALLSTAR* with Dario, Federica, and Fabio.

Contents

List of Figures

Chapter 1

Introduction

In the last ten years, the advances in DNA sequencing technologies have allowed to precisely depict the landscape of somatic alterations in large cohorts of tumors for various cancer types. Projects such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the International Cancer Genome Consortium (ICGC) (The International Cancer Genome Consortium, 2010) provide valuable resources to identify somatic alterations directly related to tumor development and evolution (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). The amount of available data is increasing year after year, with the clear intent of promoting data analysis to sustain progress in the field of computational oncology.

Tumor Heterogeneity

The explosion of available data is, in fact, necessary to allow the observation of the outstanding variability of cancer. It is clearly understandable that intrinsic differences between various cancer types exist, because of the particular biological processes associated to the site of development. For this reason, cancer is well-known as an heterogeneous disease.
However, an additional peculiarity is the presence of variability also among same-type cancers and across different patients. In fact, the study of the alteration landscape in large cohorts has shown that cancer is characterized by various layers of heterogeneity. The term *intra*-tumor heterogeneity refers to the existence of many cancer cell clones within a single tumor mass, whereas *inter*-tumor heterogeneity describes the occurrence of distinct genetic alterations in metastatic tumors from a single patient. There is growing evidence showing that mutational subclones, within the same tumor, present individual characteristics, suggesting that tumor sub-classification based on heterogeneity *tout-court* is not a straightforward task.
Thus, there is a strong necessity for a deeper investigation to improve patient stratification and consequent response to therapy.

Correlation is not Causation

A number of computational tools have been designed to try to identify the alterations that drive the insurgence and development of tumours while tackling inter-tumor heterogeneity (Cortés-Ciriano et al., 2022). These tools are based on the detection of various types of signals (Cibulskis et al., 2013;

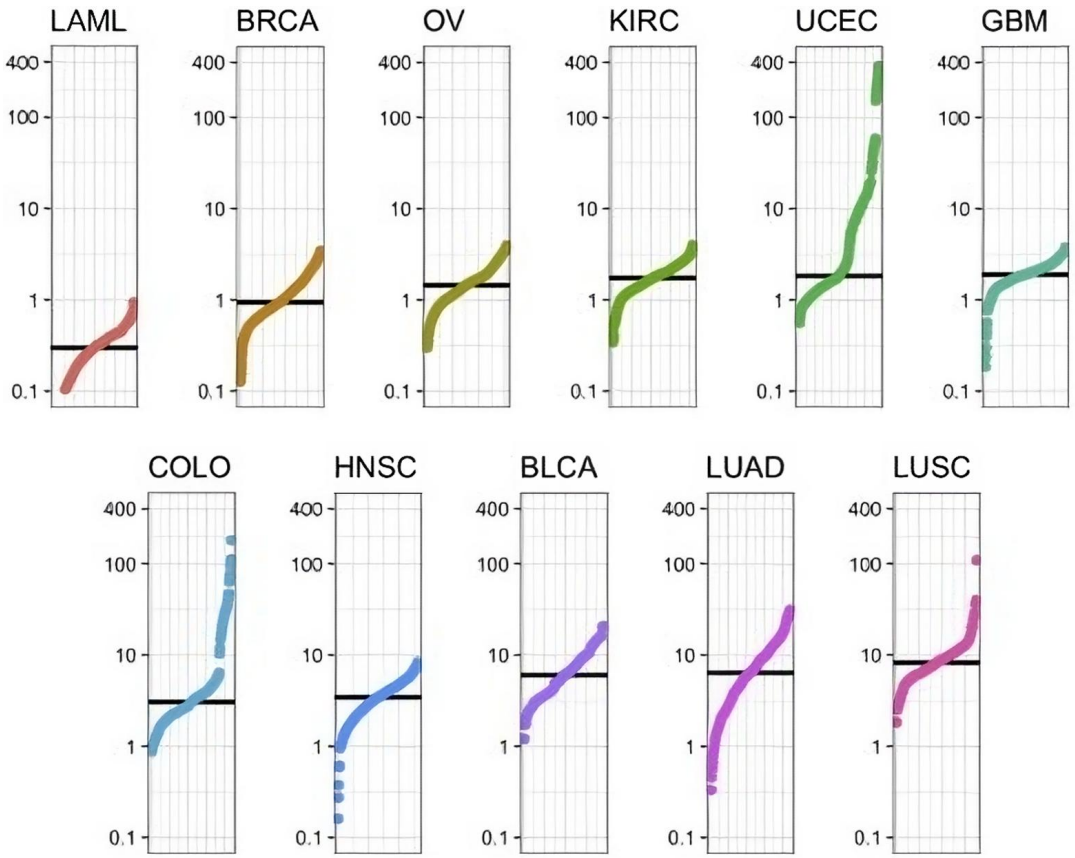

FIGURE 1.1: From Piraino et al. (2019), a visual quantification of tumor heterogeneity, showing mutation rates per megabase across different cancer types. The black bar is the median mutation rate of each cancer type. There is considerable variation in mutation rates both within and between cancer types. BRCA, breast adenocarcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; UCEC, uterine corpus endometrial carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; COLO, colon and rectal carcinoma; BLCA, bladder urothelial carcinoma; KIRC, kidney renal clear cell carcinoma; OV, ovarian serous carcinoma; LAML, acute myeloid leukemia.

Vandin, Upfal, and Raphael, 2012; Mularoni et al., 2016; Arnedo-Pac et al., 2019) and the integration of different prior and/or clinical information (Cowen et al., 2017; Reyna et al., 2020; Sarto Basso, Hochbaum, and Vandin, 2019), but a common feature of these tools is that they detect alterations that are *correlated* with cancer phenotypes. That is, they identify alterations, or groups of alterations, that are significantly enriched in a group of patients or significantly associated with a (clinical) phenotype.

While the identification of alterations correlated with cancer phenotypes provides interesting insights into cancer initiation and progression, it does not guarantee that *causal* relations between somatic mutations and cancer are reported. In fact, correlation is *not* causation. The co-occurrence of two events can be an indicator of some sort of relationship between them, but it definitely cannot imply a consequential dependency of one another. However, it is very easy to fall into the trap, because the correlation is undeniably strong, or simply because it feels logical.
For example, the correlation value between obesity and Type-2 diabetes is high, therefore one could conclude that being obese causes the pathology. Nevertheless, it is more likely that both conditions are caused by common factors, such as sedentary lifestyle, or an unhealthy diet. A graphical representation of the most-likely relationship between these variables can be found in Figure 1.2.
Although researchers are usually very careful when stating their correlation-based results do not suggest a causal relationship, a superficial interpretation or report to a common public, as it usually happens with medical findings and newspapers, may lead to large-scale misunderstandings. A certain food can quickly become *anti-cancer* if its consumers tend not to develop the disease. Analogously, coming back to the main topic of this thesis, a gene or alteration can be labeled as leading causes of a certain cancer type, if their correlation is strong enough. Hence it comes the necessity to define causality and a solid framework to investigate it.
While experimental and clinical validation is a necessary step to demonstrate the significance of alterations, tools reporting causal relations with guarantees on the quality of their findings would greatly reduce the resources needed to identify relevant alterations in follow-up experimental and clinical studies.

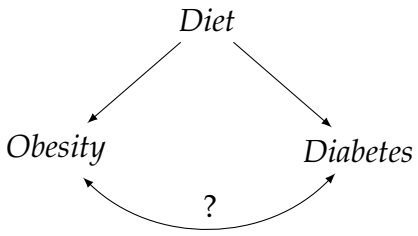

FIGURE 1.2: Graphical representation of a common misunderstanding: obesity does not cause diabetes, since they are both caused by unhealthy lifestyle and diet.

ALLSTAR's Contributions

This thesis describes *ALLSTAR*, a novel tool to identify reliable causal relations between somatic mutations and cancer phenotypes. *ALLSTAR* identifies causal relations in the form of *rules* highlighting combinations of mutations with the highest average effect on the phenotype. In this regard, its main contributions are fourfold and they will be extensively described in Chapters 3, 5 and 6.
Firstly, it will be proven the underlying computational problem is NP-hard.
Secondly, the necessity to properly correct for multiple hypothesis testing when identifying *reliable* causal rules will be highlighted.
Thirdly, *ALLSTAR* itself will be characterized as an effective branch-and-bound algorithm to identify the k rules with the highest reliable average effect on the phenotype, with guarantees on the family-wise error rate (FWER) of the output. *ALLSTAR* identifies rules where genes are connected in a large interaction graph provided in input, and employs an iterative procedure leading to the identification of diverse rules, which highlight different causal relations potentially linked to cancer heterogeneity.
Fourthly, an extensive evaluation of *ALLSTAR* on both synthetic data and real-world cancer data will be depicted. The results show that *ALLSTAR* is effective in identifying causal rules associated with the phenotype and that it reports well-supported as well as potentially novel causal relations between somatic mutations and cancer phenotypes.

Chapter 2

Preliminaries

The previous chapter had the objective to introduce and motivate the necessity to develop a causal tool in the field of computational oncology. This chapter aims to provide the reader with the necessary background to understand the theoretical framework of *ALLSTAR*.

This thesis presents an algorithm based on a solid mathematical background. While the notation could be heavy to read, this chapter provides the preliminary knowledge that is necessary to comprehend the pillars of causality. Whereas possible, I will try to break down each notation into simple and intuitive concepts to bridge the gap between theory and practical implementation. Section 2.1 introduces the state-of-the-art definition of causal rules, Section 2.2 their application to observational datasets, while Section 2.3 describes the improvement given by the introduction of the *reliable* causal framework, which has been adopted from Budhathoki, Boley, and Vreeken (2021), as a starting point for this thesis.

2.1 Causal Rules

Let $\mathcal{D}$ be a dataset of observations for two sets of variables $\mathbf{X} = \{X_1, ..., X_n\}$ and $\mathbf{Z} = \{Z_1, ..., Z_m\}$, and for a target variable Y. $\mathbf{X}$ is the set of features of interest, for which *causal* relations with respect to the target variable Y need to be identified, while $\mathbf{Z}$ is the set of confounder variables. In this thesis the focus is on *rules* $\sigma = \pi_1 \wedge \pi_2 \wedge ... \wedge \pi_\ell$, defined as conjunctions of propositions π_i on the set of variables $\mathbf{X}$ (e.g. $\pi_1 \equiv X_3 = 1$, $\pi_2 \equiv X_5 = 1$, and $\sigma = \pi_1 \wedge \pi_2$). In other words, combinations of features of interest $\mathbf{X}$, each taking a specific value, contribute to the definition of rules σ. Each rule refers to specific chunks of observations in the dataset $\mathcal{D}$, when propositions are verified.
A rule σ is *true* ($\top$) for an assignment $\mathbf{x} = \{x_1, ..., x_n\}$ if every proposition π_i in σ is verified under the assignment $\mathbf{x}$, and it is *false* ($\bot$) otherwise.

The general objective of rules' evaluation is to calculate (or, at least, estimate) their effect on the target. Algorithms based on association, or general dependence, between rule and target measure the effect by taking into account the observed conditional distribution,

$$P(Y|\sigma = \top) = \sum_{\sigma(\mathbf{x})=\top} P(Y|\mathbf{X} = \mathbf{x}). \tag{2.1}$$

In this thesis, and for *ALLSTAR*'s core, among various definitions of causality, I consider causality in the context of the do-calculus (Pearl, 2009) for the identification of causal relations in non-parametric models.
Let $do(X_i = x_i)$ be the **atomic intervention** operator (Pearl, 2009), which changes the value of the variable X_i to x_i while keeping the values of all the other variables fixed. Informally, I am interested in finding rules, potentially including multiple variables of interest, which cause a specific change in the value of the target Y when (atomic) interventions are performed. More formally, a rule σ defines a so-called **stochastic policy** Q_σ, i.e., a probability distribution over the interventions (see Budhathoki, Boley, and Vreeken, 2021 for more details), which integrates $do(X = x)$ operations. Thus, the post-intervention distribution of Y under the stochastic policy Q_σ is formally defined as follows:

$$P(Y|do(Q_\sigma)) = \sum_{\sigma(\mathbf{x})=\top} P(Y|do(\mathbf{X} = \mathbf{x}))Q_\sigma(do(\mathbf{X} = \mathbf{x})). \tag{2.2}$$

As stated above, the main interest is the **effect** of a **causal rule** on Y taking value y, which is defined as:

$$e^y(\sigma) = p(Y = y|do(Q_\sigma)) - p(Y = y|do(Q_{\bar{\sigma}})) \tag{2.3}$$

where p represents the probability mass function and $\bar{\sigma}$ the equality $\sigma = \perp$.

$e^y(\sigma)$ can take value in the interval $[-1, 1]$, and, intuitively, a positive value means that the rule σ has a positive causal influence on Y. A positive effect is interpretable as as an increase in the probability that Y takes value y when the rule σ is true, compared to any other configuration (i.e., where σ is false).

2.2 Observational Data and Causal Effect

Randomized controlled trials (RCT) are the gold standard to determine whether a variable of interest is causal with respect to a target. However, RCT are often rather expensive and/or impractical. Besides that, an increasing amount of observational datasets are becoming more and more available over the years, suggesting the necessity to develop algorithms able to exploit them.

In the previous section, I have already provided the definitions of observed conditional distribution and post-interventional distributions. These two entities tend to differ when evaluating observational datasets, due to the presence of confounding variables $\mathbf{Z}$, or simply confounders, that influence both intervention variables $\mathbf{X}$ and the target Y. From observational data, it is only possible to estimate observational distributions (i.e., $p(Y|X_i = x_i)$) and not interventional distributions (i.e., $p(Y|do(X_i = x_i))$) as in RCT, but the two are equivalent if *spurious correlations* between $\mathbf{X}$ and Y are removed.

To understand this concept, it is worth moving the discussion to a graph-based domain. A *graph*, or network, is a set of objects (*nodes*) that are connected together. The connections between the vertices are called *edges*. Edges

can be directed or undirected. Directed edges are interpretable as causal dependencies between two nodes.

Formally, a Bayesian network (BN) is defined as a tuple $< \mathcal{G}, p >$ where $\mathcal{G} =< \mathbf{V}, \mathbf{E} >$ is a directed acyclic graph for which $\mathbf{V} = \mathbf{X} \cup \mathbf{Z} \cup \{Y\}$ and there is an edge from $V_i \in \mathbf{V}$ to $V_j \in \mathbf{V}$ only if V_i is a cause of V_j w.r.t. Pearl's do-notation (Pearl, 2009), and p is a probability distribution function over $\mathbf{V}$. Any undirected path with an incoming edge towards $\mathbf{X}$ connecting Y and $\mathbf{X}$ is a *spurious path* when assessing the causal influence of $\mathbf{X}$ on Y. As proved in Budhathoki, Boley, and Vreeken (2021), spurious correlations between $\mathbf{X}$ and Y are removed in BNs that satisfy the *admissible input structure* for causal rule discovery, defined by the following constraints:

1. there are no directed edges from Y to any $X_i \in \mathbf{X}$;
2. there are no outgoing edges from any $X_i \in \mathbf{X}$ to any $Z_j \in \mathbf{Z}$;
3. there are no edges between variables $X_i \in \mathbf{X}$;
4. there are no edges between $X_i \in \mathbf{X}$ and any unobserved variable U.

In simpler terms, the target variable Y cannot be a cause of any $X_i \in \mathbf{X}$, none of the variables $X_i \in \mathbf{X}$ can be a cause of any $Z_j \in \mathbf{Z}$or any other $X_j \in \mathbf{X}$, and there cannot be unobserved variable U that directly cause $X_i \in \mathbf{X}$.

When the constraints of the admissible input structure are satisfied, observational and interventional probabilities are equal by conditioning on $\mathbf{Z}$, that is

$$p(Y|X_i = x_i, \mathbf{Z}) = p(Y|do(X_i = x_i), \mathbf{Z}) \tag{2.4}$$

and averaging the observational probabilities over $\mathbf{Z}$ gives:

$$p(Y|do(X_i = x_i)) = \sum_{\mathbf{z}} p(Y|X_i = x_i, \mathbf{Z} = \mathbf{z}) p(\mathbf{Z} = \mathbf{z}). \tag{2.5}$$

In other words, if the admissible input assumptions are met, $e^y(\sigma)$ measures the average treatment effect that the variables in σ exert on the event Y taking value y without including any spurious (i.e. non-causal) statistical correlation.

2.3 Reliable Causal Rules

The estimation of probabilities from data is challenging when sample sizes are small, as the estimates obtained with naïve empirical estimators have high variance. As a consequence, rules discovered by data using such naïve empirical estimators have effects whose estimates are far from their true effects. To mitigate this phenomenon, which may lead to overfitting , Budhathoki, Boley, and Vreeken (2021) proposes a *reliable* estimator for the effect of causal rules.

The *naïve* empirical estimator of $p(Y = y|\sigma = \top)$ is defined as

$$\hat{p}(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}}{n_{\sigma=\top}}, \tag{2.6}$$

where $n_{\sigma=\top}$ is the number of instances for which $\sigma = \top$ (i.e, σ is true), and $n_{Y=y,\sigma=\top}$ is the number of instances for which $Y = y$ and $\sigma = \top$. Analogously, the equality

$$\hat{p}(Y = y|\sigma = \bot) = \frac{n_{Y=y,\sigma=\bot}}{n_{\sigma=\bot}}$$

stands for the cases when the rule is not verified.
In extreme cases (e.g. $\sigma = \bot$ for all instances) such quantities are ill-defined, therefore the Laplace correction is applied to the estimated probability, which becomes

$$\hat{p}_c(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top} + 1}{n_{\sigma=\top} + 2}. \tag{2.7}$$

By considering all samples such that $\sigma = \top$ (and, respectively, $\sigma = \bot$), the value y is a binomial distribution with success probability $p(Y = y|\sigma = \top)$. For a given confidence level $\alpha \in (0, 1)$, by defining $\beta(\alpha)$ as the $1 - \alpha/2$ quantile of a standard normal distribution, the confidence bound for Equation 2.7 proposed by Budhathoki, Boley, and Vreeken (2021) is then

$$\left[\hat{p}_c(Y = y|\sigma = \top) - \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}}, \hat{p}_c(Y = y|\sigma = \top) + \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}}\right]. \tag{2.8}$$

Such bound allows to compute the effect of **reliable causal rules**, defined as the lower bound of the effect of causal rules. Since the main objective is always to retrieve a quantitative measurement of the causal relationship between variables of interest and a target, the estimated reliable effect $\hat{e}^y_{rel}(\sigma)$ of a causal rule σ on Y taking value y with confidence α is defined as:

$$\hat{e}^y_{rel}(\sigma, \alpha) = \hat{p}_c(Y = y|do(Q_\sigma)) - \hat{p}_c(Y = y|do(Q_{\bar{\sigma}})) - \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}} - \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\bot}}}. \tag{2.9}$$

In the following, I will refer to $\beta(\alpha)$ as β and to $\hat{e}^y_{rel}(\sigma, \alpha)$ as $\hat{e}_{rel}(\sigma)$ to improve readability.

In the next chapter, I will make clear how the causal framework has been improved and the algorithm *ALLSTAR* implemented.

Chapter 3

Algorithm and Methodology

In the previous chapter, I introduced the state-of-the-art mathematical knowledge that is necessary to understand the foundations of *ALLSTAR*. In this chapter, I take a step further by introducing the core methodology and the main contributions of the algorithm, as well as the translational field of application. *ALLSTAR* is the first tool able to find the ranked top-k rules $\sigma_1^*, ..., \sigma_k^*$ (e.g., combinations of alterations) with the highest positive reliable causal effect ($ATE > 0$) on a cancer phenotype Y with statistical guarantees on the result. This project is able to bring multiple advances to the state-of-the-art, which can be summarized as follows:

- Proof of NP-hardness for the general computational problem;
- Development of an efficient and scalable Branch-and-Bound algorithm;
- Statistical guarantees for Multiple Hypothesis Testing (MHT);
- Integration of aprioristic knowledge (protein-protein interaction network) to make the algorithm biologically-informed;
- Extensive evaluation of the algorithm performances through synthetic experiments (see Chapter 5);
- Application to Breast Cancer data and analysis of the result via oncological experts (see Chapter 6).

3.1 Translation of the Causal Framework

As it already has been made clear in the introductory chapters, the development of *ALLSTAR* has always had translational oncology as field of application. Towards this aim, the causal framework described in Chapter 2 needed both suitable data and meaningful biological questions, to exploit its full potential.
For the former point, I decided to work with publicly available datasets, to put the accent on the robustness of the algorithm and guarantee reproducibility of the results. Concerning the latter point, the causal framework offered the possibility to extract rules (i.e., combinations of features) scoring the highest causal effect on a relevant target. This suggested the idea of mining causal

combinations of somatic alterations with the highest effect on a specific cancer phenotype.
Variables **X**, **Z**, and Y are defined as follows:

- the set **X** of features includes somatic alterations (i.e., SNVs, loss of heterozygosity, hypermethylation) in a set of genes, and the observations are provided by a binary matrix describing the status (present or not) of such alterations in a cohort of patients;
- the set **Z** of confounders includes relevant germline mutations (in driver genes, BRCA1 and BRCA2) and clinical information (i.e. race, age, sex, history of previous malignancy, etc), and the observations are provided by a corresponding matrix of relevant clinical variables;
- the target Y is a phenotype of interest, such as histological or molecular marker-derived cancer subtypes.

Recalling Figure 1.2, Figure 3.1 describes the translational causal framework, focused on retrieving

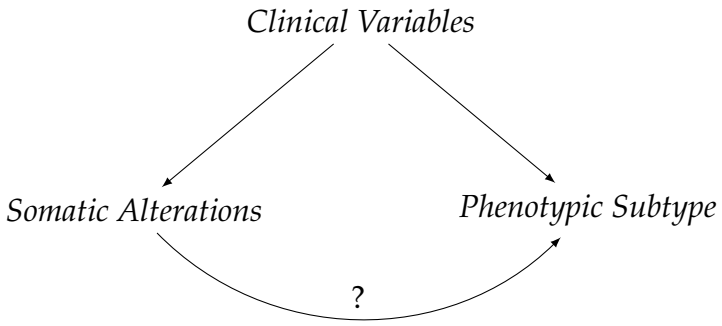

FIGURE 3.1: Graphical representation of the translational causal framework, applied to a cancer mutational settings.

Such a setting needs also a translation of the constraints required by an admissible input structure for causal rule discovery. They now recite as follows:

1. the target subtype variable Y does not cause somatic alterations;
2. there is no somatic alteration that is a cause of any clinical variable identified as confounder;
3. there are no causal relations between somatic alterations;
4. there are no causal relations between somatic alterations and relevant unobserved variables.

Assumptions 1, 2, and 4 are satisfied by a proper choice of target variable Y, of confounders **Z**, and the features **X** to include in the study. Assumption 3 is instead supported by the fact that somatic alterations arise as independent observations in the genome (even in normal cells), even if specific somatic

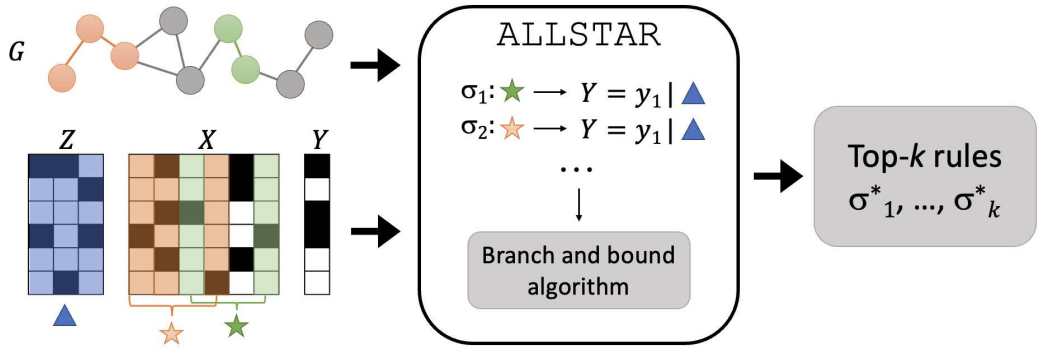

FIGURE 3.2: An illustration of *ALLSTAR* framework. From a dataset comprising a set of confounders **Z**, treatments **X**, and a target Y, *ALLSTAR* uses a branch and bound approach to discover the top-k rules $\sigma_1^*, ..., \sigma_k^*$ with the highest reliable causal effect. *ALLSTAR* exploits a gene-gene interaction network G to focus on biologically meaningful rules.

alterations may modify the overall distribution of alterations in the genome (e.g., due to their impact on processes involved in mutagenesis). In fact, according to current models of cancer as an accumulation of somatic mutations, such mutations are happening randomly, under an increasing state of instability, and the disease develops only when a sufficient number of alterations providing selective advantage has accumulated. In this sense, there is no causality between the appearance of somatic alterations, since all alterations appear as a result of a random process. Therefore, I believe that Assumption 3 is well justified. In such setup, each rule represents the observation of a specific set of gene alterations that occur simultaneously, and the rule effect is a measure of the influence of such pattern on having a specific cancer type.

3.2 The *ALLSTAR*'s Manifesto and Implementation

This section is meant to present the core of the algorithm *ALLSTAR* (reliable cAusaL ruLe discovery between Somatic muTations and cAnceR phenotypes) for causal rule discovery with guarantees on its results. The underlying problem of estimating the rule with highest causal effect is NP-hard even if the probability distributions are known a priori. This means that there is no known algorithm able to solve this problem in polynomial time, thus the solution can only be retrieved in exponential time. The proof of NP-hardness, formulated by Dario Simionato who is one of the contributors to this project, is available in the Appendix A. As a consequence, it is necessary to optimize the algorithm in order to reduce the complexity: for example, *ALLSTAR* exploits a gene-gene interaction network to focus on sets of functionally related genes and prune the search space. The pseudocode of *ALLSTAR* is in Algorithm 1.

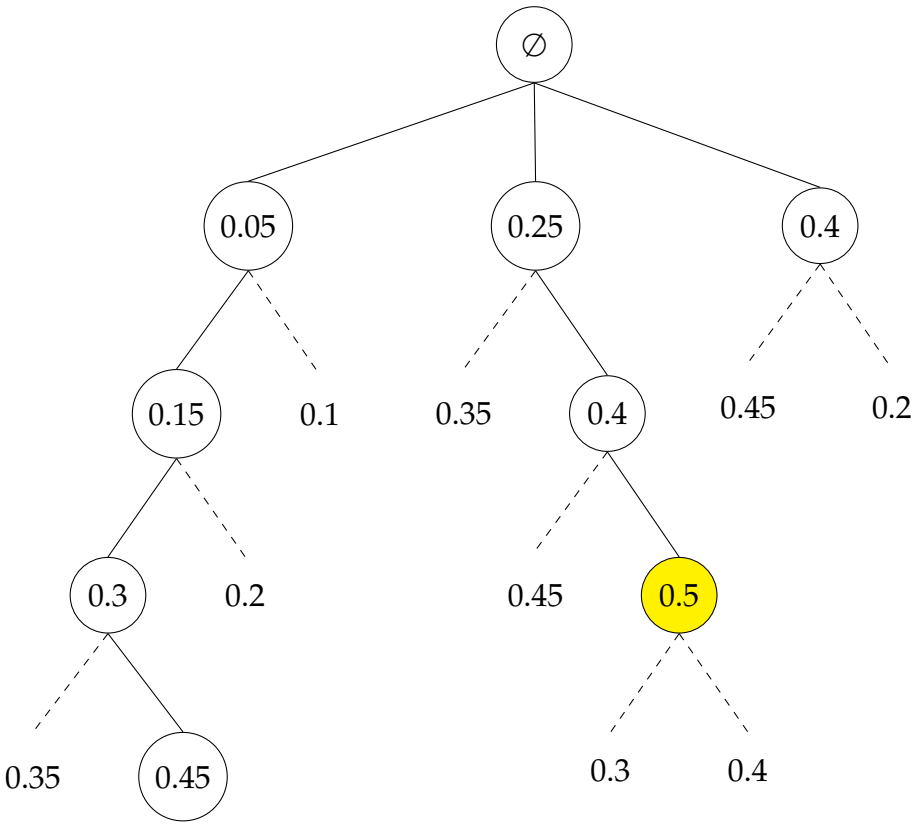

FIGURE 3.3: Tree-like explanation of how a Branch-and-Bound algorithm works. The depth of the tree represents the parameter ℓ, which is the maximum length of rules to be considered in the search space. Each *child* node is the extension of the *parent* rule by one feature, taken out of $\mathbf{X}$; the numerical value of a node is its score according to the objective function $f(x)$. In this example, $f(x)$ is assumed subject to maximization to find the optimal solution. This approach reduces the search space of possible solutions by discarding the paths that are less likely to produce an improvement to the best result yet found. The dashed edges suggest the branches likely to lead to suboptimal solutions, thus pruned, while the yellow node represents the optimal solution produced as output of the algorithm. The algorithm tries to extend the length ℓ of the rule, while following only promising branches and improving the solution score.

At its core, *ALLSTAR* (see Figure 3.2) employs a Branch-and-Bound approach to discover the rule with the highest causal effect, while limiting to rules with at most ℓ alterations. Following the demonstration of NP-hardness, the Branch-and-Bound was chosen as method to solve the optimization problem to avoid the need to evaluate all the possible combination of features (i.e., rules): its goal is to find the best candidate value that maximizes (or minimizes, depending on the problem) a real-valued function $f(x)$, called objective function, among a set of admissible solutions by recursively splitting the search space into smaller spaces (*branching*) and keeping track

of bounds on the maximum (or minimum) that it is trying to find (*bounding*). These bounds are then exploited to *prune* the search space, eliminating candidate solutions that it can prove will not contain an optimal solution. A simplified graphical representation of this principle can be found in Figure 3.3.

Moreover, since in practice the general interest is in finding multiple and diverse rules with positive reliable effect and with functionally related alterations, *ALLSTAR* uses an iterative approach to identify at most k rules, where k is a parameter provided by the user, and an interaction graph G to consider only rules with functionally related alterations.

Specifically, *ALLSTAR* takes several inputs:

- a set $\mathbf{X}$ of alterations,
- a set $\mathbf{Z}$ of confounders,
- a value y of interest for the target variable Y,
- the maximum length ℓ of rules,
- a confidence level α,
- a graph G whose vertices are the alterations in $\mathbf{X}$ and whose edges represent some relation between alterations (e.g., an edge represents the interaction between the proteins where the alterations are found),
- the maximum number k of rules to be reported in output,

Algorithm 1: *ALLSTAR*

Input: alterations $\mathbf{X}$, confounders $\mathbf{Z}$, value y of target Y, max. rule length ℓ, confidence α, graph $G = (\mathbf{X}, E)$, integer k, clean-up threshold t

Output: top-k reliable causal rules

```
1  N ← calculateRulesNumber(G, ℓ); α_c ← α / N;
2  output ← ∅; Q ← empty FIFO queue;
3  for i ← 1 to k do
4      ê_max ← −∞; σ_max ← ∅;
5      for X_j ∈ X do Q.enqueue("X_j = 1");
6      while |Q| > 0 do
7          σ ← Q.dequeue();
8          if upperBoundRelATE(σ,y,Z,α_c) > ê_max then
9              ê_σ ← computeRelATE(σ, y, Z, α_c);
10             if ê_σ > ê_max then ê_max ← ê_σ; σ_max ← σ;
11             for σ' ∈ expand(σ, G, ℓ) do Q.enqueue(σ');
12     if ê_max > 0 then output ← output ∪ {σ_max};
13     update(X, σ_max, t);
14 return output;
```

- a clean-up threshold $t \in [0,1]$ that controls the diversity of the rules reported in output.

In output, *ALLSTAR* produces at most k rules containing up to ℓ alterations, with the highest reliable effect and where each rule consists of alterations that form a connected subgraph of G. In addition, each reported rule comprises alterations that appear in a set of patients different from the alterations in other reported rules, where the difference is controlled by the parameter t.

3.2.1 Family-Wise Error Rate and Confidence Level α

When two variables are tested for independence, they are usually considered dependent if the p-value of the corresponding test is below a certain threshold α. It is easy to see that such procedure guarantees that if the variables are indeed independent, then the probability of a false discovery, that is, mistakenly rejecting their independence, is at most α. Instead, if a number N of tests is performed, and the same threshold α is used for each of the N tests, the expected number of false discoveries is increased up to αN. In Chapter 2, I pointed out the contribution of Budhathoki, Boley, and Vreeken (2021), who introduced the *reliable* estimator $\hat{e}^y_{rel}(\sigma)$, a biased estimator for $e^y(\sigma)$ with lower variance. It accounts for statistical noise using a confidence interval that contains the true rule effect with confidence $1-\alpha$, where $\alpha \in (0,1)$ is user-defined. However, such an estimator is correct for the effect estimation of just one rule, but it may lead to false positives if multiple hypotheses (i.e. multiple rules) are analyzed. In this thesis' scenario, in which the aim is discovering the top-k rules with the largest effect among a high number of combinations, this approach is no longer applicable.

Rigorously, the Family-Wise Error Rate (FWER) is the probability of returning in output at least one false positive (FP) when conducting multiple hypothesis testing, and it can be defined as

$$FWER = P(FP \geq 1) = 1 - P(FP = 0). \tag{3.1}$$

Therefore, by keeping $FWER \leq \alpha$ it is possible to control the overall probability of returning false positives at tests' family-level. There are various ways to apply this, the one I chose for *ALLSTAR* is *Bonferroni correction* (Bonferroni, 1936). Substantially, a corrected threshold $\alpha_c = \alpha/N$ is considered for each hypothesis where N is the number of (potential) hypotheses tested. A simple union bound shows that the resulting FWER is at most α.
Among all methods, Bonferroni correction is considered as one of the most conservative ways to limit the FWER: in other words, the risk is to considerably reduce the estimated causal effect, when a high number of rules is evaluated. In a field like cancer biology, I felt the need to model *ALLSTAR* in such a way for a very practical reason: by returning as output rules with strong causal signal, the immediate advantage is to reduce the number of downstream investigations to the very essential. In fact, as it will be discussed in Chapter 6, the results need to be evaluated by domain experts, in order to be validated. *ALLSTAR* is able to point the finger towards potential

interesting targets, but biological processes addressed by the rules in output, should always be investigated further, possibly with *in vitro* experiments.

3.2.2 A Priori Knowledge: the Protein-Protein Interaction Network & Graph G

One of the most effective ways to bridge the gap between computer science and its implementation to biological scenarios is to *inform* the algorithm about aprioristic knowledge, relevant to the application domain. Protein-Protein interaction (PPI) networks play a crucial role in enhancing our understanding of biological processes and how various proteins/genes relate to each other. By unveiling functional links, these networks reveal the intricate molecular and cellular mechanisms that govern the health and disease states of organisms. As a result, this information can be exploited to guide the search for optimal results along functionally meaningful paths. In the case of *ALLSTAR*, it can be optionally fed with a suitable PPI and generates an interaction graph G that is used to extend rules with $\ell > 1$. Thus, the Branch-and-Bound core of *ALLSTAR* can evaluate only combinations of altered genes that represent a link in G. The immediate benefit of this approach consists in forcing the algorithm to produce domain-relevant results to drive the post-analysis discussion into an appropriate track. An additional silver lining of the PPI exploitation consists in the fact that the search space is dramatically reduced in the first place, leading to significant reduction of the waiting times to obtain the results.

3.2.3 The Clean-up Threshold t

As pointed out by Fisher, Pusztai, and Swanton (2013), the increasing evidence that solid tumors may consist of subpopulations of cells with distinct genomic alterations within the same tumor, a phenomenon known as *intra-tumor heterogeneity*, has been ascertained by more and more experimental results over the years. This genetic complexity of malignant tumors is likely to have profound implications for the characterization and understanding of cancer itself, as well as biomarker discovery, especially in the era of personalized treatment. Consequently, there is emerging evidence suggesting a relationship between intra-tumoral heterogeneity and clinical outcomes. To tackle this problem and to identify a diverse and more informative set of rules, the set $\mathbf{X}$ of alterations is updated after each rule is extracted. This is done with function `update`$(\mathbf{X}, \sigma_{\max}, t)$ (line 13), hereby described. Such function removes from the set of alterations $\mathbf{X}$ the ones that either appear in the rule $\sigma_{\max}$ or are very similar to at least one alteration in $\sigma_{\max}$. The similarity is defined according to the normalized city-block Manhattan distance, defined for two vectors $\mathbf{a}$ and $\mathbf{b}$ in n dimensions as

$$d_M(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i| . \tag{3.2}$$

In particular, `update`($\mathbf{X}$,$\sigma_{\max}$,t) removes from $\mathbf{X}$ all alterations in $\sigma_{\max}$ and the ones with distance d_M less than t from at least one alteration in $\sigma_{\max}$, where the distance between the vectors describing the appearance of alterations in patients is considered and t is a user-defined threshold.This function therefore allows to recover non-overlapping rules over the whole alterations' search space, focusing on a wide spectrum of potential genetic targets causally related with the outcome of interest.

3.2.4 *ALLSTAR*'s Methodological Theorem

The following theorem proves that *ALLSTAR* produces in output a set of rules with a rigorous bound on its FWER, where a false positive is defined as a rule σ reported in output but with effect $e(\sigma) \leq 0$. The proof is provided below.

Theorem 1. *ALLSTAR($\mathbf{X}$, $\mathbf{Z}$, y, ℓ, α, $G = (\mathbf{X}, E)$, k, t) outputs a set of rules with $FWER \leq \alpha$.*

Proof. It is noticeable that each iteration of the for loop at line 6 considers an increasingly small subset of $\mathbf{X}$ and therefore the total amount N of candidate causal rules that may be evaluated by *ALLSTAR* (i.e. the total number of hypotheses tested in the worst scenario) is equal to the total number of rules that can be evaluated on the first iteration of the loop. In particular, the number of all the different rules of max length ℓ (i.e. N, line 1) is equivalent to the number of distinct connected subgraphs in G of length at most ℓ since *ALLSTAR* exploits G to expand a rule σ to a more specific $\sigma' \supset \sigma$ by adding a proposition $X_i = 1$ only if X_i is not already present in σ and it is connected to at least one treatment of σ.

It is proved in the next paragraph that, by setting $\alpha_c = \alpha/N$ (line 1), *ALLSTAR* returns a false positive with probability at most α. It is supposed that a false positive rule σ_{FP} (i.e. such that $e(\sigma_{FP}) \leq 0$) is returned in output by *ALLSTAR*. A necessary condition for this to happen is to add σ_{FP} to the top-k rules found (line 12) which in turn happens only if its estimated effect $\hat{e}_\sigma$ (calculated in line 9) is greater than 0 (line 12). By construction of the confidence intervals with confidence α_c, a rule with $e(\sigma_{FP}) \leq 0$ may have its estimated effect $\hat{e}_\sigma > 0$ with probability at most α_c. Since there are at most N rules under study, in the worst case the probability of having at least a false positive estimate is $N \times \alpha_c = \alpha$ which implies that the algorithm does not output any false positive with probability of at least $1 - \alpha$. □ □

3.3 A Step Further: an Improved Bound

While the parallel implementation of *ALLSTAR* employs the branch-and-bound approach proposed in Budhathoki, Boley, and Vreeken (2021), an additional contribution to this project is constituted by the development of an improved (i.e., tighter) upper bound on the reliable causal effect of a rule.

Such bound is best suited for single-core runs since it requires a data structure shared among processors. It relies on the key observation that one rule σ' is more specific of every rule in $\Omega_p = \{\sigma' \setminus \{\wedge\pi_k\} | \forall \pi_k \in \sigma'\}$.

Consider a rule $\sigma = \pi_1 \wedge ... \wedge \pi_i$ and a more specific one $\sigma' = \sigma \wedge \pi_j$. Budhathoki, Boley, and Vreeken (2021) defined the upper bound $\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z})$ to the reliable effect estimate $\hat{e}_{rel}(\sigma')$ of σ' as a function of the number of instances n (in the $\mathbf{Z}$ strata), the number n_1 of instances with $Y = y$, and the number a_σ of instances for which σ holds and $Y = y$, as

$$\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_\sigma \in \{0,1,...,a_\sigma\}} \frac{a'_\sigma + 1}{a'_\sigma + 2} - \frac{n_1 - a'_\sigma + 1}{n - a'_\sigma + 2} - \frac{\beta(\alpha)}{2\sqrt{a'_\sigma + 2}} - \frac{\beta(\alpha)}{2\sqrt{n - a'_\sigma + 2}} \tag{3.3}$$

which upper bounds the effect of σ' by exploiting the fact that a_σ will upper bound the number $a_{\sigma'}$ of instances for which the σ' holds and $Y = y$, given that σ' is more specific than σ. It is arguable that σ' not only is more specific than σ, but also than every rule in the set $\Omega_p = \{\sigma' \setminus \{\wedge\pi_k\} | \forall \pi_k \in \sigma'\}$ of all possible rules chosen from σ' removing the proposition π_k. The proposed estimator must hold for each rule in Ω_p therefore I propose a tighter optimistic estimator that considers $a_{min} = \min_{\sigma_j \in \Omega_p} a_{\sigma_j}$ as

$$\bar{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_{\Omega_p} \in \{0,1,...,a_{min}\}} \frac{a'_{\Omega_p} + 1}{a'_{\Omega_p} + 2} - \frac{n_1 - a'_{\Omega_p} + 1}{n - a'_{\Omega_p} + 2} - \frac{\beta(\alpha)}{2\sqrt{a'_{\Omega_p} + 2}} - \frac{\beta(\alpha)}{2\sqrt{n - a'_{\Omega_p} + 2}} \tag{3.4}$$

Notice that if a rule $\sigma_{rem} \in \Omega_p$ has been pruned by the breadth-first branch and bound algorithm, then $\bar{\tau}_{\sigma'}(\sigma, \mathbf{z}) = -\infty$ can be set since the condition in line 8 does not hold for any such σ', given that it is more specific than σ_{rem}.

Chapter 4

Literature Review

After introducing the methodological perks of *ALLSTAR*, this chapter depicts the state of the art to contextualise where the algorithm stands in the current field of computational oncology.

Randomized Controlled Trials & Observational Studies

The optimal statistical setting to infer causal relationships between variables is offered by randomized controlled trials (RCTs), as seen in Concato, Shah, and Horwitz (2000) and Rosenbaum, Rosenbaum, and Briskman (2010). To this extent, RCTs are considered the *gold standard* for causality in the biomedical field. Their foundation consists in subdividing the population under test into a *treatment* group, whose members are all given the cause (e.g., a medicine) under test, and a *control* group, not receiving the cause and producing natural effects. If the basic principles of RCTs are respected, such as random assignment of the cause, and a suitable choice of the placebo for the controls, the effect of the cause is expected to be precisely estimated, without falling into correlation-bound spurious relationships (Cartwright, 2010).

In reality, RCTs are most of the time extremely expensive, ethically challenging, or, worse, unfeasible. For example, in the setting of this thesis as it will be clarified in the methodological chapter, building an RCT with somatic alterations as causes to investigate, poses the challenge to plant mutations into patients of the treatment group. On the other hand, the increasing amount of observational data being collected provides the opportunity to mine such data to identify possible interesting relations, to be confirmed with follow-up experimental evaluation.

The Landmark for Causal Rules Mining

Mining association *rules*, or combinations, aims at discovering frequent patterns from a dataset (Agrawal, Imieliński, and Swami, 1993; Kotsiantis and Kanellopoulos, 2006). However, in recent years, a lot of attention has been devoted towards mining *causal* rules (Silverstein et al., 2000) from observational data. Recently, Budhathoki, Boley, and Vreeken (2021) proposed a novel estimator of a rule's effect, taking into account the uncertainty of the estimates derived from data, and developed a branch and bound algorithm for the discovery task. This thesis proposes a tool advancing their framework, which is tailored for scenarios with low number of variables, by implementing a

correction for controlling the FWER in a multiple hypothesis testing setting. This is a fundamental feature of cancer studies given the high number and typologies of alterations found in tumors.
Moreover, *ALLSTAR* comes with improved performances due to a novel and tighter upper bound to the reliable effect of a rule, and due to the incorporation of prior knowledge in the form of an interaction graph. This feature allows to reduce the search space while focusing on functionally related alterations.

State-of-the-Art Causal Tools & Multi-Omics

In the past ten years, given the theoretical advantages over correlational approaches, causality-based tools have started to rise, also in the field of computational biology. As stated in Chapter 1, the constant upgrade of sequencing technologies and consequential increase of accessible *omics* data pushed the development of tools and pipelines able to solve more and more complex biological problems. Various causal tools are now publicly available, but *ALLSTAR* differentiates from existing approaches due to the possibility to mine causal combinations of alterations with respect to a target phenotype and to output their impact in terms of effect. For instance, *ALLSTAR* focuses on estimating the impact of genomic alterations on a tumor subtype, unlike bayesian approaches such as Zhang, Burdette, and Wang, 2014 that learn a causal graph from The Cancer Genome Atlas (TCGA) mutation data to identify genetic causes relevant to ovarian cancer, but without considering their effect on a target variable. The final outcome is an interesting directed graph, comprising and refining multiple relations between genes, but it fails in specifying a genetic target for follow-up analyses.

Other causal tools, instead, such as Cifuentes-Bernal et al. (2022), leverage the increasing availability of single-cell RNAseq data and the estimates of pseudo-time derived from such data to identify causal relations at the transcriptomic level. The pseudo-time reconstruction is an in silico procedure that places each cell on a time axis, producing a temporal-like ordering. Since temporal relationships are usually considered valid when identifying causes and effects, these tools are thus included within the causality realm, even if they draw conclusions from similarities at gene-level between pseudo-temporally-ordered cells' transcriptional counts.

A step towards the identification of causal relations between multi-omics data and a target variable (e.g., phenotype) has been made by the tools Aristotle (Mansouri et al., 2022) and CauMu (Liu et al., 2022), both identifying single features (i.e., alterations, or genes) linked to the phenotype. *ALLSTAR* takes another step by providing an efficient approach to identify rules comprising multiple features, which is an important characteristic given the high levels of heterogeneity found in cancer. Moreover, Aristotle focuses on the significance of the relation by computing a corresponding p-value, rather than returning their effect as done by *ALLSTAR*.

Chapter 5

Synthetic Experiments and Algorithm Evaluation

This Chapter is meant to assess *ALLSTAR*'s performance. Before applying the algorithm on real data, it is good practice to build meaningful experiments on synthetic data to motivate its perks and prove its claims on the robustness of the output. The synthetic experiments can be summarized as follows:

1. comparison of *ALLSTAR* with standard correlational approaches[1];
2. evaluation of the impact of the multiple hypothesis testing correction employed by *ALLSTAR*;
3. effectiveness of using graph G to reduce the number of rules to evaluate;
4. assessment of *ALLSTAR*'s ability to recover diverse rules, involving multiple alterations, planted in a large, noisy dataset, and robustness with regards to Assumption 3 violations (see Sec. 3.1) under suitable settings of update threshold t;
5. confirmation of the stability of our algorithm on different combinations of the user-defined parameters ℓ, G, and t, to highlight how increasing the number of combinations decreases the recovered effect without leading to any false discovery;
6. establishment of computational performances against a brute-force approach.

Every synthetic dataset resembles the structure of real cancer data, with mutated genes as treatments $\mathbf{X}$ and a binary outcome Y. For simplicity, $\mathbf{Z} = \emptyset$ was set in these analyses. For each experiment, 10 datasets for every tested sample size (25, 50, 75, 100, 250, 500, 1000, 5000, 10000, and 25000) were randomly sampled. In each dataset, most alterations are drawn randomly with probability 0.5 and independently of the outcome Y. In some datasets, alterations with a causal relation to the target Y were planted; such alterations constitute the rules of interest to assess *ALLSTAR*'s performance. Their relationships with Y are described in the section related to each experiment.

[1] It has been not possible to compare with Aristotle due to issues with its implementation, available at https://github.com/MehrdadMansouri/Aristotle.

5.1 Comparison with Correlational Approaches

In the preliminary experiment, I compared *ALLSTAR* with standard correlational approaches, to understand whether the results obtained are the same. The main purpose does not consist in motivating the choice of causal methods over correlational ones, rather assessing a different output between the approaches to support the novelty of *ALLSTAR*'s contributions. In particular, I considered a dataset of 100 samples with one alteration and a target; I omitted confounders since they would be a superflous addition for the sake of the experiment. All possible permutations of the alteration's distribution across samples were generated, hence each permuted dataset referred to a different rule comprising the alteration and the target (i.e., a uniquely supported rule, since each permuted dataset is different in the alteration's distribution). Three metrics were then computed: *i)* the reliable effect by *ALLSTAR*, *ii)* the p-value from the Fisher exact test, and *iii)* the odds ratio. More specifically, in the two latter cases, the values were obtained by taking into consideration the contingency table and by evaluating the distribution of σ (that for each sample it is either $\top$ or $\bot$) and the linked distribution of the event "$Y = y$", associated to each rule as in Budhathoki, Boley, and Vreeken, 2021. The results were then sorted according to each computed value and the three rankings compared. Figure 5.1 shows the rankings of the top-10 rules ranked by highest reliable effect: the top rules obtained by *ALLSTAR* have a much lower ranking as if they were ranked by p-value or odds ratio. For example, 4 of the top-10 rules according to the reliable effect are not in the top-10 by p-value or by odds ratio, with one rule appearing in the 18^{th} position of the ranking by p-value. In general, while there is an overall concordance in terms of Kendall-tau coefficient (Kendall, 1938) between the ranking by reliable effect and the other measures (Kendall-tau coefficient 0.79 correlation between the odds ratio and effects; Kendall-tau coefficient 0.9 between p-values and effects), the reliable effect provides different top rules (which are the most interesting ones for any practical purpose) than standard correlation approaches. In fact, in a real-world scenario, downstream analyses (such as *in vitro* confirmation experiments) are performed by oncologists and biologists on a limited part of the results, ideally the highest scoring rules. To simulate a situation like the one described, I took a step further and replicated the same experiment in a real-data example that includes confounders.

It appears that findings in the synthetic setting are exacerbated in the real-world scenario. A breast cancer-related dataset was chosen as reference, presenting 622 alteration profiles, 7 confounders and a binary outcome[2]. *ALLSTAR* and a python implementation of the Cochran-Mantel-Haenszel test (CMH) were both run on the dataset, ranking reliable effects and p-values for rules built on every combination of one confounder, two alterations and the outcome. The scatter plot describing the two rankings' comparison for

[2] A dataset was chosen among the ones that will be described in Chapter 6. Specifically, it was selected the dataset with the 300 most frequent somatic mutations, the 300 most frequent loss of heterozygosity profiles (LOHs), and the patterns of 22 frequently hypermethylated genes, as $\mathbf{X}$, and the Triple-Negative binary molecular classification, as target Y.

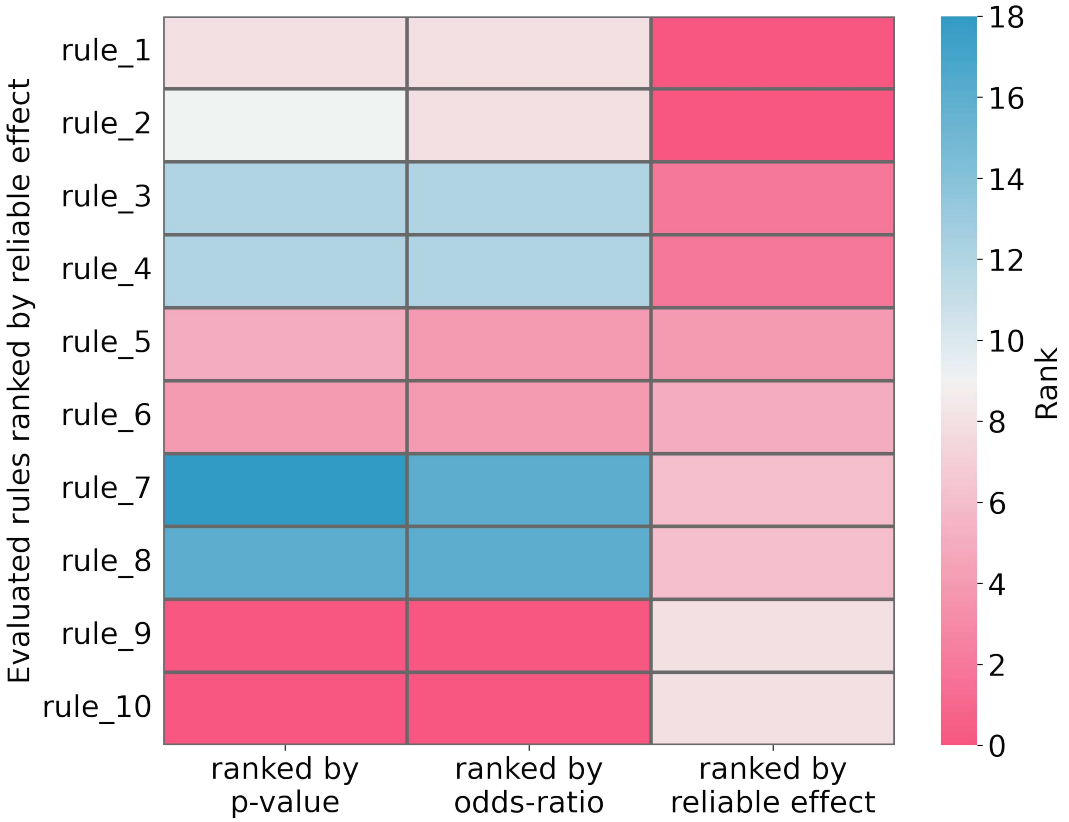

FIGURE 5.1: Ranking comparison of the top-10 rules with the highest effect computed on a dataset of 100 samples with regards to three different metrics (p-value, odds-ratio and reliable effect). Each row corresponds to a rule and each column corresponds to its ranking (with 1st scores being the highest) with regards to p-value, odds-ratio and *ALLSTAR* reliable effect, respectively. Color-scale representing ranking position on the right.

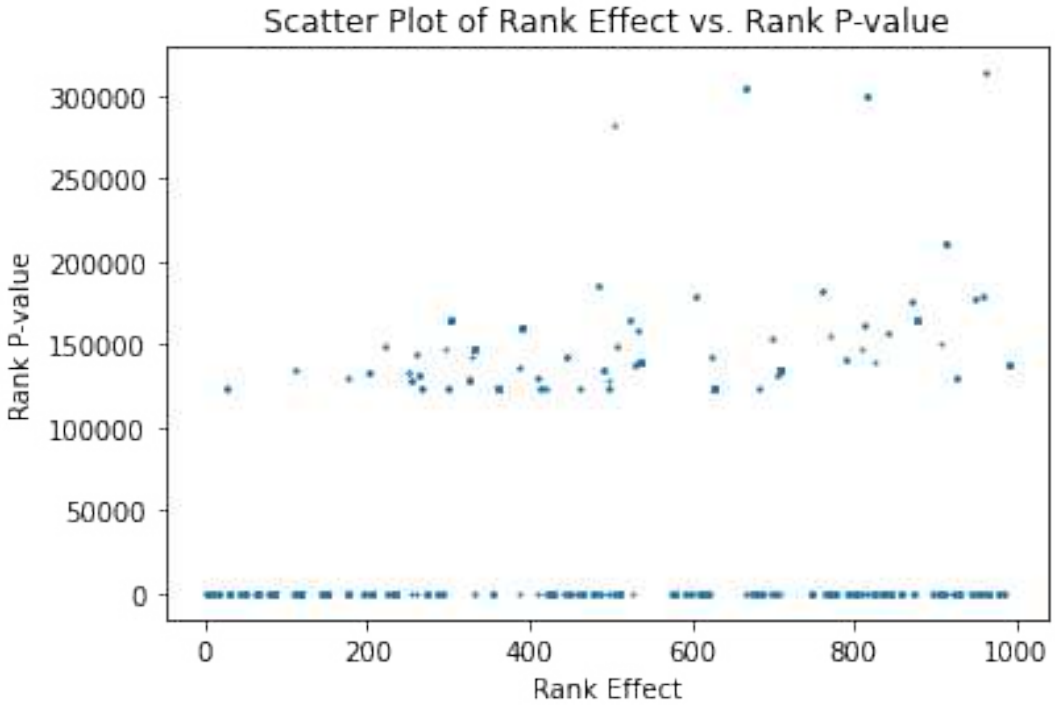

FIGURE 5.2: Comparison of the rankings in terms of reliable effect (*ALLSTAR* output, x-axis) and p-value (CMH test output, y-axis) for real-world data. Each dot corresponds to one of the top-1000 rules ranked by reliable effect.

the first 1000 rules sorted by reliable effect, is shown in Fig. 5.2. It seems clear that a considerable amount of rules ranked among the top-1000 in the effect ranking, when assessed using the correlation-based method CMH, are placed well beyond the 100000^{th} position. Additionally, the p-value rankings compressed to the bottom of the plot are actually all valued 1. CMH is not able to differentiate all these rules, giving them a p-value of zero, which hints at their possible significance, but fails at prioritizing the combinations of genes that may be relevant to the target outcome.

5.2 Impact of Multiple Hypothesis Testing (MHT) Correction

In this experiment, the impact of correcting for multiple hypothesis testing on false positives was investigated. Only random alterations were considered in each sample, hence no causal rule (i.e. any rule with a positive effect) with respect to the outcome was planted. In this setup, three different estimates of the (reliable) effect were taken into account: the version based on the naïve

estimate of probabilities, the reliable approach proposed in Budhathoki, Boley, and Vreeken (2021)[3], and the one used by *ALLSTAR*.For the last two estimates, which are considering a confidence level, the value $\alpha = 0.05$ was considered. The estimation of probabilities from data is challenging when sample sizes are small, as the estimates obtained with naïve empirical estimators have high variance. As a consequence, rules discovered by data using such naïve empirical estimators have effects whose estimates are far from their true effects. In particular, the naïve approach estimates the effect (i.e., empirical probabilities estimated from data and without any correction) as the difference

$$\hat{e}(\sigma) = \hat{p}(Y = y|\sigma = \top) - \hat{p}(Y = y|\sigma = \bot),$$

while the reliable approach proposed in Budhathoki, Boley, and Vreeken (2021) considers $\hat{e}_{rel}(\sigma)$ (i.e., adding confidence bounds) but without correcting for multiple hypothesis testing, as it is done instead in *ALLSTAR*.

Both the naïve and reliable approach incorrectly return at least one rule with a positive effect for *every* dataset (i.e., corresponding to a FWER of 1), while *ALLSTAR* is the only one correctly returning zero false positives. These results reinforce that the multiple hypothesis correction on *ALLSTAR*'s reliable effect is a crucial component to avoid false discoveries.

5.3 Perks of Using an Interaction Graph G

The second experiment assesses the effectiveness of using the interaction graph G in *ALLSTAR* when identifying causal rules composed of multiple alterations. A total of 22 alterations were used to sample multiple datasets. 7 of these alterations are part of a rule causally related to the target Y and constitute a connected subgraph of G. *ALLSTAR* was run with various values of the maximum rule length ℓ as inputs. As expected, the estimate of the effect converges to the true effect for all values of ℓ, and the estimate obtained using the interaction graph G is significantly better than the one when no prior knowledge is considered. Moreover, as additional benefit, the use of G drastically reduces the runtime, due to a reduction in the number of candidate rules.

Figure 5.3 shows the results obtained passing G in input (dotted line) and the results obtained when no prior knowledge on gene interaction is considered (dash-dotted line), obtained by passing a fully connected graph in input to *ALLSTAR*. In particular, both the effect estimation of the implanted rule and the runtime were considered. For example, with 25000 samples and $\ell = 7$, the runtime using G is of few seconds, while almost 3 minutes are required when no prior knowledge is considered. This shows that the interaction graph leads to significant improvements in terms of the estimate of the true effect and of runtime.

[3]The code is available on the bitbucket repository at https://bitbucket.org/realKD/ does not run properly, therefore I implemented an equivalent, but functioning version.

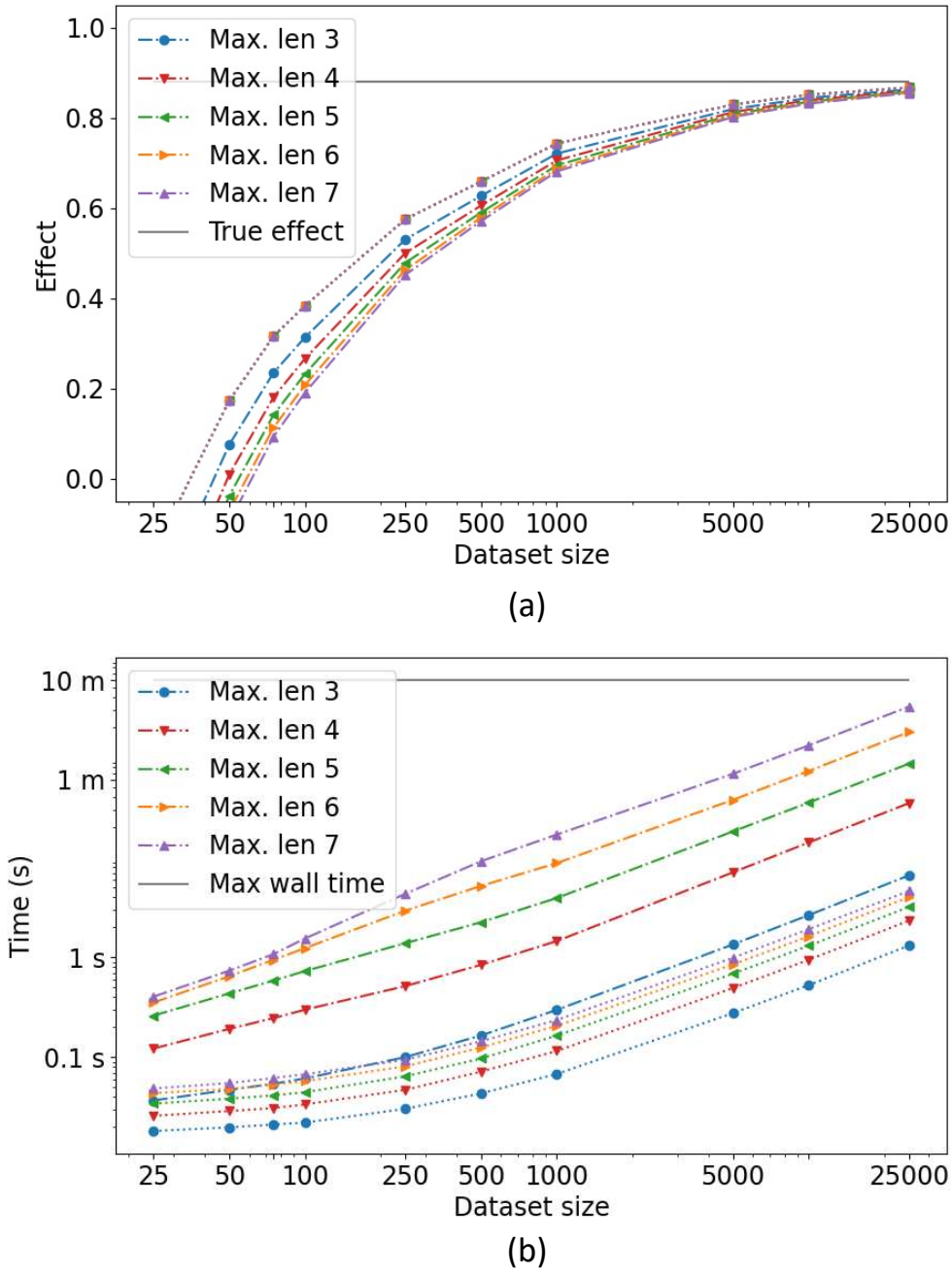

FIGURE 5.3: Mean planted reliable rule effect (a) and mean runtimes (b) over multiple dataset sizes on 10 runs. In each plot, the dotted lines represent *ALLSTAR* results passing a protein-protein interaction G in input, and dash-dotted lines represent the approach with a fully connected graph (i.e., no prior knowledge).

5.4 Ability to Tackle Inter-Tumor Heterogeneity

A key feature claimed by *ALLSTAR* is the power to tackle cancerous scenarios characterized by high inter-tumor heterogeneity. To prove this feature, an adequate experiment was set up to assess the ability of recovering planted rules that cover a wide spectrum of diverse functional processes. As a plus, this experiment suggests the robustness of the algorithm even when some admissible input structure assumptions (see Sec. 3.1) are not satisfied. Datasets were simulated with 3 planted rules (of 5 genes in total) and 100 random alterations. For each alteration in a planted rule, a correlated alteration was generated with 97.5% of values identical to the planted alteration. This allows to assess whether *ALLSTAR* correctly reports only the causal alterations or returns an output contaminated by the correlated (but not causal) ones.

Both a variant of *ALLSTAR* (that ignores the Manhattan distance-based updating procedure, see function `update`$(\mathbf{X}, \sigma_{\max}, t)$ in Algorithm 1), and *ALLSTAR* itself with $t = 0.05$ were run. The top-3 rules in output of the two analyses were then compared. *ALLSTAR* reports the planted rules and correctly disregards the rules comprising the correlated alterations. The variant of *ALLSTAR* not using the Manhattan distance-based updating procedure, instead, produces, among the top-3 rules, rules containing the correlated alterations. The immediate positive side of this result is the ability of the algorithm to produce diversity within the output: in fact, most genetic alterations that are related to the same biological mechanism tend to either have the same mutational pattern or be characterized by mutual exclusivity. In any of these two cases, the pairwise correlation is extremely high, *ALLSTAR* is able to manage this peculiar situation.

Additionally, another way of seeing the failure to apply the update threshold t is that *ALLSTAR* returns correlations instead of causal relations: this is due to the fact that Assumption 3 for the admissible causal structure is not satisfied. On the other hand, the Manhattan distance-based updating procedure allows to remove the spuriously-linked variables and to report only causal relations. These results show that the use of the Manhattan distance-based updating procedure is also important to focus on causal alterations only.

In particular, Figure 5.4 shows the graph of the data generative Bayesian Network (BN) associated with this experimental setup in *(a)* against the one assumed by *ALLSTAR* in *(b)*. The main difference between the two BNs is to be found in the relationship between $X_i, 1 \leq i \leq 5$ and their clones $X_{i(clone)}$. In particular, X_i always blocks spurious correlation paths (more on this and *d-separation* in Pearl, 2009) from $X_{i(clone)}$ in 5.4*(a)* but not in 5.4*(b)*, therefore if we (incorrectly) assume the underlying graph to be as the latter, in order to still have correct results we should be able to have some other heuristic mechanism (i.e., the threshold-based cleaning procedure) that removes the clones in order to exclude them from the analysis. Another difference between the two BNs in Figure 5.4 lies on the links between the external variables $(E_i, 1 \leq i \leq 100)$ and Y. Such links imply some form of (possible) dependence whose strength is defined by the probability distribution functions

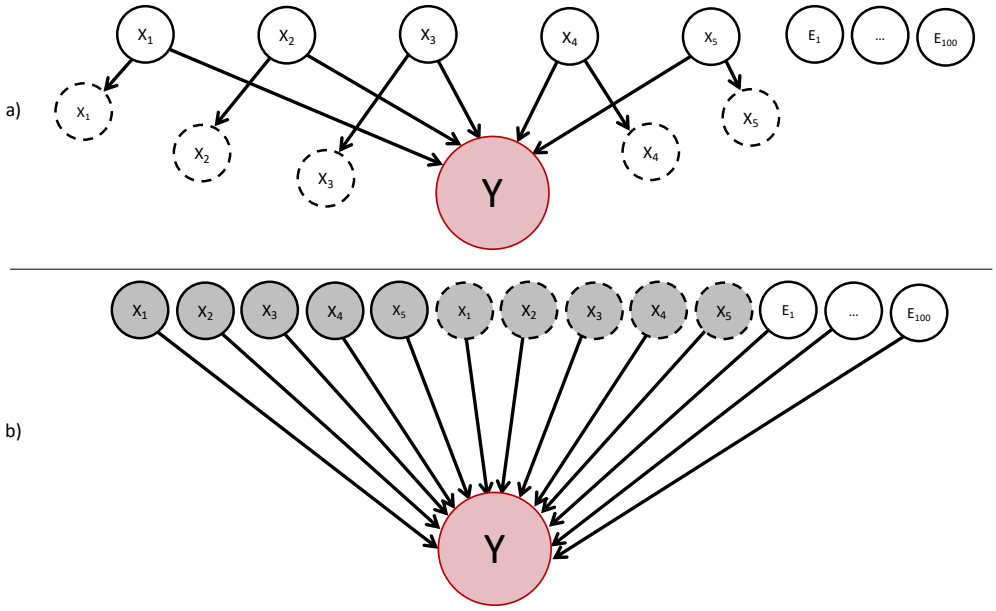

FIGURE 5.4: Data generative Bayesian Network *(a)* and assumed graph *(b)* of synthetic experiment in Section 5.4. In the second plot, clone variables (i.e. those which definition depends on other variables) are shown with a dashed border, and variables that output rules with a positive effect by *ALLSTAR* without cleaning procedure are represented in grey.

inferred by the observational dataset. *ALLSTAR* however, is able to confidently ignore such spurious correlations due to the use of the *reliable* effect estimator and its ability to deal with multiple hypotheses testing (see more on Theorem 1 proof).

For reproducibility's sake, the equations for sampling data from the graph of Figure 5.4(a) are the following:

$$
\begin{aligned}
X_1 &\sim \mathcal{B}(0.5) \\
X_2 &\sim \mathcal{B}(0.4) \\
X_3 &\sim \mathcal{B}(0.7) \\
X_4 &\sim \mathcal{B}(0.65) \\
X_5 &\sim \mathcal{B}(0.15) \\
X_{i(clone)} &\sim X_i \oplus \mathcal{B}(0.025) \\
E_i &\sim \mathcal{B}(0.5) \\
Y &\sim (X_1 \wedge X_2) \vee (X_3 \wedge X_4) \vee X_5
\end{aligned}
$$

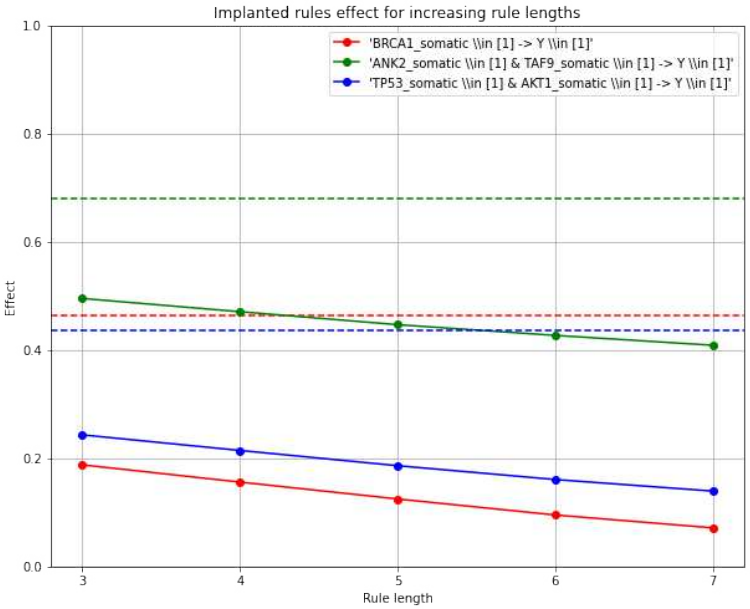

FIGURE 5.5: Average effect returned by *ALLSTAR* (solid lines) and theoretical value (dashed lines) for the 3 implanted rules of the last syntetic experiment (see Sec. 5.4). Results have been averaged on datasets with 1000 samples, and the variability in each run results is negligible.

5.5 Stability Analysis

The stability of *ALLSTAR*'s results were experimentally assessed with respect to the user-defined parameters ℓ, t, and $G = (\mathbf{X}, E)$. The two tests are reported below.

In the first of stability analyses, we run *ALLSTAR* on the datasets of the synthetic experiment of Sec. 5.4, with ℓ taking values from 3 to 7, and $t = 0.05$. Figure 5.5 shows the average effect returned by *ALLSTAR* (solid lines) for the three implanted rules by varying ℓ, as well as their theoretical value (dashed lines). Results have been averaged for all datasets of 1000 samples, and their variability across the runs is negligible. As expected, the rule effect returned by *ALLSTAR* decreases as the rule length increases because the number of hypotheses to test increases and therefore the Bonferroni correction becomes stricter. Moreover, despite increasing ℓ allows *ALLSTAR* to evaluate more rules, our algorithm did not return any false positive.

We then analyzed results variability with regards to changes of t by running *ALLSTAR* on the same setup of the previous experiment, and setting t to $0.01, 0.025, 0.05, 0.075$, and 0.1. We finally set $k = 4$ to assess the ability of *ALLSTAR* to avoid returning duplicated rules. *ALLSTAR* returned duplicate rules consistently among all the runs for $t = 0.01$, among 6 of 10 runs for $t = 0.025$, and did not return any duplicated rule for all the other values of t tested. This is an expected behavior in this data generative scenario since each rule differs from its clone on 2.5% of samples on average (see equations in Sec. 5.4).

5.6 Computational Performances

The last performed experiment aims at comparing the computational performances of *ALLSTAR* against a brute-force algorithm that exploits G to select the candidate rules to study, but calculates them all without exploiting the Branch-and-Bound. 10 synthetic datasets were created with 1000 samples from the following distributions:

$$\begin{aligned} X_1 &\sim \mathcal{B}(0.15) \\ E_i &\sim \mathcal{B}(0.1), 1 \leq i \leq 600 \\ Y &\sim X_1 \vee \mathcal{B}(0.05) \end{aligned}$$

and the rule with the highest effect ($k = 1$) was reported by setting the target value $Y = 1$. Both algorithms ran on 60 cores of a computing cluster and the runtimes were tracked without considering the time required to calculate the Bonferroni correction (i.e. function `calculateRulesNumber` of *ALLSTAR*) as the only focus is to compare the performances of the two rule discovery approaches[4]. Figure 5.6 compares the average runtimes in seconds of both approaches (y axis is log-scaled) over increasing maximum rule lengths ℓ. As expected, *ALLSTAR* is faster than the brute force approach due to the speedup given by its Branch-and-Bound, and their difference increases with the number of rules under study, therefore it increases monotonically with ℓ. As a reference, the brute force algorithm is more than 3 times slower than *ALLSTAR* when discovering rules setting $\ell = 4$, and nearly 20 times slower for $\ell = 5$.

[4] As a reminder, this procedure would be a prerequisite for both algorithms, therefore it would just add a bias term to both runtimes under analysis.

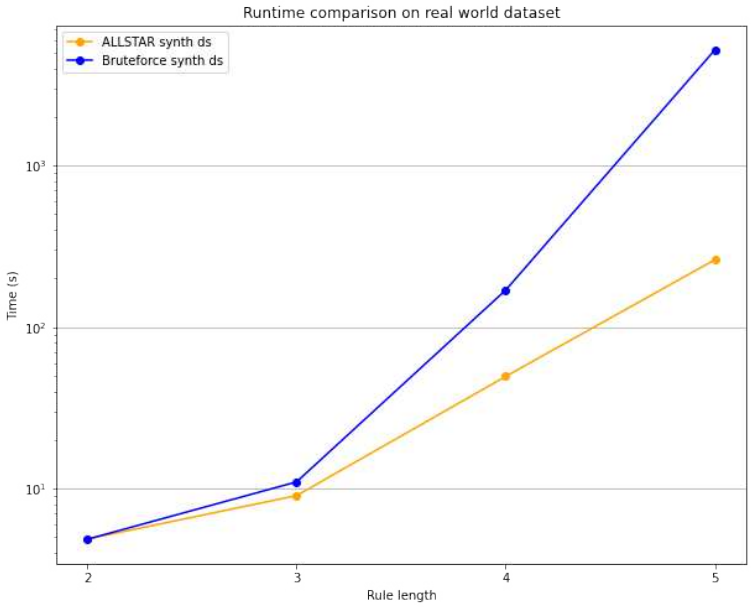

FIGURE 5.6: Average runtime comparison between *ALLSTAR* and a brute-force algorithm on 10 synthetic datasets over different rule lengths ℓ. Y-axis is logarithmically scaled, and variability across runs with the same ℓ is negligible (and therefore not plotted).

Chapter 6

Biological Experimental Results

6.1 Cancer Data and Interaction Network

ALLSTAR was tested on publicly available breast cancer (BRCA) data from TCGA. In particular, the main source is the TCGA-BRCA repository which offers public clinical and somatic mutational data, for a total of 1096 samples. The complete molecular and histological landscape of TCGA-BRCA patients was downloaded, comprising also the subtype classification of TCGA-BRCA based on the 50-gene PAM50 model (Parker et al., 2009). Germline mutational patterns for TCGA patients in BRCA1 and BRCA2 were retrieved from Kraya et al. (2019). Two additional alteration types that play a significant role in cancer were integrated: loss of heterozygosity (LOH) information from Riaz et al. (2017) and reported by Bodily et al. (2020), and hypermethylation from Xena Functional Genomics Explorer data (Goldman et al., 2015) and reported in Bodily et al. (2020). Not for every sample the corresponding target value was available, therefore the total number of observations dropped slightly. The final datasets comprised a number of samples ranging from 898 to 935, depending on the target variable of interest. As an input graph for *ALLSTAR* any protein-protein interaction network can be used: these results were generated exploiting the most recent Functional Interaction (Wu, Feng, and Stein, 2010) gene network from Reactome[1], which comprises almost 14,000 genes and more than 250,000 edges.

6.2 Real-World Cancer Datasets

ALLSTAR was run on breast cancer data described in Section 6.1. This section provides the details on how the datasets were built, the parameters used in the analyses with *ALLSTAR*, the results that were obtained, and their biological relevance.

The first step consisted in identifying from data described in Section 6.1 the variables to included into treatments $\mathbf{X}$, confounders $\mathbf{Z}$, and outcomes Y. In each run, a common set of confounders was considered, while treatments and outcome are combined in different ways to focus on certain cancer mechanisms. In particular:

[1]The repository *FIsInGene* (version 2021) is publicly available at `https://reactome.org/download/tools/ReatomeFIs/FIsInGene_122921_with_annotations.txt.zip`.

- 7 confounder variables **Z** were included, which are: *gender*, *race*, *age at diagnosis*, *menopause status*, *history of another previous malignancy*, and the *presence of a germline mutation in genes BRCA1 and BRCA2*.
- a total of 622 alterations were selected, including the 300 most frequently somatically mutated genes, the 300 most frequent LOHs, and 22 frequently hypermethylated genes. Two types of analyses were processed: one where each alteration corresponds to a treatment (element of **X**), and one where treatments **X** were the 300 most frequently altered genes by considering a gene mutated if any of the 3 alterations above is present.
- As target Y, three sub-typing classifications were accounted for: an histological categorization (Ductal, Lobular, and Other carcinoma), an expanded molecular one, based on gene expression (Basal, HER2E, Luminal-A, Luminal-B, and Normal-like), and a specific binary molecular classification (Triple-Negative, or not).

6.3 Results

ALLSTAR was tested under multiple settings on the differently combined datasets: maximum rule length ℓ spanned from 2 to 4, while $k = 3$, and $t = 0.01$. Informally, the maximum number of variables comprising each tested rule could be 4, the top-3 rules for each experiment were retrieved and the clean-up (similarity) threshold after each iteration of the algorithm was set at 99%. Data requirements increase exponentially with the size of **Z**, and, therefore, for each dataset *ALLSTAR* was run multiple times, passing a different subset of **Z** of cardinality at most 1. On each run, $\alpha = 0.05/(|\mathbf{Z}| + 1)$ was set to bound the FWER of all the tests on the same dataset below 0.05. Finally, both the presence ($GENE_{alteration} = 1$) and the absence ($GENE_{alteration} = 0$) of treatment were taken into consideration. Table 6.1 contains the highest-scoring results, separated into two blocks: on top, it shows the best rules with no confounders, while below it presents the best rules when conditioning on confounders. It is noteworthy that patients in every dataset are all affected by cancer, therefore every reported rule implicitly conditions on such event.

6.3.1 The Role of CDH1

The first three rules by effect include gene CDH1, which is a recurrently mutated gene in breast cancer and whose impact has been recognized as substantial (Pereira et al., 2016) in lobular histological subtype (McCart Reed et al., 2021; Erber and Hartmann, 2020), consistent with rules *a* and *b*; steadily, rule *c* states that the absence of an alteration in CDH1, given a breast cancer diagnosis, increases the chances of developing a ductal subtype, antagonist to the lobular one. Moreover, the combination of mutated CDH1 with unaltered ANK2 and SCN5A (rule *b*) provides an additional perspective on the

ID	Rule	Effect
a	$CDH1_{som} = 1 \rightarrow$ Lobular	0.470
b	$CDH1_{som} = 1 \wedge ANK2_{som} = 0 \wedge SCN5A_{som} = 0 \rightarrow$ Lobular	0.430
c	$CDH1_{som} = 0 \rightarrow$ Ductal Carcinoma	0.401
d	$ITGB3_{alt} = 1 \wedge RHOA_{alt} = 1 \wedge MAP3K1_{alt} = 1 \rightarrow$ Basal	0.342
e	$TP53_{som} = 1 \wedge ATRIP_{loh} = 1 \wedge ERBB2_{loh} = 1 \rightarrow$ Basal	0.300
f	$ITGB3_{alt} = 1 \wedge MAP3K1_{alt} = 1 \rightarrow$ Basal	0.297
g	$TP53_{som} = 0 \rightarrow$ Luminal-A	0.289
h	$RB1_{loh} = 1 \wedge PHB_{loh} = 1 \wedge LIMD1_{loh} = 1 \rightarrow$ Basal	0.271
i	$TP53_{som} = 0 \wedge BRCA1_{meth} = 0 \rightarrow$ Luminal-A	0.268
j	$ERBB2_{alt} = 1 \wedge MST1_{alt} = 1 \rightarrow$ Basal	0.245
k	$MST1_{loh} = 1 \wedge ERBB2_{loh} = 1 \rightarrow$ Basal	0.243
l	$RB1_{loh} = 1 \wedge PHB_{loh} = 1 \rightarrow$ Basal	0.242
m	$STAT3_{alt} = 1 \wedge ERBB2_{alt} = 1 \wedge WNT5A_{alt} = 1 \rightarrow$ Basal	0.241
n	$TP53_{alt} = 1 \wedge RB1_{alt} = 1 \wedge NGFR_{alt} = 1 \rightarrow$ Basal	0.241
o	$PIK3CA_{som} = 0 \wedge RHOA_{loh} = 1 \wedge NGFR_{loh} = 1 \rightarrow$ Basal	0.240
p	$TP53_{som} = 1 \wedge NME1_{loh} = 1 \rightarrow$ Basal	0.230
q	$TP53_{loh} = 1 \wedge PRKCD_{loh} = 1 \wedge NME1_{loh} = 1 \rightarrow$ Basal	0.226
r	$PDX1_{alt} = 1 \wedge SPOP_{alt} = 1 \rightarrow$ Basal	0.203
s	$TP53_{som} = 1 \wedge ERBB2_{loh} = 1 \wedge PRKCD_{loh} = 1 \rightarrow$ TripleN	0.195
t	$ITGB3_{alt} = 1 \wedge RHOA_{alt} = 1 \wedge MAP3K1_{alt} = 1 \rightarrow$ TripleN	0.184
u	$TP53_{som} = 0 \wedge BRCA2_{som} = 0 \rightarrow$ Luminal-A | gender	0.243
v	$CDH1_{som} = 1 \wedge AKT1_{som} = 0 \rightarrow$ Lobular | age_at_diagnosis	0.229
w	$ERBB2_{alt} = 1 \wedge RHOA_{alt} = 1 \rightarrow$ Basal | $BRCA2_{germ}$	0.202
x	$TP53_{som} = 0 \wedge BRCA2_{som} = 0 \wedge BRCA1_{meth} = 0 \rightarrow$ Luminal-A | gender	0.197
y	$TP53_{som} = 0 \wedge RB1_{som} = 0 \wedge BRCA1_{meth} = 0 \rightarrow$ Luminal-A | $BRCA2_{germ}$	0.191
z	$TP53_{som} = 1 \wedge ERBB2_{loh} = 1 \rightarrow$ Basal | history_other_malignancy	0.175

TABLE 6.1: Best rules, with (bottom) and without (top) confounder's conditioning, ordered by descending effect. Rules' description is as follows: $GENE1_{alteration} = [0,1] \wedge GENE2_{alteration} = [0,1] \wedge \ldots \rightarrow$ Target subtype | Confounder.

mechanisms regulating the lobular subtype: ANK2 is typically downregulated in breast cancer, while SCN5A is upregulated in almost every neoplastic process. However, SCN5A is known to mediate the epithelialmesenchymal transition (EMT), a biological trait underpinning cancer aggressiveness: the absence of a mutation in this gene can be interpreted as a normal state for EMT, aligned with the mild characteristics of the lobular subtype (Gradek et al., 2019; Luo et al., 2020). Additionally, the rule including the absence of mutation in AKT1 in lobular carcinoma (rule v) is coherent, since this gene is strongly associated with ductal differentiation (Hinz and Jücker, 2019). As a plus, this rule is strengthened by the conditioning on the confounder "age at diagnosis", which removes spurious correlations.

6.3.2 Cancer Promotion and Precocious Metastasization

When considering a gene altered in the presence of either a somatic, LOH or hypermethylation event, strong effects are linked to the molecular basal-like subtype. *ALLSTAR* reports the combination of aberrations occurring in ITGB3 and MAP3K (rule f) as strongly causal of the aforementioned subtype, in agreement with literature: Sesé et al. (2017), Fuentes et al. (2020) and Li et al. (2022) converge on this conclusion due to their cancer-promoting activity and inclusion in the metastatic process. Hence, ITGB3 and MAP3K have recently gained attention relatively to basal-like breast cancer, but their combination is yet to be investigated. Even more interesting is the extension of this causal rule with the alteration of RHOA (rule d): the score of this expanded rule is even higher in association with basal-like subtype and it can be explained by the association of the outcome with precocious metastasization in accordance to RHOA's anti-metastatic function (Kalpana et al., 2019; Kalpana et al., 2021; Privat et al., 2020).

6.3.3 A Constant in Cancer: TP53

Considering the decomposed treatments, rules pertaining to even more specific mechanisms are retrieved. Besides the strong positive effect of mutated TP53, which is well ascertained in non-luminal breast cancer (Bertheau et al., 2013; Abubakar et al., 2019), even more relevant is the causal effect increase in combination with the LOH event in ATRIP (rule e). When stable, this gene is responsible for anti-proliferative signal mediation (Venere et al., 2007), but its impairment's effect is not well established in the literature. The interaction with mutated TP53 can open up new scenarios but it needs to be investigated in vivo.

6.3.4 The RB1-PHB *partnership*

Another combination found with strong roots in literature is LOH in RB1 and PHB (rule l), as well explained by Wang et al. (1999a), Wang et al. (1999b), and Wang and Faller (2008): RB1 is an important tumor suppressor gene (Herschkowitz et al., 2008), while PHB mediates anti-proliferation signaling (Jupe et al., 1996; Nuell et al., 1991; Sato et al., 1993; Sato et al., 1992), therefore their combined action, if altered, in basal-like tumors is easily explainable. The addition of the LOH in LIMD1 (rule h) is less established in breast cancer, being more associated with lung carcinoma, but its oncosuppressive role, and the correlation between LOH and mitosis, make it a potential key player in basal subtype (Huggins and Andrulis, 2008; Spendlove et al., 2008).

6.3.5 Potential Novel Targets: The Case of ERBB2

HER2-positive, basal-like, and triple-negative breast cancer are consistently determined by aberrations occurring in MST1 (Ercolani et al., 2017; Jin et al., 2021). Our findings (rules j, k) coherently overlap this knowledge, extending it by pairing MST1 and ERBB2 within the same positively-scored

rule. ERBB2 is a member of the epidermal growth factor (EGF) receptor family and its overexpression in 20-30% of invasive breast carcinomas leads to increased chemoresistance to certain chemotherapeutic agents (Tan and Yu, 2007). Its mutational impact is undefined in literature, as only ERBB2's expression abnormalities have been encountered in breast malignancies, especially in triple-negative/basal-like. Our result in this particular case is partially coherent but can enable further studies into the MST1-ERBB2 interaction in terms of mammalian carcinoma profiling. Conversely, the joint action between ERBB2, STAT3, and WNT5A (rule *m*) is more explainable. STAT3 has a pivotal role in the initiation, progression, metastasis, and immune evasion of triple-negative breast cancer (Manore et al., 2022; Qin et al., 2019), while WNT5A reduces the clonogenicity, invasiveness, migration, and proliferation of carcinoma cells, and it is also considered a therapeutic target (Kobayashi et al., 2018). The rule that *ALLSTAR* returned is not specific, as it emerged from the aggregated dataset, but it suggests a strong mutational involvement of these three genes in basal breast cancer. This being said, ERBB2 is recalled in rule *w* with RHOA: both genes offer potential reasons to be partnering in the determination of Basal subtype, but there is no clinical evidence of their combination, let alone an involvement of BRCA2 germline mutation as a confounder to condition over.

This rule is a clear example of potential relations that need to be evaluated in future studies. It is not a surprise that various rules with a high conditional effect, which is one of the main contributions of choosing a causal approach, are related to one of the most debated genes, ERBB2, suggesting its direct involvement in breast cancer carcinogenesis (see also rule *z* in combination with TP53).

6.3.6 Long Genes and Mutation Frequency

An additional point favorable to our methodology consists in the rules that have *not* appeared among the highest-scoring ones: long genes such as TTN, HMCN1, or DMD, usually harbor several mutations simply due to their size. *ALLSTAR* seems robust to this drawback, even if those genes are in the top-20 of the most somatically mutated ones in TCGA data. As a term of comparison, Saravia et al. (2019) perform a chi-square test to detect meaningful mutations in triple-negative breast cancer, identifying TTN, HMCN1, and DMD, among others, as statistically significant players in recurrent patterns of genomic alterations with a potential contribution to tumor evolution. The authors themselves acknowledge the possibility their findings may be false positives and our results support this hypothesis.

6.3.7 Pathway Enrichment with *DAVID*

As a further functional evaluation, for each analysis, the set of genes obtained by merging the alterations reported in any of the rules from *ALLSTAR* were considered. Subsequently, pathway enrichment analysis was performed with DAVID (Huang, Sherman, and Lempicki, 2009) to find statistically overrepresented biological functions (encoded in the KEGG database, Kanehisa and Goto (2000)) in each of these sets of genes. As the significance cut-off for pathways' p-value, 0.05 was conventionally selected . After that, the occurrence of each pathway, when significantly enriched, was counted over all the results of the analyses. The most represented pathway, occurring in 65% of the sets, is neurotrophin signaling pathway, whose relevance as a potential therapeutic target for breast cancer has been previously ascertained in preclinical studies (Hondermarck, 2012). Interestingly, the breast cancer pathway (*KEGG: hsa05224*) occupies one of the top spots with a 57% of occurrence, alongside others known to be relevant pathways such as Rap1 and PI3K-Akt signaling (Gil, 2014; Zhang et al., 2017). Additionally, the fluid shear stress and atherosclerosis (*KEGG: hsa05418*) scored an occurrence of 60%: the impact of this process in breast cancer, and in oncogenesis in general, is still unclear. However, this result seems to endorse some preliminary findings: according to Choi et al. (2019), in addition to promoting hematopoietic growth, biomechanical forces seem to be significant microenvironmental variables in the generation of cancer stem-like cells (CSLCs) or tumor-initiating cells (TICs) in cancer metastasis.

6.4 Stability on Real-World Data

Lastly, a stability-like experiment was performed on these data. Since most of the interactions between genes that are publicly available are cursed by a serious level of uncertainty, the impact of using a high confidence Protein-Protein Interaction network, as input knowledge, was evaluated. Conveniently, the original PPI from Reactome (see Section 6.1) is featured with a score ($s \in [0, 1]$) for each pair of genes, representing the confidence of their edge in the interaction network. Therefore a *certain* subnetwork could be extracted from the original one. To build the experiment, every pair with a score lower than 1 was removed, thus keeping only high confidence links, and *ALLSTAR* run with the same data inputs and parameters as described in Section 6.3, with the exception of the PPI. The results obtained running *ALLSTAR* with the two input PPIs were compared. Keeping Table 6.1 as reference, a total of 7 rules out of 26 (27% of the reference), and specifically, rules b, h, l, m, r, u, and v, were not retrieved in this analysis. These results show that most of the rules found by *ALLSTAR* including lower confidence interactions, are still reported using only high-confidence interactions. To be more precise, the excluded rules r and u were not extensively characterized in the results section due to lack of literature support. Conversely, rules b, h, l, m, and v were labeled as potentially novel discoveries: as motivated in Section 6.3, these rules refer to proteins whose impact on breast cancer is debated.

Even if their role in breast cancer physiology is not specifically supported by sufficient literature, the underlying biological mechanisms are explainable, either because of their genetic properties and functionalities, or the existence of an analogous biological process in other cancer types. Overall, these results show that *ALLSTAR* can focus on well-characterized mechanisms by including only high-confidence interactions, but also that *ALLSTAR* can be used to pinpoint potential novel discoveries by including lower-confidence interactions.

6.5 Translation to Medical Practice

The aforementioned results offered by *ALLSTAR* are promising, given their wide overlap with breast cancer literature. Moreover, they seem to satisfy the prerequisites I conveyed in the first chapter of this thesis: in addition to being able to describe known biological mechanisms, *ALLSTAR* improves the state of the art in the field of computational oncology and suggests potentially novel combinations of alterations, defining a certain cancer subtype. I would stress the latter point: as already pointed out, cancer is an extremely intricate disease, and one of the keys to better understand it is represented by the unveiling of new mechanisms that drive its development and progression.
This project was built towards this goal: in Subsection 6.3.5, I acknowledged the potential role of ERBB2 in breast cancer physiopathology. The combinations involving this gene, as well as other ones not overlapping literature, represent the real added value of this project, and, if confirmed by subsequent *in vitro* experiments, may improve patients' stratification, as well as response to personalized therapy. As a plus, the algorithm is generalized for cancer and scalable, therefore it is not restricted to a certain tumor type. In fact, the ultimate objective of this thesis is to offer a powerful tool that focuses on promising genetic targets for patients affected by diverse cancer subtypes.

Chapter 7

Conclusions

This thesis aimed at presenting the development of a novel tool, *ALLSTAR*, able to infer combinations of somatic alterations that are causal to a specific cancer phenotype. This is in contrast to previous approaches focusing on correlations. Despite the high number of required computations, *ALLSTAR* can take as input mutational data measured in large cohorts of cancer patients, thanks to its optimized branch-and-bound approach. This novel tool reports the highest-scoring rules defined on several interactions and integrates prior information in the form of a graph to focus on functionally related alterations. It also uses an iterative procedure to identify diverse rules to tackle tumor heterogeneity.

Throughout the project, one of the biggest challenges has been the translation of a rigid mathematical framework, offered by the causal notation, into a completely different context. In fact, this project drew inspiration from Budhathoki, Boley, and Vreeken (2021), an excellent theoretical work that presented simple, yet effective results on a rather limited dataset, the *Titanic* database. The authors offered intuitive results supported by logical reasoning: in synthesis, being either a woman, a child, or simply a rich person was a strong causal feature for survival, and it makes complete sense. But the logic behind cancer development and evolution is way more complicated to grasp, especially due to the inner intricacy of the disease. Where do clinical variables stand within the causal framework? What about germline and somatic alterations? Data types best fitting the model are not as intuitive.
As an additional challenge, the code provided by Budhathoki, Boley, and Vreeken (2021) had trouble returning correct answers when fed with toy synthetic datasets. Even after contacting the authors, I was not able to obtain the expected results the theory should guarantee. Honestly, this bump in the road has been a real blessing in disguise, since it allowed the possibility to build from the ground up a tool that not only advances the state of the art, but also translates and adapts its powerful means to a relevant subject. For example, the integration of the protein-protein interaction graph and the clean-up procedure are practical solutions to the specific problem at hand, and their addition has been eased by this possibility to start from scratch.

Furthermore, the extensive experimental evaluation, both with synthetic and real-world data, shows that *ALLSTAR* is an efficient and effective tool, and that it is able to identify well-supported causal relations from cancer data. In Section 6.3, I presented these relations from a biological point of view, obviously with the support of an experienced oncologist. Among the

best rules (see Table 6.1), most of them either are known in breast cancer physiology, or capture known mechanisms, while some other alterations' combinations take a stand in debated topics significant to the oncologic subject. Thus the results are satisfactory in terms of relevance to the field of application, and they offer potentially good targets for an eventual *in vivo* follow-up analysis.

Despite building up from solid theory and existing research, there is still room for improvement. Concerning the methodological part, in Section 3.2.1, I introduced the method to correct for multiple hypotheses testing, the Bonferroni correction, and mentioned its conservative nature. In fact, it assumes the worst-case scenario between every generated null hypothesis, when computing the bound, and the estimate of the effect is then extremely affected by the correction term. On the one side, rules with a positive score are guaranteed to be significant with confidence α, and the overall results is an essential list containing the very best it can be found. On the other side, *ALLSTAR* could be discarding potentially interesting interactions, especially in the case of datasets with a very high number of possible rules to investigate. A possible future direction consists in integrating *Rademacher averages* as bounding technique to control the FWER. Leaving the rigorous mathematical description to the reader (Bartlett and Mendelson, 2002), Rademacher averages come from statistical learning theory and they are frequently applied to quantify the complexity of a family of functions: by doing so, they offer a probabilistic method to limit the variance between the empirical means of the functions in the family and their expected values. The immediate advantage is the computation of a *data-dependent bound* rather than an absolute one. To the best of my knowledge, their application to biological problems has not been proven, yet.

Another future possibility, on the biological side, is the evaluation of *ALLSTAR*'s performances on a wider Pan-Cancer cohort of patients. This could unveil novel combinations and mechanisms, highly specific for a certain phenotype in a specific cancer. One possible challenge could be the necessity to line up a conspicuous number of domain experts to interpret the results, at least one for each cancer type. Additionally, the integration of more omics data, such as RNAseq, could be an interesting upgrade, but it definitely would require a careful homogenization with the causal framework assumptions, as described in Section 3.1.

In conclusion, this thesis presented an innovative traslational tool that combines elements of statistics, probability, and informatics, with the direct application to oncology. *ALLSTAR* proposes a reliable causal methodology to retrieve genetic targets for cancer phenotypes, starting from raw somatic alteration data that are publicly available online. The statistical guarantees have been extensively demonstrated with multiple synthetic experiments, therefore I strongly believe the results on real data could offer both confirmation of previous knowledge and potentially novel insights on breast cancer physiology, differentiation and, hopefully, treatment.

Appendix A

Supplemental Methods

A.1 Computational Problem Definition

We now define the computational problem at the core of finding causal rules. In particular, we consider the problem of finding the rule with the largest positive effect on a target variable, defined as follows.

Definition 1. ***Max Positive CRD problem.*** *Consider variables $\mathbf{Z} \cup \mathbf{X}$ and a target variable Y. Find the rule σ^* with i) $e(\sigma^*) > 0$ and ii) $\sigma^* = \arg\max_\sigma e(\sigma)$.*

The *Max Positive CRD* problem is a simplified version of the problem of finding the rule with largest positive effect from data, since it assumes that one has access to the *exact* probabilities for the events of interests, while, in practice, such probabilities are estimated from an observational dataset (see Section 2). Nonetheless, we prove that the problem above is computationally difficult. In particular, we prove that finding the causal rule with the maximum effect is NP-hard, even when no confounder is considered (i.e., when $\mathbf{Z} = \emptyset$) and the true probabilities are described by a Bayesian Network, that is a convienent mathematical way to represent causal relations between variables. Formally, Bayesian network (BN) is defined as a tuple $< \mathcal{G}, p >$ where $\mathcal{G} =< \mathbf{V}, \mathbf{E} >$ is a directed acyclic graph for which $\mathbf{V} = \mathbf{X} \cup \mathbf{Z} \cup \{Y\}$ and there is an edge from $V_i \in \mathbf{V}$ to $V_j \in \mathbf{V}$ only if V_i is a cause of V_j w.r.t. Pearl's do-notation (Pearl, 2009), and p is a probability distribution function over $\mathbf{V}$.

We now define the problem aforementioned problem, that we call the `MaxCRD` problem.

Definition 2. ***MaxCRD Problem.*** *Given a Bayesian Network B, output $\top$ if the rule $\sigma^* = \arg\max_\sigma |e(\sigma)|$ with the highest absolute effect has a non-zero effect.*

A.2 Proof of NP-Hardness for `MaxCRD` Problem

The following theorem proves that the `MaxCRD` problem is computationally difficult.

Theorem 2. *MaxCRD is NP-hard.*

Proof. We prove that `MaxCRD` is NP-hard by reducing from SAT. The proof is divided in two steps: first we show a polynomial-time reduction of an input of SAT to an input of `MaxCRD`, and then we show that solving `MaxCRD` on such

input allows to derive a solution to SAT in time polynomial on the original instance.

We start by describing the reduction from SAT. Let $\psi(\mathbf{X})$ be a boolean formula over variables in $\mathbf{X}$. Let us define $\mathcal{G} =< \mathbf{V}, \mathbf{E} >$ with $\mathbf{V} = \mathbf{X} \cup \{Y\}$ and $\mathbf{E} = \{X_i \rightarrow Y | X_i \in \mathbf{X}\}$. Let us define each $X_i \sim \mathcal{B}(0.5)$ be a Bernoulli distribution with probability $p(X_i = 0) = p(X_i = 1) = 0.5$. Let Y take values in $\{0, 1\}$ and let $p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n) = 1$ if and only if $\psi((x'_1, ..., x'_n)) = \top$ else $p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n) = 0$[1]. We then define the BN $B =< \mathcal{G}, p >$ as the reduced input for `MaxCRD`.

We now prove that solving `MaxCRD` on the reduced input leads to solving SAT in polynomial time on the original instance by proving that (i) if `MaxCRD`(B) = $\top$ then $\psi(\mathbf{X}) = \top$ and (ii) if `MaxCRD`(B) = $\bot$ then we can build a polynomial-time algorithm that solves SAT.

Let us prove (i). If `MaxCRD`(B) = $\top$ then $\exists \sigma | e_{corr}(\sigma) \neq 0$ that is $p(Y = y | \sigma = \top) - p(Y = y | \sigma = \bot) \neq 0$. By construction, we have two cases: $y = 1$ or $y = 0$. If $y = 1$ then $\psi(\mathbf{X})$ is satisfiable by construction since at least one between $p(Y = 1 | \sigma = \top)$ and $p(Y = 1 | \sigma = \bot)$ is positive. (Note that $\sigma = \bot$ corresponds to all assignments of variables $\mathbf{X}$ for which rule σ is not satisfied, and $p(Y = 1 | \sigma = \bot) > 0$ if and only if at least one such assignment lead to $Y = 1$, that by definition implies that such assignment satisfies $\psi(\mathbf{X})$.) If $y = 0$ then we notice that the same rule evaluated on $y = 1$ has a non-zero effect given that $y = 1$ is $y = 0$'s complementary event therefore $p(Y = 0 | \sigma = \bot) = 1 - p(Y = 1 | \sigma = \bot)$ (and the same holds for $\sigma = \bot$).

Let us prove (ii). If `MaxCRD`(B) = $\bot$ then $\forall \sigma$ we have $p(Y = y | \sigma = \top) = p(Y = y | \sigma = \bot) = p(Y = y)$ that is the value of Y is independent on $\mathbf{X}$ assignments. This means that $\psi(\mathbf{X})$ is either a tautology or a contradiction[2] and by evaluating $\psi(\mathbf{X})$ on any assignment we can distinguish between the two cases. □ □

As stated before, in practice we do not have access to the exact probabilities and, therefore, to the exact effect $e(\sigma)$ for a rule σ. We are therefore interested in finding the rule with the largest positive *reliable* effect from an observational dataset, which we formalize in the problem below.

Definition 3. ***`Max Reliable Positive CRD` Problem.*** *Consider an observational dataset $\mathcal{D}$ on variables $\mathbf{Z} \cup \mathbf{X} \cup \{Y\}$ and a confidence level $\alpha \in (0, 1)$. Find the rule σ^* such that i) $\hat{e}_{rel}(\sigma^*) > 0$ and ii) $\sigma^* = \arg\max_\sigma \hat{e}_{rel}(\sigma)$.*

[1] Note that the probability distribution function is fully specified since $p(Y = 0 | X_1 = x'_1, ..., X_n = x'_n) = 1 - p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n)$.

[2] If not, then it would be possible to discover a rule with non-zero effect $\sigma : X_1 = x'_1 \wedge ... \wedge X_n = x'_n$ on all elements of $\mathbf{X}$. By construction, in fact, $p(Y = 1 | \sigma = \top) \in \{0, 1\}$ since it evaluates on just one element, and $p(Y = 1 | \sigma = \bot) \neq p(Y = 1 | \sigma = \top)$ otherwise the value of Y would be constant.

A.3 *ALLSTAR*: Detailed Methodology

A.3.1 Algorithm Description

ALLSTAR starts by computing the total number of *candidate* rules of length at most ℓ (that is the number of connected subgraphs in G of length at most ℓ) and then calculates the correct threshold α_c for each confidence bound using Bonferroni correction (line 1). The rule discovery is then performed in k iterations (line 6). In each iteration, a breadth-first search (BFS) of the lattice defined by set of all possible rules with at most ℓ alterations is performed by using a FIFO queue `Q` and its (standard) operations `enqueue` and `dequeue`. During the BFS, the best rule $\sigma_{\max}$, and its maximum reliable estimated effect $\hat{e}_{\max}$, discovered during the exploration are maintained. After the initialization of $\sigma_{\max}$ and $\hat{e}_{\max}$ (line 4), the queue `Q` is initialized by inserting the rules containing a single alteration (line 5). (Note that *ALLSTAR* can also consider the *absence* of an alteration as part of a rule (i.e. $X_i = 0$); for clarity's sake, this is not reported in Algorithm 1.) The BFS then proceeds by extracting the current rule σ (line 7) until `Q` is not empty (line 6). When a rule σ is extracted from `Q`, an upper bound to its reliable effect is computed with the function `computeRelATE`$(\sigma, y, \mathbf{Z}, \alpha_c)$. If such upper bound is greater than $\hat{e}_{\max}$ (line 8) then the (exact) reliable effect estimate $\hat{e}_\sigma$ of σ is computed (line 9), and the values $\hat{e}_{\max}, \sigma_{\max}$ are updated if $\hat{e}_\sigma > \hat{e}_{\max}$ (line 10). Then, the rules that are obtained by expanding σ, obtained with the function `expand`(σ, G, ℓ), are added to the queue (lines 11-11). `expand`(σ, G, ℓ) returns all rules (with at most ℓ alterations) that are obtained by adding to σ one alteration that must be connected in G to at least one alteration of σ. When the BFS completes, the best rule $\sigma_{\max}$ is added to the output set if its estimated reliable effect is positive (line 12), and the set $\mathbf{X}$ of alterations is updated (line 13) to avoid discovering highly-overlapping, redundant, rules (see below). At the end, the set of at most top-k rules is reported in output (line 16).

A.3.2 Subroutines

ALLSTAR exploits three subroutines `calculateRulesNumber`, `upperBoundRelATE`, and `computeRelATE` that will be briefly explained in the following (and whose Python code is available online): **`calculateRulesNumber`** takes as input a graph G and the maximum rule length ℓ and outputs the number of connected subgraphs of length at most ℓ between elements in G. It is used to calculate the total number of possible rules under study, which is the amount of test performed in the worst case, and the pseudocode is described in Algorithm 2.
`upperBoundRelATE` takes as an input a rule σ, the value y for target Y, a set of confounders $\mathbf{Z}$, and a threshold α_c corrected for multiple hypotheses testing, and it outputs the tight optimistic upper bound to the effect for the rule proposed by Budhathoki, Boley, and Vreeken, 2021. It is used by the branch-and-bound algorithm for deciding whether to compute study a specific branch (i.e. all children of a specific rule) or to avoid the computation because the

Algorithm 2: `calculateRulesNumber`

Input: Graph $G = (\mathbf{V}, \mathbf{E})$, maximum rule length ℓ
Output: Number N of connected subgraphs of length at most ℓ between elements in G

$P \leftarrow \emptyset$;
$Q \leftarrow \emptyset$;
for $X \in \mathbf{V}$ **do**
  $P \leftarrow P \cup \{X\}$;
  $Q \leftarrow Q \cup \{X\}$;
for $i \leftarrow 1$ *to* $\ell - 1$ **do**
  $L \leftarrow \emptyset$;
  for $q \in Q$ **do**
    for $X \in q$ **do**
      for $e \in \mathbf{E}$ **do**
        if $X \in e$ & $e \setminus \{X\} \notin q$ **then**
          $L \leftarrow L \cup \{q \cup \{e \setminus \{X\}\}\}$
  Remove duplicates from L;
  $P \leftarrow P \cup \{L\}$;
  $Q \leftarrow L$;
return *size(P)*;

best solution found in such branch would never improve the current best solution (i.e., the incumbent) $\hat{e}_{max}$.
More specifically, let us consider a rule σ and a more specific rule $\sigma' = \sigma \wedge \pi_j$. Let us define the quantity $\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z})$ on the elements for which $\mathbf{Z} = \mathbf{z}$ holds as

$$\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_\sigma \in \{0,1,\dots,a_\sigma\}} \frac{a'_\sigma + 1}{a'_\sigma + 2} - \frac{n_1 - a'_\sigma + 1}{n - a'_\sigma + 2} - \frac{\beta(\alpha_c)}{2\sqrt{a'_\sigma + 2}} - \frac{\beta(\alpha_c)}{2\sqrt{n - a'_\sigma + 2}} \tag{A.1}$$

where $\beta(\alpha_c)$ is the $1 - \alpha_c/2$ quartile of the standard normal distribution, n is the number of instances taken into account (i.e. with $\mathbf{Z} = \mathbf{z}$), n_1 of which have $Y = y$, and a_σ is the number of instances for which σ holds, $\mathbf{Z} = \mathbf{z}$ and $Y = y$. The upper bound is then defined as

$$U(\sigma') = \sum_{\mathbf{z}} (\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) \hat{p}(\mathbf{Z} = \mathbf{z})) \tag{A.2}$$

where $\hat{p}(\mathbf{Z} = \mathbf{z})$ is the empirical probability of $\mathbf{Z}$ taking value $\mathbf{z}$. Differently from Budhathoki, Boley, and Vreeken, 2021 our bound uses a confidence level $\alpha_c = \alpha / N$, where N is the total number of rules considered by the algorithm, to account for the multiple hypothesis testing problem.
`computeRelATE` takes as an input a rule σ, the value y for target Y, a set of confounders $\mathbf{Z}$, and a threshold α_c and calculates the reliable effect of the rule $\hat{e}_{rel}(\sigma)$ as described in Section 2.3. More specifically, let us define $\hat{p}(Y =$

$y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}}{n_{\sigma=\top}}$ where $n_{\sigma=\top}$ is the number of instances for which $\sigma = \top$ (i.e, σ is true), and $n_{Y=y,\sigma=\top}$ is the number of instances for which $Y = y$ and $\sigma = \top$. Analogously we have $\hat{p}(Y = y|\sigma = \bot) = \frac{n_{Y=y,\sigma=\bot}}{n_{\sigma=\bot}}$. In extreme cases (e.g. $\sigma = \bot$ for all instances) such quantities are ill-defined, therefore the Laplace correction is applied to the estimated conditional probability, which becomes $\hat{p}_c(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}+1}{n_{\sigma=\top}+2}$. The returned value $\hat{e}^y_{rel}(\sigma)$ is then defined as

$$\begin{aligned}\hat{e}^y_{rel}(\sigma) = \sum_{\mathbf{z}} \Bigg[\Bigg(& \hat{p}_c(Y = y|\mathbf{Z} = \mathbf{z}, \sigma = \top) + \\ & - \hat{p}_c(Y = y|\mathbf{Z} = \mathbf{z}, \sigma = \bot) + \\ & - \frac{\beta(\alpha_c)}{2\sqrt{n_{\mathbf{Z}=\mathbf{z},\sigma=\top}}} - \frac{\beta(\alpha_c)}{2\sqrt{n_{\mathbf{Z}=\mathbf{z},\sigma=\bot}}} \Bigg) \hat{p}(\mathbf{Z} = \mathbf{z}) \Bigg].\end{aligned}$$

Bibliography

Abubakar, Mustapha et al. (2019). "Clinicopathological and epidemiological significance of breast cancer subtype reclassification based on p53 immunohistochemical expression". In: *NPJ breast cancer* 5.1, pp. 1–9.

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami (1993). "Mining association rules between sets of items in large databases". In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216.

Arnedo-Pac, Claudia et al. (2019). "OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers". In: *Bioinformatics* 35.22, pp. 4788–4790.

Bartlett, Peter L and Shahar Mendelson (2002). "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov, pp. 463–482.

Bertheau, Philippe et al. (2013). "p53 in breast cancer subtypes and new insights into response to chemotherapy". In: *The Breast* 22, S27–S29.

Bodily, Weston R et al. (2020). "Effects of germline and somatic events in candidate BRCA-like genes on breast-tumor signatures". In: *PloS one* 15.9.

Bonferroni, Carlo (1936). "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, pp. 3–62.

Budhathoki, Kailash, Mario Boley, and Jilles Vreeken (2021). "Discovering reliable causal rules". In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, pp. 1–9.

Cartwright, Nancy (2010). "What are randomised controlled trials good for?" In: *Philosophical studies* 147.1, pp. 59–70.

Choi, Hye Yeon et al. (2019). "Hydrodynamic shear stress promotes epithelial-mesenchymal transition by downregulating ERK and GSK3β activities". In: *Breast Cancer Research* 21.1, pp. 1–20.

Cibulskis, Kristian et al. (2013). "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples". In: *Nature biotechnology* 31.3, pp. 213–219.

Cifuentes-Bernal, Andres M et al. (2022). "Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories". In: *Briefings in Functional Genomics* 21.6, pp. 455–465.

Concato, John, Nirav Shah, and Ralph I Horwitz (2000). "Randomized, controlled trials, observational studies, and the hierarchy of research designs". In: *New England journal of medicine* 342.25, pp. 1887–1892.

Cortés-Ciriano, Isidro et al. (2022). "Computational analysis of cancer genome sequencing data". In: *Nature Reviews Genetics* 23.5, pp. 298–314.

Cowen, Lenore et al. (2017). "Network propagation: a universal amplifier of genetic associations". In: *Nature Reviews Genetics* 18.9, pp. 551–562.

Erber, Ramona and Arndt Hartmann (2020). "Histology of luminal breast cancer". In: *Breast Care* 15.4, pp. 327–336.

Ercolani, Cristiana et al. (2017). "Expression of phosphorylated Hippo pathway kinases (MST1/2 and LATS1/2) in HER2-positive and triple-negative breast cancer patients treated with neoadjuvant therapy". In: *Cancer Biology & Therapy* 18.5, pp. 339–346.

Fisher, Rosie, Lazos Pusztai, and C Swanton (2013). "Cancer heterogeneity: implications for targeted therapeutics". In: *British journal of cancer* 108.3, pp. 479–485.

Fuentes, Pedro et al. (2020). "ITGB3-mediated uptake of small extracellular vesicles facilitates intercellular communication in breast cancer cells". In: *Nature communications* 11.1, pp. 1–15.

Gil, Eva Maria Ciruelos (2014). "Targeting the PI3K/AKT/mTOR pathway in estrogen receptor-positive breast cancer". In: *Cancer Treat. Rev.* 40.7, pp. 862–871.

Goldman, Mary et al. (2015). "The UCSC cancer genomics browser: update 2015". In: *Nucleic acids research* 43.D1, pp. D812–D817.

Gradek, Frédéric et al. (2019). "Sodium channel Nav1. 5 controls epithelial-to-mesenchymal transition and invasiveness in breast cancer cells through its regulation by the salt-inducible kinase-1". In: *Scient. Rep.* 9.1, pp. 1–14.

Herschkowitz, Jason I et al. (2008). "The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas". In: *Breast Cancer Research* 10.5, pp. 1–13.

Hinz, Nico and Manfred Jücker (2019). "Distinct functions of AKT isoforms in breast cancer: a comprehensive review". In: *Cell Comm. and Sign.* 17.1, pp. 1–29.

Hondermarck, Hubert (2012). "Neurotrophins and their receptors in breast cancer". In: *Cytokine & growth factor reviews* 23.6, pp. 357–365.

Huang, Da Wei, Brad T Sherman, and Richard A Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources". In: *Nature Prot.* 4.1.

Huggins, Christopher Jack and Irene L Andrulis (2008). "Cell cycle regulated phosphorylation of LIMD1 in cell lines and expression in human breast cancers". In: *Cancer letters* 267.1, pp. 55–66.

Jin, Xiang et al. (2021). "MST1 inhibits the progression of breast cancer by regulating the Hippo signaling pathway and may serve as a prognostic biomarker". In: *Molecular Medicine Reports* 23.5, pp. 1–12.

Jupe, Eldon R et al. (1996). "Prohibitin in breast cancer cell lines: loss of antiproliferative activity is linked to 3'untranslated region mutations." In: *Cell Growth Differ.* 3, p. 4.

Kalpana, Gardiyawasam et al. (2019). "Reduced RhoA expression enhances breast cancer metastasis with a concomitant increase in CCR5 and CXCR4 chemokines signaling". In: *Scientific reports* 9.1, pp. 1–12.

Kalpana, Gardiyawasam et al. (2021). "The RhoA dependent anti-metastatic function of RKIP in breast cancer". In: *Scientific reports* 11.1, pp. 1–14.

Kanehisa, Minoru and Susumu Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1, pp. 27–30.

Kendall, Maurice G (1938). "A new measure of rank correlation". In: *Biometrika* 30.1/2, pp. 81–93.

Kobayashi, Yoshie et al. (2018). "Wnt5a-induced cell migration is associated with the aggressiveness of estrogen receptor-positive breast cancer". In: *Oncotarget* 9.30.

Kotsiantis, Sotiris and Dimitris Kanellopoulos (2006). "Association rules mining: A recent overview". In: *GESTS International Transactions on Computer Science and Engineering* 32.1, pp. 71–82.

Kraya, Adam A et al. (2019). "Genomic Signatures Predict the Immunogenicity of BRCA-Deficient Breast CancerImmunogenetic Signatures of BRCA1/2 Breast Cancer". In: *Clinical Cancer Research* 25.14, pp. 4363–4374.

Li, Cheukfai et al. (2022). "Spectrum of MAP3K1 mutations in breast cancer is luminal subtype-predominant and related to prognosis". In: *Oncology Letters* 23.2, pp. 1–12.

Liu, Yijun et al. (2022). "Identification of key somatic oncogenic mutation based on a confounder-free causal inference model". In: *PLoS Computational Biology* 18.9, e1010529.

Luo, Qianxuan et al. (2020). "The functional role of voltage-gated sodium channel Nav1. 5 in metastatic breast cancer". In: *Front. in Pharmaco.* 11.

Manore, Sara G et al. (2022). "IL-6/JAK/STAT3 Signaling in Breast Cancer Metastasis: Biology and Treatment". In: *Frontiers in Oncology* 12.

Mansouri, Mehrdad et al. (2022). "Aristotle: stratified causal discovery for omics data". In: *BMC bioinformatics* 23.1, pp. 1–18.

McCart Reed, Amy E et al. (2021). "The genomic landscape of lobular breast cancer". In: *Cancers* 13.8, p. 1950.

Mularoni, Loris et al. (2016). "OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations". In: *Genome biology* 17.1, pp. 1–13.

Nuell, MJ et al. (1991). "Prohibitin, an evolutionarily conserved intracellular protein that blocks DNA synthesis in normal fibroblasts and HeLa cells". In: *Molecular and cellular biology* 11.3, pp. 1372–1381.

Parker, Joel S et al. (2009). "Supervised risk predictor of breast cancer based on intrinsic subtypes". In: *Journal of clinical oncology* 27.8, p. 1160.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Pereira, Bernard et al. (2016). "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes". In: *Nature communications* 7.1, pp. 1–16.

Piraino, Scott W et al. (2019). "Mutations: Driver versus passenger". In.

Privat, Maud et al. (2020). "A high expression ratio of RhoA/RhoB is associated with the migratory and invasive properties of basal-like Breast Tumors". In: *International journal of medical sciences* 17.17, p. 2799.

Qin, Jiang-Jiang et al. (2019). "STAT3 as a potential therapeutic target in triple negative breast cancer: a systematic review". In: *Journal of Experimental & Clinical Cancer Research* 38.1, pp. 1–16.

Reyna, Matthew A et al. (2020). "Pathway and network analysis of more than 2500 whole cancer genomes". In: *Nature communications* 11.1, pp. 1–17.

Riaz, Nadeem et al. (2017). "Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes". In: *Nature Comm.* 8.1, pp. 1–7.

Rosenbaum, Paul R, PR Rosenbaum, and Briskman (2010). *Design of observational studies*. Vol. 10. Springer.

Saravia, César H et al. (2019). "Patterns of mutation enrichment in metastatic triple-negative breast cancer". In: *Clinical Medicine Insights: Oncology* 13.

Sarto Basso, Rebecca, Dorit S Hochbaum, and Fabio Vandin (2019). "Efficient algorithms to discover alterations with complementary functional association in cancer". In: *PLoS computational biology* 15.5.

Sato, Takaaki et al. (1992). "The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer". In: *Cancer research* 52.6.

Sato, Takaaki et al. (1993). "The human prohibitin (PHB) gene family and its somatic mutations in human tumors". In: *Genomics* 17.3, pp. 762–764.

Sesé, Marta et al. (2017). "Hypoxia-mediated translational activation of ITGB3 in breast cancer cells enhances TGF-β signaling and malignant features in vitro and in vivo". In: *Oncotarget* 8.70.

Silverstein, Craig et al. (2000). "Scalable techniques for mining causal structures". In: *Data Mining and Knowledge Discovery* 4.2, pp. 163–192.

Spendlove, Ian et al. (2008). "Differential subcellular localisation of the tumour suppressor protein LIMD1 in breast cancer correlates with patient survival". In: *International journal of cancer* 123.10, pp. 2247–2253.

Tan, Ming and Dihua Yu (2007). "Molecular mechanisms of erbB2-mediated breast cancer chemoresistance". In: *Breast Cancer Chemosens.*

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793.

The International Cancer Genome Consortium (2010). "International network of cancer genome projects". In: *Nature* 464.7291, pp. 993–998.

Vandin, Fabio, Eli Upfal, and Benjamin J Raphael (2012). "De novo discovery of mutated driver pathways in cancer". In: *Genome research* 22.2, pp. 375–385.

Venere, Monica et al. (2007). "Phosphorylation of ATR-interacting protein on Ser239 mediates an interaction with breast-ovarian cancer susceptibility 1 and checkpoint function". In: *Cancer research* 67.13, pp. 6100–6105.

Wang, Sheng and Douglas V Faller (2008). "Roles of prohibitin in growth control and tumor suppression in human cancers". In: *Translat. Oncogeno.* 3, p. 23.

Wang, Sheng et al. (1999a). "Prohibitin, a potential tumor suppressor, interacts with RB and regulates E2F function". In: *Oncogene* 18.23, pp. 3501–3510.

Wang, Sheng et al. (1999b). "Rb and prohibitin target distinct regions of E2F1 for repression and respond to different upstream signals". In: *Molecular and cellular biology* 19.11, pp. 7447–7460.

Weinstein, John N et al. (2013). "The cancer genome atlas pan-cancer analysis project". In: *Nature genetics* 45.10, pp. 1113–1120.

Wu, Guanming, Xin Feng, and Lincoln Stein (2010). "A human functional protein interaction network and its application to cancer data analysis". In: *Genome biology* 11.5, pp. 1–23.

Zhang, Qingyang, Joanna E Burdette, and Ji-Ping Wang (2014). "Integrative network analysis of TCGA data for ovarian cancer". In: *BMC systems biology* 8, pp. 1–18.

Zhang, Yi-Lei et al. (2017). "Roles of Rap1 signaling in tumor cell migration and invasion". In: *Cancer biology & medicine* 14.1, p. 90.