
UNIVERSIDADE NOVA DE LISBOA
itqb nova

NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

NOVA
MEDICAL SCHOOL

NOVA
IMS
Information
Management
School

DESDE
INST
ME
UNIVER

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

SPECIALIZATION MULTI-OMICS FOR LIFE AND HEALTH SCIENCES

NOVA University Lisbon

September, 2025

Benchmarking Causal Reasoning Algorithms for enhanced Drug Discovery
Insights from Clarivate's pre-competitive Algorithm Benchmarking Consortium

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.3.1) [0].

Dedicatory lorem ipsum.

Acknowledgements

Acknowledgments To Filipo Ciceri as my supervisor during my all internship in Clarivate, which not only included the generation of this thesis, but all the projects developed during my stay, thank you for all the knowledge and support at any hour. To Alexandr Ishkin, the project manager that agree with my collaboration in this projects, thanks for the opportunity for learning so much in such an interesting and complex project. To Cecilia and Sahra for beliving in my capabilities. To every individual from Discovery and Translational Science Team, without any exeption, for the knowlegde, and the help, the friendship every single day. To my family, Mama, Papa, Avos, Madri Ana, Tete, Afilhado Lux, Gui for always beliving me and suporting unconditionally every single step that I give. To all my friends from Braga, from Lisbon, from Barcelona.

”

“You cannot teach a man anything; you can only help him discover it in himself.”

— Galileo, Somewhere in a book or speach
(Astronomer, physicist and engineer)

Abstract

Regardless of the language in which the dissertation is written, usually there are at least two abstracts: one abstract in the same language as the main text, and another abstract in some other language.

However, this order can be customized by adding one of the following to the file `5_packages.tex`.

```
\ntsetup{abstractorder={<LANG_1>,...,<LANG_N>}}
\ntsetup{abstractorder={<MAIN_LANG>={<LANG_1>,...,<LANG_N>}}}
```

For example, for a main document written in German with abstracts written in German, English and Italian (by this order) use:

```
\ntsetup{abstractorder={de={de,en,it}}}
```

Concerning its contents, the abstracts should not exceed one page and may answer the following questions (it is essential to adapt to the usual practices of your scientific area):

1. What is the problem?
2. Why is this problem interesting/challenging?
3. What is the proposed approach/solution/contribution?
4. What results (implications/consequences) from the solution?

Keywords: One keyword, Another keyword, Yet another keyword, One keyword more, The last keyword

Resumo

Independentemente da língua em que a dissertação está escrita, geralmente esta contém pelo menos dois resumos: um resumo na mesma língua do texto principal e outro resumo numa outra língua. No entanto, esse pedido pode ser personalizado adicionando um dos seguintes ao arquivo `5_packages.tex`.

```
\abstractorder(<MAIN_LANG>):={<LANG_1>,...,<LANG_N>}
```

Por exemplo, para um documento escrito em Alemão com resumos em Alemão, Inglês e Italiano (por esta ordem), pode usar-se:

```
\ntsetup{abstractorder={de={de,en,it}}}
```

Relativamente ao seu conteúdo, os resumos não devem ultrapassar uma página e frequentemente tentam responder às seguintes questões (é imprescindível a adaptação às práticas habituais da sua área científica):

1. Qual é o problema?
2. Porque é que é um problema interessante/desafiante?
3. Qual é a proposta de abordagem/solução?
4. Quais são as consequências/resultados da solução proposta?

Palavras-chave: Primeira palavra-chave, Outra palavra-chave, Mais uma palavra-chave, A última palavra-chave

Contents

List of Figures

Glossary

MetaBase MetaBase from Clarivate [0] *30)*

Acronyms

ABC Algorithm Benchmarking Consortium *2)*

CBDD Computational Biology for Drug Discovery *30)*
CMap Connectivity Mapping *2)*

DEGs Differentially Expressed Genes *2)*

R R Programming Language *19)*

1

Introduction

This section expounds the underlying motivation, rationale and goals for the study, emphasizing its significance in the field. It provides context by giving some background on the supporting company and the initiative. Furthermore, it outlies a read's guide of this thesis.

1.1 Motivation and Goals

The Research and development (R&D) of new drugs is a fast-growing area that has also experienced significant growth in complexity in recent years. Due to the time-consuming, costly and multidisciplinary nature of the process, drug discovery remains a challenging domain. Over half of clinical trial failures are attributed to inefficiency, underscoring the importance of identifying and validating pharmacological targets, and highlighting the lack of knowledge of the drug's mechanism of action (MoA) as one of the major barriers in drug discovery. The thorough understanding of the MoA represents a critical initial step in this process, and computational methods can accelerate this by providing more efficient and cost-effective alternatives to traditional approaches. These approaches can accurately do target identification and prioritization, thereby reducing the need for lengthy experimental trials.

A key to understanding a compound's MoA lies in transcriptomics data, which captures the molecular changes triggered by a perturbagen and reflects the system's changes in the gene expression profiles. While traditional RNA sequencing methods remain too costly for large-scale expression signatures, recent high-throughput technological advances, such as the L1000 assay, enable the cost-effective generation and analysis of large-scale omics datasets. Several existing databases provide public access to transcriptomic data from experiments involving diverse chemical and genetic perturbagens across different cell lines. These data can be exploited using various computational tools to establish the causes of specific gene expression changes in a biological system. Three primary approaches have emerged: causal reasoning, connectivity mapping and enrichment tools.

Causal reasoning, a topology-based method, utilizes a list of perturbation signatures and a biological interaction network to determine potential causes for the observed gene

expression profile. The network is defined as signed and directed graph describing relations between nodes (e.g., proteins). Efforts to assemble causal molecular relations have increased, resulting in several publicly accessible databases, such as OmniPath, which offers curated prior knowledge networks, however, some causal information remains commercially available, such as MetaBaseTM developed and curated by Clarivate.

The Connectivity Mapping (CMap) method stems from the efforts to collect and analyze perturbation signatures. It employs similarity scoring to compare a set of known MoA/compound reference signatures with a query gene expression signature resulting from a perturbagen. The principle behind CMap: the higher the similarity between the query and the reference signature, the more likely it is that the mechanism underlying the observed gene expression changes is related to a known perturbation.

On the other hand, enrichment tools take perturbation signatures as query input and utilize prior knowledge, such as regulon network or collections of perturbation-induced Differentially Expressed Genes (DEGs), as a reference. The primary function of these tools is to assess whether certain regulons or gene sets (e.g., those associated with transcription factors (TFs)) are significantly enriched in the perturbed data. Several algorithms have been developed based on this approach, each producing an enrichment score.

1.2 Scope

This project was conducted within the framework of the Algorithm Benchmarking Consortium (ABC), a subscription-based initiative led by Clarivate for pharmaceutical companies. ABC is dedicated to evaluating a wide range of computational tools for a variety of applications in the life sciences and healthcare field. The topic for this thesis is the development of the ABC's tenth use case – Causal Regulation – which focuses on benchmark and identify the most optimal tools tailored to specific needs within the drug discovery process by identifying key regulators from transcriptomics data and prior knowledge graphs. This project was conducted within the framework of the Algorithm Benchmarking Consortium (ABC), a subscription-based initiative led by Clarivate for pharmaceutical companies. ABC is dedicated to evaluating a wide range of computational tools for a variety of applications in the life sciences and healthcare field. The topic for this thesis is the development of the ABC's tenth use case – Causal Regulation – which focuses on benchmark and identify the most optimal tools tailored to specific needs within the drug discovery process by identifying key regulators from transcriptomics data and prior knowledge graphs.

1.3 Parallel Contributions

This study expands the state of art in causal reasoning using gene expression data and causal graphs by presenting a robust framework and methods for benchmarking various algorithms designed for this purpose. Beyond this primary focus, several parallel projects

with real-world challenges and novel data were developed, and enriched the consultant experience allowing for an expansion in the expertise of bioinformatics. The parallel projects, spanning various domains of computational biology, include:

Skin Microbiome Atlas The skin microbiome atlas project involved extensive scientific literature review and dataset curation, followed by a systematic re-analysis of available datasets to ensure consistency and reliability. This was done by pre-processing the raw sequencing data (. . . understand how deep I can go In terms of details – confidentiality issues)

More projects ... More projects ...

1.4 Structure

This study is organized in ? chapters. Chapter **??** introduces

2

Literature Review

This section provides an overview of the relevant research related to the topic of the project. Starting with highlighting the importance of understanding the mechanism of action for enhanced drug discovery. The following sections detail two key components for MoA elucidation through computational analysis. First, the transcriptomics data and networks serve as a foundation for the analysis. Next, it outlines the computational methods used to apply various scoring algorithms: topology-, similarity-, and enrichment-based algorithms. Together, these components form the basis of our systematic evaluation of tools for elucidating compound MoA. Finally, an overview of the best practices to perform a benchmarking study is provided.

2.1 Drug discovery: the importance of the compound's mechanism of action

[0]

The development of new drugs is a highly complex process in the R&D field. The high prevalence of complex diseases today, which collectively account for 70% of all deaths in Europe and affect approximately 25% of the population, is one of the challenges faced by this industry [12]. In addition, statistics show that de novo drug discovery has become an extensive and costly process, taking around 13 years and with a cost as high as $2 billion to develop a new drug, with clinical trials lasting an average of approximately 95 months and non-clinical phases lasting 31 months [13-15]. These challenges have led to fewer drug approvals by regulatory bodies, resulting in a significant gap between therapeutic demand and available treatments. Hence, as the current treatments become less effective, there is a strong interest in finding alternatives to optimize critical steps in the drug development pipeline and developing more advanced therapeutic methods [16]. Efforts to address these challenges are evident in the growing number of studies, both in industry and academia. However, this is still not enough to meet the growing need for new drugs. Drug repositioning (DR) has emerged as a promising cost-effective strategy to tackle the constraints faced by traditional DD by reducing the initial cost to 1/3 and the duration to 3-9 years. It continues to gain increasing attention, as nearly 30% of the drugs and vaccines

approved by the FDA are derived from this method [17, 18]. The fundamental goal of DR is to broaden the scope of the known, safe, and previously approved drugs for other diseases. This is a particularly interesting method for both perspectives of drugs and diseases. By offering ways not only to investigate treatments that have been put on hold because of failed clinical trials [19], but also for diseases where no effective treatment is known, especially rare diseases, which big pharma is not particularly interested in investing in when using traditional blinded methods, given the low financial return. Many studies have demonstrated the success of establishing new drug-disease relationships [20]. A well-known example is Sildenafil; initially identified in the 1980s as a candidate to treat angina pectoris, it was approved by the Food and Drug Administration (FDA) in 1998 to treat erectile dysfunction and later in 2005 to treat pulmonary arterial hypertension [17, 18, 21]. Another classic example is Thalidomide, originally used for sedation and morning sickness, and afterward repurposed for multiple myeloma, leprosy [17, 18], and to minimize the hippocampal neuronal loss [22]. Moreover, the low success approval rate (5%) of cancer treatments that enter phase I clinical trials led to increased attention in DR for oncology, resulting in several promising findings [18, 23]. Noteworthy cases include the schizophrenia drug Spiperone, which has been studied for its ability to induce apoptosis in colorectal cancer (CRC) cells [24], and Raloxifene, indicated for osteoporosis, which proved to be effective in reducing invasive breast cancer risk in postmenopausal women [18, 25]. Understanding how cellular signaling (Figure 1) is modulated upon disturbance within an organism is essential for identifying potential drug targets and finding new indications for an existing drug or a disease that a known drug could treat. When a drug enters a biological system, it typically interacts directly or indirectly with cellular targets and therefore regulates the activity of signaling networks and pathways, a process known as the mechanism of action (MoA) [26, 27]. These interactions impact the course of drug R&D, from initial investigation to clinical trials, and deep comprehension of these mechanisms can help uncover important biomarkers, anticipate early adverse effects, and even synergistic effects resulting from drug combinations. Nevertheless, FDA approval can be obtained without knowing the drug's MoA if the drug exhibits safety and efficacy [28, 29]. Yet, not knowing the mechanisms of the compounds can be extremely disadvantageous, as demonstrated with Dimebon, which could have taken a different course if its MoA had been initially characterized. Originally developed as an antihistamine, it later entered clinical trials, with the MoA still unknown, as a treatment for Alzheimer's but failed in the third phase of the study for not affecting cognition, since it was the activation of histamine and serotonin receptors that caused the initial observed cognitive efficacy and not the stabilization of mitochondria as first hypothesized [28, 30]. Although we refer to the target(s) of a compound as a direct interaction, this is not the case. Subsequently, there is a series of interactions that modulate the different levels of biological data, and what is "detected" at a given moment does not always linearly reflect what happened previously. Indeed, the basic definition of MoA is just the tip of the iceberg, given that the chain of reactions triggered involves various molecules and forms the cell

signaling cascade. This cascade is characterized by the pathways through which the signal moves within the cell and which lead to certain cellular responses. These pathways can also interact with each other through crosstalk [31], forming a network, which, although interconnected, are distinct concepts. The impact of a certain compound in the complex cell signaling cascade can be defined and observed on a system level by different multi-omics layers, such as genomics, transcriptomics, proteomics, metabolomics, and even phenomics, each offering an alternative perspective on the compound's bioactivity. However, finding experimentally which targets, signaling proteins, and biological pathways (MoA) are being modulated by an uncharacterized compound can be a huge barrier in the process of drug discovery. Owing to the advantages of clearly describing the MOA more quickly and affordably, along with the increasing availability of high-throughput data, in silico methods have become an attractive option. These methods act as a pre-filtering process, helping to identify and produce mechanistic hypotheses for further experimental validation [28]. Many accessible computational resources integrate "omics" data with prior knowledge graphs, such as gene regulatory networks, to enhance and pinpoint the drug's potential cellular targets. However, choosing the most suitable data and computational tool to employ in each situation is not always simple, so it is important to determine exactly the scientific question that needs to be addressed. By doing this, researchers may choose the most appropriate data type and bioinformatics tools to efficiently study the compound's mechanism of action.

2.2 Transcriptomics data

Transcriptomics data provide a comprehensive view of gene expression level changes in response to a compound. Following the compound's perturbation, this data will reflect the differential mRNA expression by capturing modulated signaling and transcription factor activity changes triggered by the perturbagen, therefore, it is a type of data that plays a crucial role in understanding a compound's MoA. After a cellular disturbance, changes in gene expression levels give rise to a perturbed gene expression signature [32]. In 2000, a reference database was built from a compilation of Saccharomyces cerevisiae gene expression signatures derived from pharmacological and genetic perturbations, and the same authors even projected that to generate a wider collection of reference signatures, less expensive gene expression experiments would be required [33-35]. Since then, and with the growing interest in drug discovery optimization, several databases have emerged to aggregate and publicly provide perturbagen transcriptomic signatures. These databases make it possible to extract information on how gene expression is modified by certain manipulations and treatments (**??**). DrugMatrix was the first larger molecular toxicology database, created in 2006 by Iconix Pharmaceuticals, later acquired by the National Institute of Environmental Health Sciences (NIEHS), and has been publicly available since 2011 [33, 36, 37], comprises the gene expression response obtained through microarray to more than 600 perturbagens in rat tissues. Additionally, it provides information on chemical

treatments related to histopathology, hematology, and clinical chemistry, enabling the investigation of certain types of toxicity [38]. However, studies in vivo limit the number of disturbances that can be studied, entail high costs that make it impractical to generate data on a large scale, as well as all the associated ethical implications [9]. In addition, transcriptional changes are usually specific to each cell, so a precise analysis of transcriptomic changes before and after perturbations should be carried out, considering the cell type, which requires even more resources [39]. Connectivity mapping (CMap) is a resource that emerged from an attempt to address the lack of a systematic way to establish connections between the MoA of chemical compounds, diseases, and biological processes through pattern matching between signatures derived from perturbations applied to different cell lines. The first version of CMap (CMap 1.0) generated 453 signatures derived from 164 distinct perturbations of small molecules applied at two time points under certain concentrations to four human cell lines (MCF7, PC3, HL60, and SKMEL5) [9]. Although the goal could be achieved using signatures from various -omics layers, to capture cellular response at different levels, this resource focused only on mRNA expression data through DNA microarrays. While the concept behind this database has become extensively utilized, the data generated by this pilot project was too small for the potential application of this tool. The perturbations were few and undiversified in terms of perturbations and cell lines. The recent advances in high-throughput technology, allowed large data acquisition, as is the case with the L1000 assay. The Library of Integrated Network-Based Cellular Signatures (LINCS) program extended the CMap to a second version (CMap 2.0 or LINCS L1000) that measured the expression of 978 "landmark" genes, key representatives of various biological processes, in cells treated with different perturbagens [40]. Using the Luminex L1000 platform, these landmark genes are directly measured, and computational methods and in silico imputation then infer the expression of 11350 additional genes, resulting in a wider reconstruction of the genome profile [47]. In a preliminary phase, LINCS released over 6000 signatures from around 1300 small compounds, many of which were FDA-approved [33], and today it includes more than one million gene expression profiles from over 20000 chemical and genetic perturbagens, tested at multiple time points and doses across various human cell lines. Currently, the dataset is more than a thousand times larger than the CMap pilot dataset. The perturbagens include small-molecule compounds and genetic treatments, such as gene knockdowns using short hairpin RNA (shRNA) and/or CRISPR and induced over-expression (OE). LINCS L1000 provides data in five levels. Levels 1 to 4 contain data at different pre-processing stages, and Level 5 contains the final signatures, where replicates, usually three per treatment, are combined into a single differential expression vector. This level is recommended for most downstream analyses [40]. The extension of data by LINCS L1000 has broadened the association between changes in gene expression caused by certain disorders, enhancing the repositioning of drugs and contributing to the generation of testable hypotheses about the MoA of less characterized compounds. However, the additional genes are imputed and not measured directly, which can lead to some inaccuracy. In addition, the inherent complexity of cellular

responses must be considered, since gene expression snapshots may not fully capture the dynamic nature of biological processes and do not always correlate perfectly with protein expression due to post-translational modifications [27]. Despite these limitations, several computational methods have been developed to apply these data for leveraging the drug discovery process. However, the interpretation of the results must be careful to consider the static and imputed nature of the data. These methods will be described later in this chapter. The Cancer Drug-induced Gene Expression Signature Database (CDS-DB) [41] is an interactive, user-friendly resource, released in September 2023, aiming to provide data on how cancer therapies affect gene expression in patient samples. It compiles gene expression profiles from 78 patient-derived paired pre- and post-treatment datasets, from GEO and ArrayExpress databases, with manually curated clinical information. These source datasets have been organized into 219 CDS-DB datasets, composed of the gene expression levels from paired pre- and post-treatment patient samples, with the same factors (therapeutic regimen, administration dosage, cancer subtype, sampling location, time, and drug response status). From those, datasets with at least two patients have been used to generate differential expression analyses, resulting in 181 dataset-level gene perturbation signatures. In addition, 2012 patient-level gene perturbation signatures have been derived by comparing pre- and post-treatment profiles from individual patients (e.g., baseline vs. 14-day; baseline vs. three months). All transcriptomic data in CDS-DB were uniformly re-processed from raw files (microarray or RNA-seq), the metadata was manually curated, and the terminologies for drugs, cancers, and genes were harmonized. This database is an important resource for MoA elucidation studies by providing well-curated gene expression data for various cancer types and treatments, and by distinguishing between dataset-level signatures (group-level differential expression) and patient-level signatures (individual patient responses) [41]. Nonetheless, CDS-DB only provides data from cancer patient samples, and in LINCS, the cancer cell lines represent nearly all the gene expression profiles, so they are not ideal for addressing the challenges related to transcriptional responses in non-cancer cells [39]. ChemPert [39], emerged as a manually curated resource that maps the relationships between chemical perturbations, their protein targets, and downstream transcriptional signatures in non-cancer cells. It provides a user-friendly interface that includes two sections: the database and a web analysis tool. The database has three main components: (1) Direct signaling protein targets of chemical perturbagens, curated from Drug Repurposing Hub, DrugBank, and STITCH v5.0; (2) Initial gene expression profiles of non-cancer cells before exposure, extracted from GEO, ArrayExpress, and LINCS L1000 (Level 3 data); (3) Transcriptional responses after perturbation, categorized as upregulated or downregulated. The ChemPert database encompasses over 82000 transcriptional signatures from the exposure to 2566 chemical compounds across 167 different non-cancer cell types, lines, and tissues. It includes target data for 57818 chemical compounds, capturing activation, inhibition, or unknown effects. Additionally, ChemPert also offers two built-in analysis tools: (1) Given a perturbagen and the initial gene expression profile, the users can predict how transcription factors will respond; (2)

Given a specific transcriptional response the users can identify the potential perturbagens based on that input data. Bulk transcriptomic databases average gene expression across cells, potentially masking important heterogeneity responses between different cells to perturbations, such as the existence of cell subpopulations that can survive chemotherapy [42]. To capture these variations, single-cell perturbation sequencing methods have emerged. Techniques like Sci-Plex (for chemical perturbations) and Perturb-seq (for genetic perturbations) leverage mass screening technologies in combination with single-cell resolution to provide a more detailed view of cellular responses [43]. Although traditional single-cell RNA sequencing (scRNA-seq) is essential for analyzing such heterogeneity, its high cost per sample remains a barrier. Sci-Plex [42] was introduced to overcome this limitation by combining two techniques: nuclear hashing and combinatorial indexing-based RNA sequencing (sci-RNA-seq) to assess the global transcriptional responses to chemical perturbations at single-cell resolution [44, 45]. Nuclear Hashing labels cell nuclei with unique DNA barcodes before pooling, allowing multiple treatment conditions to be multiplexed in one experiment. Sci-RNA-seq uses successive rounds of combinatorial indexing to uniquely tag transcripts from individual cells, enabling high-throughput scRNA-seq at a much lower cost.

Database	Description	Reference
Orange	Fruit	It is fruit, which is full of nutrients and low in calories. They can promote clear, healthy skin and also lowers the risk for many d
Cauliflower	vegetable	It is the vegetable, which is high in fiber and B-Vitamins. It also provides antioxidants, which help in fighting or protect against

Sci-Plex was applied to three well-characterized human cancer cell lines, A549 (lung adenocarcinoma), K562 (chronic myelogenous leukemia), and MCF7 (mammary adenocarcinoma), exposed to 188 different compounds at four doses [42]. The integration of the two techniques allowed for profiling thousands of single-cell transcriptomes across nearly 5000 independent samples simultaneously in one experiment. Genome-wide Perturb-seq (GWPS) is a large-scale database that maps how genetic changes affect cell behavior by using a single-cell genetic perturbation sequencing (Perturb-seq), a CRISPR interference (CRISPRi) screening technique combined with single-cell RNA sequencing [54] [54]. In this approach, every expressed gene was silenced at the genome-scale to capture detailed transcriptional responses. This massive dataset allowed them to assign functions to previously uncharacterized genes and to identify new regulators involved in key cellular processes such as ribosome biogenesis, transcription, and mitochondrial respiration [54]. Additionally, the rich single-cell data enabled a deep exploration of complex phenomena, including RNA processing, differentiation, and stress-specific regulation, particularly in the context of aneuploidy. Comprehensive analysis of cellular changes in response to perturbations has been made possible by high-dimensional transcriptional profiling of cells. High-resolution single-cell expression data holds great promise for detecting cell-level transcriptional alterations across experimental conditions, because many disturbances only impact a portion of a certain cell type, with most of the cells remaining unaffected. Gene Expression Omnibus (GEO) is a public repository managed by the National Center for Biotechnology Information (NCBI) that archives microarray and next-generation sequencing data from various organisms, cell lines, etc [48]. The vast array of high-throughput experimental data stored in this resource frequently serves as the foundation for more specialized databases, making it difficult to mention other databases without mentioning GEO. Additionally, querying GEO often requires manual sample identification that has been subjected to disturbance and the control samples, since this database does not have an obligation to publish the metadata accurately [33]. When the data is user-submitted and publicly available, sometimes it ends up being unfeasible to use due to the lack of associated metadata. Other databases use the data originally from GEO, after curation and ensuring that all the associated metadata are provided, to overcome the lack of this in-depth information that often exists. The CRowd Extracted Expression of Differential Signatures (CREEDS) is a gene expression signature database that resulted from a crowdsourcing project to improve the annotation and reanalysis of data from the GEO [51]. By engaging over 70 participants, several datasets with single-gene, drug, and disease perturbation signatures were manually curated and validated for accuracy. These signatures were then used as a training set for machine learning models, allowing the collection to be scaled up with an automated search. CREEDS tackles the key challenge of metadata inconsistencies and lack of standardized annotation by using both manual curation and computational methods such as the characteristic direction algorithm to prioritize the DEGs [51]. The platform enables the download of human-validated gene expression signatures, with reduced error in control and perturbation selection, unlike

signatures available from fully automated signature extraction tools. PertOrg 1.0 [56] is another database built on extracted data from GEO [48] and ArrayExpress [47] databases to construct a comprehensive tool to analyze and download curated gene expression and phenotypic data from genetically modified organisms. This extensive database catalogs induced in vivo genetic perturbations across 8 diverse model organisms including mammals (mouse and rat), non-mammalian vertebrates (zebrafish), invertebrates (nematode worm and fruit fly), microorganisms (bacteria and yeast), and plant (thale cress). The database includes various types of genetic modifications, such as gene knockout (complete removal or inactivation of a specific gene), gene knockdown (partial suppression of gene expression using RNAi or similar techniques), gene overexpression (increased expression of a target gene through vector-based methods), and mutations and other genetic modifications (specific point mutations, insertions, or deletions introduced to study gene function and disease models). PertOrg has identified over 8.6 million DEGs associated with genetic modifications, derived from microarray, RNA-seq, and scRNA-seq. The database has two main built-in analytical tools, the differential gene overlapping analysis which investigates perturbation datasets significantly enriched in the user-given gene set using a hypergeometric test, and dataset enrichment analysis which identifies perturbation datasets where the user-given gene set is over-represented [56]. The huge collection of curated non-human data and all the functionalities offered make this a valuable resource, bridging the gap between genetic disorders and their phenotypic outcomes. Perturbation signatures can provide valuable information about the key targets and pathways involved in the compound's effect through the gene expression changes. However, extracting downstream information is not an easy task itself, so integrating transcriptomics data with prior knowledge networks allows the mapping of gene regulatory networks and a better understanding of the complex interactions and regulatory mechanisms behind the snapshot provided by a perturbation signature.

2.3 Prior knowledge Network

Understanding the physical molecular interactions within a biological system is crucial for contextualizing experimental data. Computational methods for elucidating the mechanisms of action allow the integration of omics data with prior knowledge of the interactions between biological entities [28]. These interactions can be represented with more or less complexity and can be included in the analysis as supplementary data sources. A prior knowledge network (PKN) is a collection of interactions where nodes represent molecular entities (such as proteins, genes, or metabolites) and edges illustrate their relationships. Understanding causal graphs is key to modeling and interpreting these networks, as they depict cause-and-effect relationships. In such graphs, nodes represent variables, while directed edges represent causal influences, indicating that a change in one variable affects another [60]. Furthermore, edges can be signed, indicating whether a causal node employs a positive or negative effect on the second variable, and weighted, to show the connection

strength [60]. In causal graphs that model biological networks, multi-edge connections are common, with two or more edges linked to the same node. Networks can be classified based on interaction types and node characteristics. Protein-protein interaction (PPI) networks show direct interactions between proteins. Gene Regulatory Networks (GRN) Figure **??** illustrate how transcription factors influence gene expression [1]. Signal transduction networks describe how cells process external signals. Metabolic networks display relationships between enzymes and metabolites. Furthermore, networks are not always composed of molecular entities, as is the case with the disease network model, which links diseases using genes and mutations as connections. These networks fit experimental data to predictions from causal graphs describing the system. The choice of the PKN should match the data type. For instance, for transcriptomic data, integrating a Protein-Gene Regulatory Network (PGN) may be beneficial, while for metabolomic data, metabolic networks are more suitable. Researchers have made significant efforts to construct regulatory networks. A primordial example was the functional characterization of yeast genes through PPI analysis. This study aimed to show the guilt-by-association principle, by inferring an unknown protein's function by looking at its interactions with nearby entities [60, 61]. The guilt-by-association principle is a foundational concept in biological networks. It suggests that genes with similar functions often interact with the same proteins or have similar expression patterns [62]. This principle also applies to drugs that cause similar transcriptional responses and may have comparable mechanisms of action (MoA) [18]. Biological interactions can be described with different levels of complexity. A network is an intricate representation of the global interactome, linking all entities in the system. These interactions are primarily established in published experimental research and can vary in the amount of associated information. Based on supporting studies, each interaction may include details about direction, signal, and confidence level. This helps filter data, creating a network with more reliable relationships. Still, networks can be noisy and incomplete, with high rates of false positives and negatives and a tendency for well-researched entities to become overrepresented [28, 63, 64]. In contrast, a pathway is a simpler version of a network. It illustrates a series of molecular interactions that begin with one entity and follow a specific signaling cascade. This arrangement helps classify entities by their common biological roles. Yet, pathways often miss crosstalk between other pathways and provide a static view of a dynamic process [28]. The entities' overrepresentation issue also applies to pathways. Another way to show interactions is, for example, through regulons. Regulons are groups of co-regulated genes controlled by a common transcription factor and are usually represented as GRN (Figure 2). The choice of network type must always be appropriate to the scientific question and the type of data with which the network is used. The interactions between biological entities and the complexity of these interactions should be considered. For example, if pathways are used instead of a full network, they might miss some interactions and changes over time. If the study targets a specific cell type or tissue, it's important to use tissue-specific networks. Databases like TissueNet [65] can provide molecular interactions specific to a particular cellular context. The use of large-scale causal

graphs for gene expression data interpretation was first introduced by Pollard et al. [64, 66]. This study aimed to infer the molecular causes of the changes in oxidative phosphorylation gene expression in skeletal muscle from type 2 diabetes (DM2) patients. For this purpose, the gene expression data were integrated with a large-scale model created from over 210,000 molecular relationships based on the DM2 literature. Computer-aided causal reasoning on these complementary data identified that the observed changes are linked to decreased glucose transport, impaired insulin signaling, and increased risk of post-transplant diabetes [66]. Given the good results obtained from supplementing the studies with PKN, the identification of interactions began to receive more attention. The development of high-throughput screening techniques such as yeast two-hybrid screening and DNA microarray [67] allowed the detection of PPI. Those interactions began to be deposited in databases that provide molecular interaction data. Nowadays, there are several public and commercial network and pathway resources. Table 2 summarizes some of the main resources of biological pathways and networks. Two resources, public and commercial, that provide composite networks are OmniPath and MetaBase™, respectively. OmniPath [68] is a freely available resource of prior knowledge in molecular biology. It combines data from over 100 resources and builds five integrated databases with different types of data: Interactions (several molecular interactions organized into sub-networks), Post-Translational Modifications (enzyme-substrate reactions), Complexes (35,000+ protein complexes), Annotations (proteins and complexes annotations, such as the function, localization, tissue, etc.) and Intercell (inter-cellular signaling roles, such as, if a protein is a ligand, a receptor, an extracellular matrix component, etc.) [68]. The interactions database is a composite signaling network that offers several manually curated subnetworks, encompassing a total of 282,504 unique interactions. Each subnetwork has different types of interactions, including post-translational interactions, transcriptional interactions, post-transcriptional interactions, and other interactions involving small molecules. The number of interactions per subnetwork is described in Figure 3. One of the GRNS that is provided by this database is the CollecTRI-derived regulons [1] Figure **??**. This collection contains high-confidence signed transcription factor (TF) - target gene interactions. These interactions were compiled from 12 resources, including information inferred from text mining, manual curations, and several publicly available databases.

MetaBase™ [69] is a proprietary, commercial database from Clarivate that offers one of the most comprehensive, manually curated systems biology datasets available. It contains over 4.2 million molecular interactions, including protein-protein, protein-RNA, compound-protein, compound-compound interactions, and transport reactions, with details on directionality, mechanisms, and effects. In addition, MetaBase provides more than 1,500 pathway maps that cover regulatory, disease, metabolic, and toxicity characteristics, alongside over 10,000 disease-related networks and 1,000+ validated networks. Each interaction is assigned a trust score that reflects its reliability, helping users distinguish well-established interactions from those obtained via high-throughput screening. MetaBase is accessible through SQL queries or via the metabaseR package in R, which

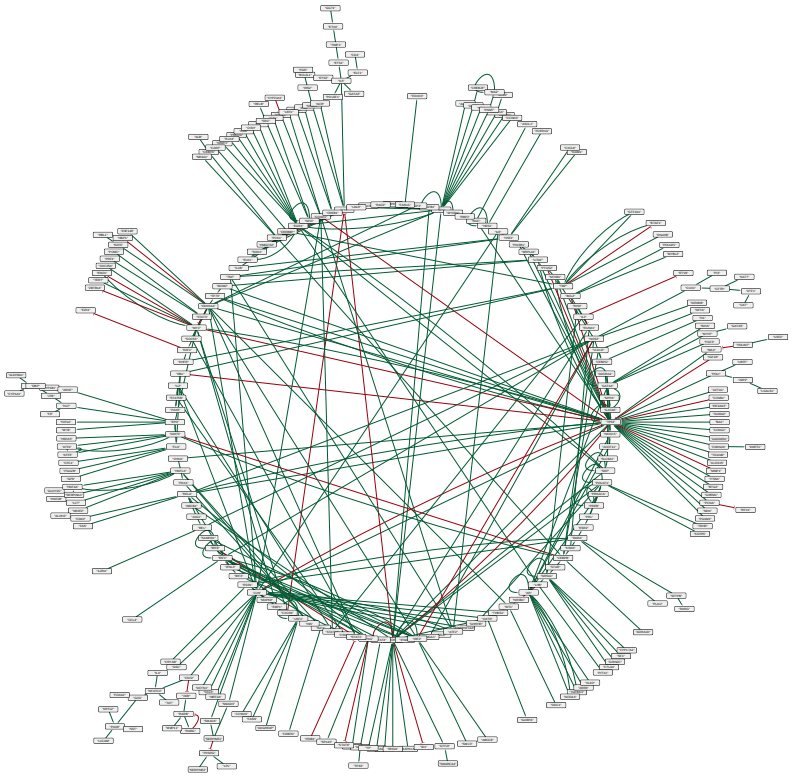

Figure 2.1: Gene regulatory network based on regulon representation of human transcriptional interactions. CollecTRI-derived regulons were extracted from the decoupleR (v. 2.12.0) package, which provided 43,178 interactions. CollecTRI collection [1] is a comprehensive, curated resource of transcription factors (TFs) and their target genes, expanding on DoRothEA. This figure illustrates a subset of those 1,000 interactions, with edge color indicating the mode of regulation (green for activation and red for repression) using RCy3 (v. 2.26.0) R package.

simplifies visualization, functional analysis, and network manipulation. Furthermore, the CBDD R package offers 73 advanced algorithm implementations for analyzing and extracting insights from networks. Integrating biological knowledge with experimental data is key to understanding how cellular regulation impacts gene expression. Known interaction networks are usually used to predict the results of regulatory events, but they can also be used in the opposite direction, to find upstream regulators that cause expression changes [64]. Computational tools play a crucial role here. They combine high-throughput omics data with established cellular interactions, like protein-protein interactions and signaling pathways, to give a broader context. While network data show the complete interactome of molecular interactions, pathway data arrange these interactions into cascades. Each of these data sources forms prior knowledge. When combined with experimental results helps to create mechanistic hypotheses about, for instance, how a perturbation works in a system. This integration of experimental and interaction data sets the stage for some of the computational methods covered in the next chapter. These methods aim to uncover the mechanisms that cause the observed transcriptomic changes.

2.4 Computational methods for MoA inference

Due to technological advances, large-scale transcriptomics datasets can now be generated affordably across many perturbations. However, extracting biological insights from this complex data can be complicated. Thus, using computational tools has become essential for analyzing the massive amounts of data available. These methods include in silico experiments that combine experimental data with prior knowledge. They fall into three main categories: topology-based, similarity, and enrichment methods [14, 84]. Choosing the appropriate tool depends on several factors, including the type of input data, runtime, computational complexity, and the specific scientific questions being addressed, along with the inherent strengths and limitations of each method [28]. For example, in Hill et al.'s study, [60] three types of topology-based algorithms were employed. Node prioritization algorithms rank the nodes in the network based on connectivity or distance from start nodes, causal regulator algorithms infer and rank upstream nodes in the network by their connectivity or distance from start nodes, and subnetwork identification algorithms extract regions of the input network that are enriched for perturbed nodes. These approaches help generate mechanistic hypotheses about the cellular targets and pathways affected by a perturbagen more accurately and easily [85].

2.4.1 Similarity-based methods for comparative analysis

2.4.2 Enrichment-based tools for downstream analysis for downstream analysis

2.4.3 Topology-based methods for upstream analysis

2.5 Benchmarking of computational methods for MoA inference

The use of computational methods to elucidate mechanisms of action is becoming increasingly indispensable for integrating and interpreting the multitude of available data. Given the plethora of existing computational tools, choosing the appropriate data and methods to answer specific scientific questions can be challenging. When a new tool is developed and published, it is usually compared with popular existing methods. For those with less experience, distinguishing the benefits of a novel tool from others that may be equally advantageous but better suited for different applications or data types can be difficult. A comprehensive benchmarking study of the tools is crucial for evaluating available methods in a standardized way, providing sufficient information to accurately choose the best tools and data for a given study [88]. A key component of a benchmarking study is the use of gold-standard datasets, against which the results obtained from a method are compared. By comparing these results with the ground truth, it is possible to evaluate performance metrics and statistical analyses, consistently distinguishing different computational algorithms based on their behavior with certain

types of data. It is widely acknowledged that evaluating the vast available methods is important for obtaining more accurate results. Therefore, following certain good practices when conducting a benchmarking study is essential. These studies can be carried out by the authors who implemented the tool, independent groups, or as organized challenges, such as those organized by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) Consortium [89]. When the authors do the evaluation, the aim is usually to demonstrate the advantages and performance improvements over other techniques. In other contexts, it is extremely important to define the scope and purpose of the benchmark. The selection of the methods should reflect the relevance of the study's objective and include publicly available implementations to ensure accessibility. Parameter optimization can significantly affect a tool's behavior, including runtime, yet finding the optimal values is not always straightforward. Thus, balancing default settings with computational efficiency is important. Regarding datasets, in the context of studying the compound's mechanism of action, it is crucial to include diverse data sources and generation methods to ensure representativeness and a credible assessment of performance. For instance, transcriptomic data, if making sense in the scope, should ideally include both bulk RNA-seq and single-cell RNA-seq data to broaden the options and use two widely used types of data. Since there are no perfect, fully curated datasets, it is necessary to ensure quality to avoid biasing the results and performance of the tools [89]. The same applies to the gold standard datasets, which serve as the ground truth and are fundamental for statistical analysis and assessing metrics, defining the essence of a benchmarking study. Some benchmarking studies arise from an effort to contextualize gene expression data with several computational algorithms. Hosseini-Gerami et al. [27] evaluated the performance of different causal reasoning algorithms to recover direct compound targets of small molecules and associated signaling pathways using gene expression data. The study compared four causal reasoning algorithms against networks from two different sources and transcriptomics data from one database. Hill et al. [60] conducted a study that provided a more comprehensive framework by analyzing a diverse range of algorithms, networks, and datasets to assess how well network-based algorithms prioritize and connect gene lists derived from transcriptomics data. This study integrated 17 algorithms, categorized into three main groups: (1) Node Prioritization Algorithms, which rank network nodes based on connectivity, (2) Causal Regulator Algorithms, which identify upstream regulators of gene expression changes, and (3) Subnetwork Identification Algorithms, which extract subnetworks linking input genes. The algorithms were applied to three PPI networks, each with different structures and levels of curation, using hundreds of datasets from four sources to cover scenarios where certain data types might be unavailable. The first network combined data from various sources, resulting in a mix of signed/unsigned and directed/undirected interactions. The second network included only signed, directed, and high-confidence interactions, while the third was a large-scale, undirected PPI network. This study exemplifies good practices while doing a comparative analysis, by including and integrating different resources, although exploring the parameter landscape for each algorithm was beyond the scope

of this work, so it was not included. Typically, a final benchmarking analysis ranks the algorithms in terms of the most appropriate use for distinct applications, and so the choice of algorithm(s) may depend on the specific use case [60]. By providing a robust assessment of the capabilities of existing algorithms, these studies leverage knowledge and provide guidelines for researchers to choose which resources should be used in certain situations [88].

3

Materials and Methods

The following section describes the workflow of the benchmarking study. It begins by describing the input data used, both transcriptomics data and prior knowledge data. Further emphasis is given to the implementation of the selected algorithms. Finally, the description of the algorithm's execution, as well as the methods used to assess their performance.

3.1 Benchmarking architecture setup

Several tools and algorithms are available for most research tasks in computational biology, and new algorithms and tools are published every week. Systematic benchmarking of tools is a time- and resource-consuming endeavor, while a lack of benchmarking carries several potential risks. Finding the right computational tool for a given research question is essential. Researchers usually carry out published benchmarking to demonstrate that their tool performs better than others. ABC is a consortium created in 2021 that aims to help members reduce RD risks, saving time and resources by distributing the effort of benchmarking computational biology algorithms. ABC is a consortium established in 2021 that aims to assist members in reducing RD risks, saving time and resources by distributing the effort of benchmarking computational biology algorithms. ABC maintains the same workflow regardless of the case study. It consists of three main steps: (1) Voting, (2) Curation, and (3) Coding. The consortium members suggest and vote on the use case (1). Once the use case is determined, the curation phase begins, where Clarivate collects the most appropriate datasets and algorithms according to the voted use case (2). Again, the members vote on the final selection of datasets and algorithms (1). Finally, the last phases - implementation, execution, and reporting - are conducted by Clarivate (3). The project description will fall within the third phase of the workflow, specifically concerning some of the algorithm's implementation and execution, where I actively participated since all the data was already collected and voted on when I started the project. A visual representation of the study workflow is provided **??** and will be explained in detail in the following sections. The entire workflow was implemented in R Programming Language (R) Statistical Software (v4.4.1) [96].

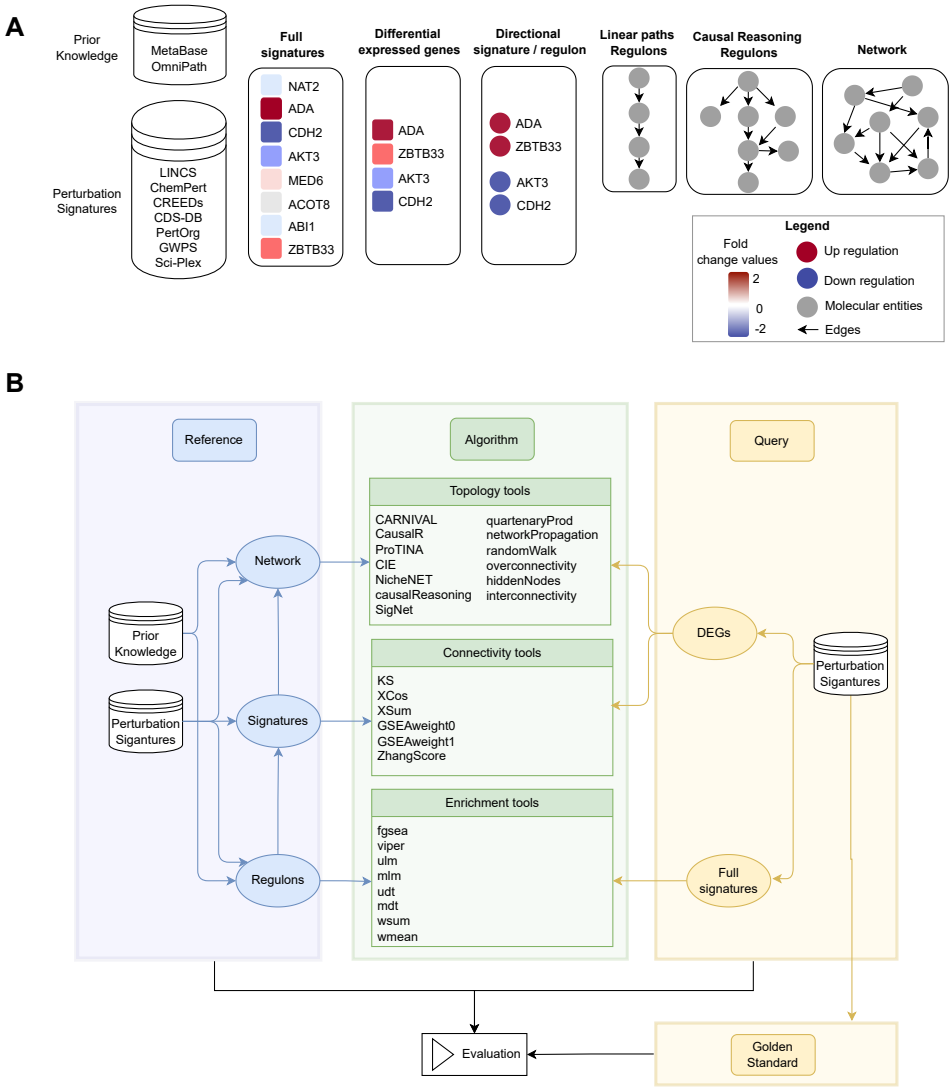

Figure 3.1: Schematic of the study architecture. A. Perturbation signatures collected from seven public sources are used in the benchmarking framework either as reference, query, and gold standard (known targets) datasets. Prior knowledge networks, used as reference, were derived from two sources: OmniPath (public) and MetaBase™ (commercial). From OmniPath, a global network, and regulons were used as references. From MetaBase, it was also used a full network, regulons, and, in addition to the regular regulons, regulons derived from linear paths. B. Three classes of computational methods were evaluated: topology-based, connectivity-based, and enrichment-based, comprising a total of 27 algorithms. Depending on the method, input data may consist of a global interactome (network), curated signaling pathways, or perturbation signatures (typically directional gene sets or full transcriptomic profiles, which can be reduced to gene sets if needed). These input types are often interrelated, and the arrows in the diagram indicate the required data transformations specific to each algorithm. The output of each method is systematically compared to the gold standard targets for evaluation.

3.2 Data Description

As represented in Figure **??**, each algorithm should receive two types of inputs: query and reference dataset. The query dataset refers to the data derived from perturbed signatures (Full profiles or DEGs lists). The reference dataset can be derived either from perturbed signatures (Full profile, DEGs, or directional/regulons) or from prior knowledge (Networks or pathways/gene sets). The databases and datasets used as perturbed signatures and as prior knowledge are described below.

3.2.1 Gene expression data: Perturbation signatures

Currently, there are several publicly available perturbation-driven gene expression datasets. This study comprehends transcriptomics datasets from 10 different public sources, summarized in Table 3. Chemical and genetic perturbagens were included, spanning bulk microarray, bulk RNA seq, and single-cell RNA seq assays. Each dataset contains more than hundreds of perturbation signatures. For each collection, the perturbagen type, the total number of unique perturbagens profiled, and the subset for which a gold standard target annotation is available were recorded. The gold standard is necessary for the evaluation. It is a set of known targets, whether drug-protein interactions or genes deliberately perturbed which were used to assess the ability of the algorithms to recover true upstream regulators from observed expression changes. The LINCS expands upon the original CMap by leveraging the cost-effective L1000 platform, which directly measures 978 "landmark" transcripts and imputes the remaining transcriptome to reconstruct genome-wide expression profiles. LINCS comprises several distinct collections of perturbations in human cell lines: over 30,000 unique small-molecule treatments, CRISPR knockouts targeting 5,156 genes, cDNA overexpression of 3,780 genes, and shRNA knockdowns of 4,854 genes. The level 5 data were retrieved from the CLUE platform (available at: https://clue.io/data/CMap2020LINCS2020). This level already contains the differential expression signatures with z-scores aggregated across biological replicates without p-values. Since each perturbagen usually appears under multiple conditions (different doses, time points, and cell lines), these were condensed into a single consensus signature per perturbagen by extracting every available gene's z-score and then using the median value across signatures. For the gold standard, directional effects were assigned as follows: for chemical perturbations, CDDI (Cortellis Drug Discovery Intelligence) database annotations (i.e. if the drug is annotated as the antagonist of target X, that target would be assigned negative effect in golden standard); for CRISPR and shRNA datasets, the target genes were assigned with inhibition effect, and for OE, each target was assigned with activation effect. ChemPert is a manually curated compendium of 82,256 transcriptional signatures derived from non-cancer cell compound perturbation experiments. Most signatures originate from bulk expression studies in various cell lines, and each is represented as a list of DEGs indicating only up- or down-regulation (no fold-change values or p-values).

From the total number of signatures, only 2,587 have distinct compounds. A set of consensus DEG lists were derived to reduce redundancy and runtime. For each compound, only genes appearing as DEGs in at least two signatures and with the same regulation direction were kept. As well as only signatures with at least 50 consensus DEGs. This resulted in 1,304 signatures which was the dataset used instead of the original ChemPert. CDS-DB contains 78 cancer patient-derived, paired pre- and post-treatment transcriptomic datasets, all with associated metadata such as drug dosages, sampling times, and locations. 181 study-level gene perturbation signatures (85 therapeutic regimens across 39 cancer subtypes) were extracted. The perturbagen consists of drugs, and the expression is measured as microarray or RNA-Seq, and each signature contains full expression with fold change and p-values. The sci-Plex dataset is based on a single-cell transcriptomics method that uses nuclear hashing. Sci-Plex dataset profiled three cancer cell lines treated with 188 small-molecule compounds. The data contains full transcriptomic signatures with around 11,000 genes each, containing dose-response effect estimates and associated p-values. Only signatures linked to compounds with known CDDI targets were kept. For each of the 135 perturbagen with a target, the gene expression responses were measured across 3 cell lines, resulting in 405 signatures. CREEDS is a crowd-sourced, manually curated collection of perturbation signatures from GEO. Includes both small-molecule and genetic perturbations in mouse and human with the expression from different bulk gene expression platforms. These signatures are represented as DEG lists indicating the regulation direction without fold change values. Only perturbations with CDDI target annotations were retained, and all mouse data were mapped to human orthologs using the metabaser package. PertOrg is a curated collection of in vivo genetic perturbation (such as knockdown, knockout, and overexpression) signatures across eight model organisms. Only mouse signatures with more than 5,000 genes were kept and mapped to human orthologs using the metabaser package. For the golden standard, perturbation effects were considered as activation for knockin, overexpression and activation, and inhibition for the remaining ones. Since PertOrg originally contained 7,398 signatures but only 2,321 distinct target genes, a filtering criteria was applied. Each signature should have at least 50 DEGs, and the target gene's fold change should be ranked in the top 5The GWPS dataset represents a large-scale effort for single-cell CRISPRi profiling across more than 2.5 million human cells. It targets 9,866 genes and was generated using the 10x Genomics platform. The dataset includes 1,946 perturbation signatures corresponding to gene knockdowns. Each signature consists of full transcriptomic profiles by z-scores without p-values. Although the DEGs per signature were also provided by the authors, only the full signatures were used in the analysis.

The Table **??** provides a summary od the datasets used in the study.

Table 3.1: Datasets summary.

Dataset	**Perturbagen**	**Perturbation**	**N. of perturbagens**	**Signature**

LINCS compounds	Chemical	Compound	3499	Full
LINCS CRISPR	Genetic	CRISPR KO	*2B, 1C*	Full
LINCS OE	Genetic	?	*A*	Full
LINCS shRNA a375	Genetic	?	—	Full
ChemPert	Chemical	2	*C*	DEGs
CREEDs	Chemical	2	*1A, 1C*	0
CDS-DB	Chemical	1	—	0
PertOrg	Genetic	1	*1A*	0
GWPS	Genetic	2	*C*	0
Sci-Plex	Chemical	3	*2B*	0

The concept of causal inference can be described as the ability of algorithms to find the target candidates of a perturbation, in this study, based on gene expression data. Each dataset described in this section feeds into the benchmarking workflow as the query or reference dataset and as a gold standard (signature associated with the set of known targets). Golden standard target annotations are mandatory, not for running the algorithms, but for the evaluation step. During the evaluation, the performance of each algorithm will be assessed based on how well the target was recovered. When using signatures derived from drug perturbation, it can be hard to identify the exact compound used just from gene expression. Instead, it's easier and more meaningful to infer the target(s) of the compound (e.g., a protein the drug binds to). Although MetaBase also contains compound information, most networks do not, but they do include gene or protein targets. Even for connectivity scoring methods, knowing the drug targets helps when querying compound perturbations versus gene perturbation references (or vice versa). Five chemical perturbation datasets (LINCS compounds, ChemPert, CDS-DB, Sci-Plex, and CREEDs) were subjected to this mapping through three approaches. (1) The authors' target information was extracted from the dataset/database whenever possible. All target gene symbols were extracted. (2) Small molecules were mapped against the drugs in Clarivate's CDDI database. (3) The target lists from the database and from CDDI were then merged to form the final set of targets for each drug or small molecule.

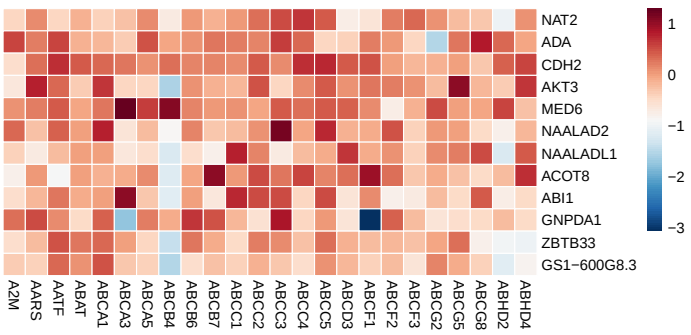

Figure 3.2: Transctipotomic signatures perturbation from LINCS shRNA, cell line a375.

3.2.2 Prior Knowledge: Interaction Networks

One type of input that can serve as a reference is prior knowledge data for contextualizing gene expression signatures. The benchmarking framework depends on three complementary types of this data: PKN (global interaction networks), regulons (regulator-target gene sets), and pathway-derived linear maps (Table 4). Although these resources vary in their coverage, they are related, as illustrated in Figure 4. Including sources of different sizes and densities is particularly important for understanding how topology-based algorithms manage that in terms of performance. Additionally, an increase in network size can introduce noise that may disturb the rationale. The interactions are driven by two databases, OmniPath [68] and MetaBase [69]. OmniPath is a public database with protein-protein, transcriptional, and RNA-related interactions. MetaBase™ is a manually curated systems-biology database, proprietary Clarivate, containing over 4million directional molecular interactions, such as protein-protein, protein-RNA, compound-protein, etc. From each of these two sources, PKN and regulons were used as input. Canonical linear pathways were extracted only from MetaBase. Four concepts that are important to distinguish in this section related to interactions are directionality, effect, mechanism, and weight. Directionality indicates the intended flow of signal, from the source to the target node. The effect (or edge type) denotes whether the interaction is inhibition (-1), unknown (0), or activation (1). Mechanism distinguishes generic molecular interactions (interactions from receptors upstream to the transcription factors downstream, 0) from transcriptional regulation edges (transcription factors with their target genes, 1). Finally, weight determines interaction confidence based, for instance, on literature support. Regardless of the database source used, the mandatory annotation for each interaction is directionality information, other columns may be present but may not be used by algorithms. OmniPath PKN was constructed by integrating signaling and TF-target interactions using OmniPathR (v. 3.14.0) R package. Signaling interactions were obtained using the $import_omnipath_interactions()function, anditwasassignedmechanism = 0, andfortranscriptionalregu$
$1.Thesewerecombinedintoasinglenetwork, andinteractionswitheffect = 0werekeptonlyifmechanism$
$0.NodesinOmniPathareproteinsorproteincomplexes(UniProtIDs)withthecorrespondinggenesymbol($
$0)andtranscriptionregulationinteractions(mechanism = 1).Onlyhigh-confidenceinteractionswithde$
$annotate.nwobj2gene()).Anotherwayofrepresentinginteractionsthatcanbeusedasreferencedataforbo$
$basedandenrichment-basedtoolsaretheregulons.Inthiscase, theywereextractedinthesamewayforboth$
$hypothesisGeneration()function, indicatingasparametershowfartolookdownstream(inthiscase, 4ste$

3.3 Algortithms

To carry out a systematic and robust comparative evaluation of inference algorithms, wrapper functions were developed to standardize the input data and output formatting in different computational approaches. Each algorithm has specific data requirements and processing methods, requiring adaptation to ensure compatibility in the common

framework. A wrapper is nothing more than a function that serves as an intermediary layer. These are important to handle data type conversions, parameter standardization, and result formatting, allowing diverse algorithms to be executed consistently regardless of their underlying implementation differences. This approach addresses the inherent complexity of having algorithms coming from different approaches. Here there are two types of wrappers: shared and individual. The shared wrapper architecture incorporates an already established package that bundles several algorithms inside, unlike the individual ones that incorporate single algorithms. The connectivity mapping approaches from the RCSM package, enrichment methods from decoupleR, and topology-based algorithms from CBDD were implemented in shared wrappers. On the other hand, causal reasoning CARNIVAL, CausalR, ProTINA, CIE, and NicheNet were incorporated in individual wrappers. Table 5 provides a complete list of algorithms together with their annotations. Some supporting helper functions were also implemented to facilitate essential data conversions across all wrappers. Those functions include mapping identifiers between transcriptomic datasets and network nodes to ensure the same IDs and converting the input data when necessary. For the query input data, the tool may need a full signature or DEGs. When DEGs are required, the full signature can be filtered using a fold change and p-value threshold or by simply taking the top threshold for differentially expressed genes by fold change magnitude. Regarding reference, the workflow can start with PKN or full signatures, and for topology, enrichment, and CMap tools, require networks, regulons/gene sets, and full signatures, respectively. To use this large variety of input data and tools, some conversions are required to run them uniformly. All the conversions are indicated by the arrows in Figure 4 B. The parameters selected for each algorithm can be found in Supplementary table 2. As the input data, the output should also have the same format, so it is possible to evaluate the performance of each algorithm. For that reason, at the end of each run, all algorithm wrappers return a table with all prioritized regulators identified without any significance filtering applied. The output contains a score column, and the larger score reflects greater confidence in this regulator being causal for observed differential expression patterns. Score may be signed if the tools can predict directionality of perturbation. In that case, regulators are ranked by absolute value of score, and activation/repression status is stored in separate column effect (-1/1 values).

The Table **??** provides a summary od the algorithms used in the study.

Table 3.2: Algorithm summary.

Tool	Algorithm	Description	Resource	Dataset	Reference
Causal Reasoning					
CARNIVAL	Description	Network	DEG/Full	[0]	
CausalR	Description			[0]	
ProTINA	Description			[0]	
CIE	Description			[0]	

NicheNet	Description			[0]
causalReasoning	Description			[0]
Signatures	Description			[0]
quaternaryProd	Description			[0]
Causal Reasoning (baseline)				
randomWalk	Description	Network	DEG/Full	[0]
networkPropagation	Description			[0]
overconnectivity	Description			[0]
hiddenNodes	Description			[0]
interconnectivity	Description			[0]
overconnectivity	Description			[0]
CMap				
KS	Description	Signatures	DEG/Full	[0]
XCos	Description			[0]
XSum	Description			[0]
ZhangScore	Description			[0]
GSEAweight0	Description			[0]
GSEAweight1	Description			[0]
Enrichment				
wmean	Description	Regulons	Signatures	[0]
fgsea	Description			[0]
viper	Description			[0]
ulm	Description			[0]
mlm	Description			[0]
udt	Description			[0]
mdt	Description			[0]
wsum	Description			[0]

3.3.1 Connectivity Mapping

The **??** represents the wrapper function framework for running connectivity mapping algorithms from the RCSM package [87]. This package provides uniform implementations of several CMap scoring methods including Kolmogorov-Smirnov (KS), and GSEA-based approaches. The function is designed to accept filtered DEG lists as query input and full perturbation signatures as reference data. If full signatures are used as the query, they are converted to DEGs using the filtering parameters (Supplementary table 2), as well as the additional parameters. RCSM R package includes a variety of algorithms already implemented, each designed to quantify the similarity or dissimilarity between

query and reference perturbation signatures. Those algorithms include the Kolmogorov-Smirnov (KS) statistic which was used in the original Connectivity Map [9]; Xcos, a cosine similarity metric between query and reference fold-changes; Xsum connectivity map statistic based on the sum of reference fold-change values of query genes; GSEAweight0 is a GSEA weighted KS ES with parameter p = 0, meaning that the fold-change magnitude is not taken into account; GSEAweight1 with parameter p = 1, fold-change magnitude contributes linearly; Zhang, a connectivity mapping score first suggested in [105]. The function handles the different algorithm requirements by preparing either separate up- and down-regulated gene lists for most methods or simple gene vectors for XCos. The function also includes optional regulator filtering for TF mode. The output is formatted to return regulator rankings with similarity scores, directional effects, and optional statistical significance measures. The results are sorted by absolute score magnitude to prioritize the most relevant regulatory relationships regardless of similarity direction. For these algorithms, the regulator score measures the similarity of the query versus the reference perturbation signature.

3.3.2 Pathway Enrichment

For running the enrichment-based algorithms, the decoupleR package [10] was used. It contains 12 algorithms already implemented to extract biological activities from omics data using prior knowledge resources (gene sets or regulons). Some of them take directionality into account (i.e., can work with regulon-gene set with activated and repressed genes). The package was initially used to benchmark approaches for TF activity inference. From those, only GSEA and the others that respect directionality were used. As for connectivity-mapping algorithms, a shared wrapper function (Figure **??**) was built to prepare the input and output data for these algorithms. It is designed to accept full signatures as query and regulon table or a gene regulatory network as reference data. If the reference is a list of signatures or DEGs, it is converted to directed regulatory networks using the common filtering parameters described above. The implementation supports TF-mode by filtering the network to keep only transcription-regulation edges. The query signatures are converted to an fold change matrix and perform ID space conversions when necessary to match network node identifiers.

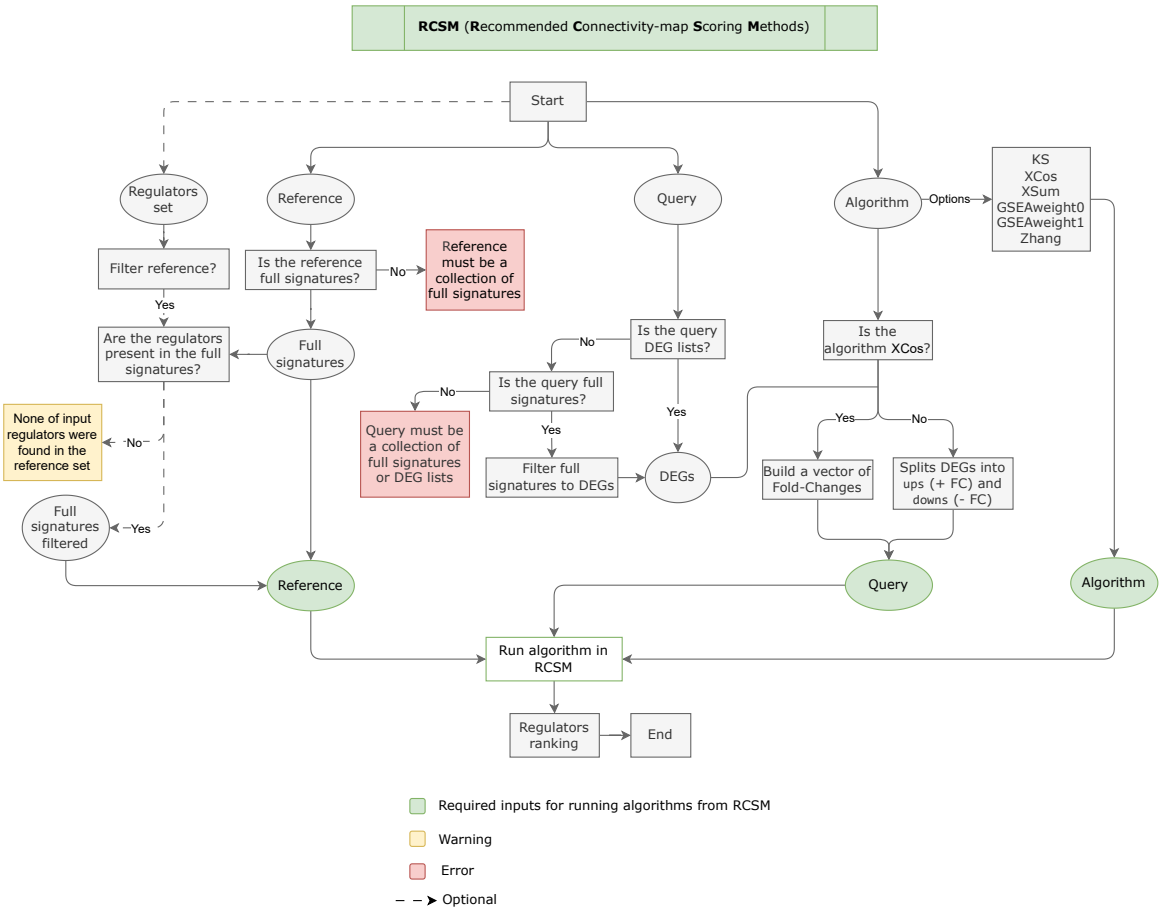

Figure 3.3: Flowchart representing the main steps for implementing connectivity mapping algorithms pre-built in the RCSM R package. The general computational pipeline for executing connectivity-based methods, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicate required inputs, while red highlight potential errors.

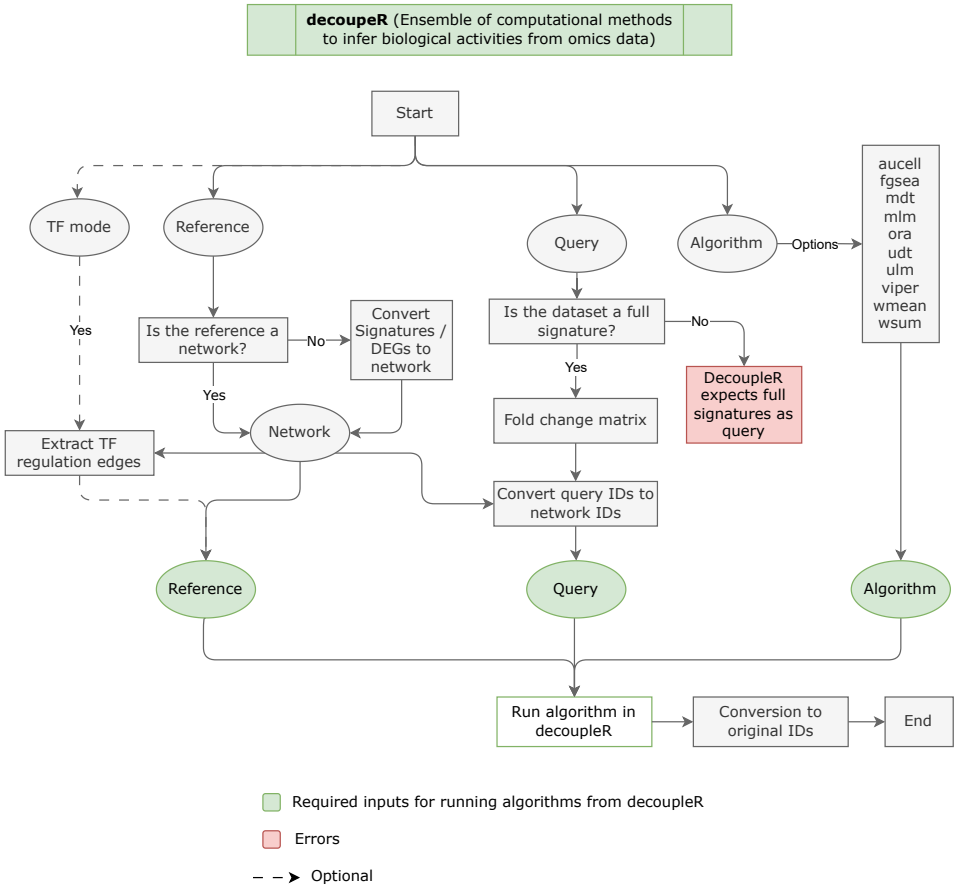

Figure 3.4: Flowchart representing the main steps for implementing enrichment algorithms pre-built in the decoupleR package. The general computational pipeline for executing enrichment-based methods, showing the main input requirements, data preprocessing. Green indicates required inputs, while red highlights potential errors.

4

Adding Support to a New School (work in progress)

The directory `uminho` contains the customization for all Schools of Universidade do Minho. This university is an example of the case where the regulations are defined at University level and all the schools apply the same thesis layout and organization. So, the all the customization is done in the file `uminho/uminho-defaults.ldf`, except the definition of the name and logo of each individual school.

This is the first occurrence of an abbreviation: Computational Biology for Drug Discovery (CBDD).

MetaBase

References

[0] M. J. Alvarez et al. "Functional characterization of somatic mutations in cancer using network-based inference of protein activity". In: *Nature genetics* 48.8 (2016-08), pp. 838–847. ISSN: 1061-4036. DOI: 10.1038/ng.3593. URL: https://europepmc.org/articles/PMC5040167.

[0] G. Bradley and S. J. Barrett. "CausalR: extracting mechanistic sense from genome scale data". In: *Bioinformatics* 33.22 (2017-06), pp. 3670–3672. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx425. URL: https://doi.org/10.1093/bioinformatics/btx425.

[0] R. Browaeys, W. Saelens, and Y. Saeys. "NicheNet: Modeling intercellular communication by linking ligands to target genes". In: *Nature Methods* (2019). URL: https://www.nature.com/articles/s41592-019-0667-5.

[0] *CBDD*. Web Page. URL: https://clarivate.com/life-sciences-healthcare/consulting-services/research-and-development-consulting/cbdd/.

[0] S. Farahmand et al. "Causal Inference Engine: a platform for directional gene set enrichment analysis and inference of active transcriptional regulators". In: *Nucleic Acids Research* 47.22 (2019-11), pp. 11563–11573. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1046. URL: https://doi.org/10.1093/nar/gkz1046.

[0] G. Korotkevich et al. "Fast gene set enrichment analysis". In: *bioRxiv* (2021). DOI: 10.1101/060012. URL: https://www.biorxiv.org/content/early/2021/02/01/060012.

[0] J. Lamb et al. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease". In: *Science* 313.5795 (2006), pp. 1929–1935. DOI: 10.1126/science.1132939. URL: https://www.science.org/doi/abs/10.1126/science.1132939.

[0] A. Liu et al. "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL". In: *npj Systems Biology and Applications* 5 (2019-12). DOI: 10.1038/s41540-019-0118-z.

[0] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf.

[0] *MetaBase*. Web Page. URL: https://clarivate.com/life-sciences-healthcare/research-development/discovery-development/early-research-intelligence-solutions/.

[0] H. Noh, J. E. Shoemaker, and R. Gunawan. "Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection". In: *Nucleic Acids Research* 46.6 (2018-01), e34–e34. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1314. URL: https://doi.org/10.1093/nar/gkx1314.

A

NOVAthesis covers showcase

This Appendix shows examples of covers for some of the supported Schools. When the Schools have very similar covers (e.g., all the schools from Universidade do Minho), just one cover is shown. If the covers for MSc dissertations and PhD thesis are considerable different (e.g., for FCT-NOVA and UMinho), then both are shown.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

A.1 A section here

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna

fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

B

Appendix 2 Lorem Ipsum

This is a test with citing something [0] in the appendix.

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I

Annex 1 Lorem Ipsum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

UNIVERSIDADE NOVA
DE LISBOA

2025

BENCHMARKING CAUSAL REASONING ALGORITHMS FOR ENHANCED DRUG DISCOVERY

Maria Inês Nunes Vilar Gomes
MASTER IN **Computational Biology and Bioinformatics**
SPECIALIZATION Multi-Omics for Life and Health Sciences

NOVA
UNIVERSIDADE DE LISBOA