
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

Draft: *September 9, 2025*

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.

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):

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.

```
\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):

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

Contents

List of Figures

Glossary

Gene signature A gene signature is a specific set of genes whose expression pattern is characteristic of a particular biological state, disease outcome, or response to treatment. Usually, it includes the name/ID of the gene, together with a value representing their relative expression (fold-changes or p-values). *14)*

Mechanism of Action (MoA) Molecular cascade by which a perturbation (such as a drug or genetic modification) produces its biological effect. The molecular cascade includes the interaction with the direct molecular target(s) and the immediate downstream cascade of events leading to a certain cellular outcome. The cascade can be reflected in changes in the gene expression. *1)*

Molecular network A graph representation of molecular interactions where nodes represent molecular entities (such as genes and proteins) and edges represent the relationships between them. Networks can be directed (causality), weighted (interaction strength), or signed (activation/inhibition), or without any of these attributes. From a full network, it is possible to extract subsets of interactions (pathways). Pathways are cascades of molecular interactions, and some databases like Kyoto Encyclopedia of Genes and Genomes (KEGG) or Reactome catalog those into specific types, such as metabolic, signaling, or regulatory pathways. Pathway enrichment analysis provides a high-level understanding of the biological processes by identifying coordinated network variations, instead of focusing on individual genes. *59)*

Transcriptomics

The measurement of gene expression levels across the genome (or a subset of genes) obtained from a biological sample. One or more messenger RNA (mRNA) molecules are produced from the transcription of each gene. Transcriptomic technologies use different approaches to measure gene expression by quantifying individual mRNA molecules. Examples of some transcriptomic technologies are microarrays, L1000, and RNA Sequencing (RNA-Seq). Each one of these techniques can generate full gene expression profiles by capturing the expression of all the genes transcribed in a given sample. Transcriptomic analyses are usually performed by sampling at different time points, conditions, or treatments. The raw data can be further processed to identify differentially expressed genes, using metrics such as log fold-change, z-scores, p-values, and q-values. *7)*

Acronyms

ABC Algorithm Benchmarking Consortium *3, 4)*
AUPR Area Under the Precision-Recall Curve *48)*

CARNIVAL CAusal Reasoning for Network identification using Integer VALue programming *2, 41, 42)*
CBDD Computational Biology for Drug Discovery *2, 37, 38, 48)*
CDDI Cortellis Drug Discovery Intelligence *36)*
CDS-DB Cancer Drug-induced Gene Expression Signature Database *9, 10)*
CIE Causal Inference Engine *2, 47)*
CMap Connectivity Mapping *2, 8, 26, 27, 39, 41)*
CRC Colorectal Cancer *6)*
CREEDS Crowd Extracted Expression of Differential Signatures *17)*
CS Concordance Score *59)*

DD Drug Discovery *1, 5, 6, 8, 28)*
DEGs Differentially Expressed Genes *2, 18, 28, 39, 41)*
DM2 Diabetes Mellitus Type 2 *20)*
DNA Deoxyribonucleic Acid *8)*
DR Drug Repositioning *5)*
DREAM Dialogue on Reverse Engineering Assessment and Methods *30)*

ES enrichment score *29)*

FDA Food and Drug Administration *6)*

GEO Gene Expression Omnibus *9, 11–13, 15, 17)*
GRN Gene Regulatory Networks *19)*
GSEA Gene Set Enrichment Analysis *26, 39, 41)*
GWPS Genome-Wide Perturb-Seq *17)*

HTS High-Throughput Screening *1)*

KS Kolmogorov-Smirnov *26*, *39*, *41)*

LINCS Library of Integrated Network-Based Cellular Signatures *8*, *10*, *12)*

MoA Mechanism of Action *1*, *6–9*, *18*, *19*, *26*, *27)*

NCBI National Center for Biotechnology Information *17)*

OE Overexpression *9)*
ORA Over-Representation Analysis *27)*

PKN Prior Knowledge Network *18*, *19*, *37*, *41)*
PPI Protein-Protein Interaction *19*, *22*, *25*, *41)*
ProTINA Protein Target Inference by Network Analysis *43)*

QS Quaternary Scoring *29)*

R R Programming Language *2*, *4*, *32*, *37*, *39*, *48)*
RCSM Recommended Connectivity-map Scoring Methods *3*, *39)*
RNA Ribonucleic Acid *18*, *37)*

shRNA short hairpin RNA *9)*

TF Transcription Factor *2*, *8*, *19–22*, *37*, *41*, *42)*
TS ternary scoring *29)*

XSum eXtreme Sum *26)*

1

Introduction

This section summarizes the study's underlying motivation, rationale, and goals, emphasizing its significance in the field. It provides context by giving some background on the supporting company and the initiative, along with other contributions. Furthermore, it outlines a reading guide for this thesis.

1.1 Significance and Objectives

Drug Discovery (DD) and development is a time-consuming, resource-intensive, multidisciplinary effort that can be challenging from many points of view. Over half of clinical trial failures are attributed to lack of efficacy, 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 to clinical efficacy [**RN3**, **RN1**, **RN2**]. As thorough understanding of the Mechanism of Action (MoA) is such a critical step, computational methods can accelerate the identification of pharmacologically active agents by providing more efficient and cost-effective alternatives to traditional approaches (for example phenotyping screening). Computation methods can accurately identify a new target or propose new indications for a known molecule, thereby reducing the time dedicated to in cell and in vitro target validation [**RN29**].

A key to understanding a compound's MoA lies in transcriptomic profiling, which captures the changes in gene expression triggered by a perturbagen. While traditional RNA sequencing methods remain too costly for large-scale expression signatures, recent High-Throughput Screening (HTS) advances, such as the L1000 assay [**RN30**], enable the cost-effective generation and analysis of large-scale omics datasets. Several existing databases provide public access to transcriptomic data from experiments exposing cell lines to range of chemical and genetic perturbagens. These datasets can be leveraged, using various computational tools, to establish the causal chain of gene expression changes triggered by a specific compound. Three primary approaches have emerged: causal reasoning, connectivity mapping, and enrichment tools [**RN31**].

Causal reasoning is a topology-based method, that determines potential causes for an

observed gene expression profile, starting from a perturbation signature and a biological interaction network. The network is defined as a signed and directed graph describing relations between nodes (e.g., proteins or genes). Efforts to compile causal molecular relation networks have increased, resulting in several publicly accessible databases (such as OmniPath), which offers curated prior knowledge networks. Among the networks commercially available [**RN32**] we can find MetaBase^{TM}, developed and curated by Clarivate.

The Connectivity Mapping (CMap) method is instead focused on collecting and analyzing perturbation signatures. In this case, a similarity score is used to compare a set of known MoA/compound reference signatures, with a query gene expression profile from a perturbagen of interest [**RN34**, **RN31**]. The principle behind CMap is that the higher the similarity between the query and the reference signature, the more likely it is that the underlying mechanism is the same.

On the other hand, enrichment tools take perturbation signatures as query input and utilize prior knowledge, such as a 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 Factor (TF)) are significantly enriched in the perturbed data. Several algorithms have been developed based on this approach, each producing specific enrichment scores [**RN35**].

With a comprehensive systematic benchmarking approach, this study guides on selecting the most suitable tool for addressing diverse research questions and evaluating a plethora of current solutions for causal regulation assessment. The evaluation process relies on three inputs:

- Gene expression signatures derived from chemical or genetic perturbation experiments.
- A prior knowledge network of the molecular interactions within the system.
- A golden standard dataset to serve as a reference for validation.

These data are used to feed the evaluated methods, serving as the reference, query, and golden standard datasets, respectively. This study analyzes the three types of tools depicted below:

Topology-based tools Eight algorithms for causal reasoning (CAusal Reasoning for Network identification using Integer VALue programming (CARNIVAL), CausalR, ProTINA, Causal Inference Engine (CIE), NicheNet, plus causalReasoning, SigNet, quaternaryProd from the Computational Biology for Drug Discovery (CBDD) R Programming Language (R) package [**RN36**] , and 5 algorithms for node prioritization (networkPropagation, randomWalk, Overconnectivity, hiddenNodes, interconnectivity - from CBDD) were also included as baseline topology-based tools for node prioritization.

Similarity-based tools Six algorithms are built into the Recommended Connectivity-map Scoring Methods (RCSM) R package.

Enrichment-based tools Eight algorithms implemented under the decoupleR package [**RN35**].

Performance is assessed in terms of results obtained, comparing against golden standard datasets, along with the robustness, and the computational efficiency (runtime and memory footprint). With this approach, the project aims at identifying tools that are the most suitable to contextualize gene expression data and proving a correct assessment of biological results.

1.2 Scope

This project was conducted within the framework of the Algorithm Benchmarking Consortium (ABC), a subscription-based initiative for pharmaceutical companies led by Clarivate. 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 ABC's tenth use case - Causal Regulation - which focuses on benchmarking tools designed to identify key regulators from transcriptomic data and prior knowledge networks.

1.3 Other Contributions

This study expands the state of the art in causal reasoning using gene expression data and causal graphs, by presenting a robust framework and a systematic algorithm benchmarking approach. The study was presented during the following poster communication:

XIV Edition of Bioinformatics Open Days . Gomes, A. Ishkin, F. Ciceri, C. Klein. Benchmarking Causal Reasoning Algorithms for enhanced Drug Discovery: Insights from Clarivate's pre-competitive Algorithm Benchmarking Consortium. XIV Edition of Bioinformatics Open Days, 26 March 2025, Braga, Portugal.

Beyond the topic of this thesis, during the MSc industry placement, I was also involved as developer in other activities aimed at identifying reliable solutions for several different external stakeholders in the pharmaceutical business. Although not related with the topic of this thesis, these experiences enriched the consultant experience, allowing for an expansion in the expertise across various domains of computational biology and data science. These projects included:

Skin Microbiome Atlas This project involved the curation and re-analysis of publicly available skin microbiome datasets. My role in this project included conducting an in-depth scientific literature review, compiling relevant datasets, and pre-processing raw sequencing data to ensure consistency and comparability across studies.

Natural Language Processing An NLP pipeline was set up for automatic text classification of epidemiology abstracts, leveraging different versions of the BERT foundational model. Model fine-tuning on a minimal dataset was attempted by generating synthetic data using paraphrasing techniques.

ABC - Spatial Niche Use Case As part of an internal case study for ABC, I implemented several Python wrappers for selected algorithms relevant to spatial niche analysis. This work mainly included the development of Python wrapper functions to run those algorithms, and the integration with an R-based pipeline and the management of Conda environments.

Google Data Extraction Tool In collaboration with another team in the company, I developed an automated script for the retrieval of pharmacological information from web sources. The pipeline involved querying URLs and keywords and extracting structured data through API calls, including a language model API.

Transcriptomic comparative analysis I conducted a comparative analysis of transcriptomic profiles from three types of cancer. The workflow included exploratory data analysis, identification of differentially expressed genes and hub genes, pathway enrichment analysis, and causal reasoning to infer upstream regulators. Additionally, a survival analysis was performed. All the analysis was also repeated to compare both the Human Papillomavirus status and tumor localization.

Proteomics analysis In a proteomics-focused project, I was responsible for carrying out exploratory data analysis and functional enrichment analysis to uncover biological insights from protein-level data.

1.4 Structure

This study is organized in five chapters.

2

Literature Review

This section provides an overview of the relevant research related to the topic of the project. First, we start by highlighting the importance of understanding the mechanism of action in drug discovery, followed by a description of the two key components for MoA elucidation (transcriptomic data and biological networks). A comprehensive summary of the computational methods used to apply various scoring algorithms (topology-, similarity-, and enrichment-based algorithms) is also presented, to provide all the basis for a systematic evaluation of tools for elucidating compound MoA. Finally, an overview of the best practices to perform a benchmarking study is also provided.

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

Developing new drugs is an extraordinarily complex process. The high prevalence of complex and polygenic diseases, which collectively account for 70% of all the deaths in Europe and affect around 25% of the population, is one of the challenges faced by this industry [**RN43**]. In addition, statistics show that *de novo* drug discovery has become an extensive and costly process, taking on average 13 years and $2 billion to develop a new drug, with most of clinical trials lasting 95 months and non-clinical development 31 months [**RN55**, **RN56**, **RN47**]. 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 [**RN44**]. Efforts to address these challenges are evident in the growing number of studies, both in industry and academia.

Drug Repositioning (DR) 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, and it continues to gain increasing attention, as nearly 30% of the drugs approved by the FDA are identified using this approach [**RN54**, **RN64**]. The fundamental goal of DR is to broaden the indication of known, safe, and previously approved drugs. From

multiple points of view, this is a particularly interesting approach. It allows to investigate therapeutic agents that have been put on hold because of failed clinical trials [**RN62**], and also it enables to identify treatment for conditions with unmeet clinical needs. This is the case of rare diseases, which are not providing sufficient returns to pharmaceutical companies to justify a conventional DD pipeline. Many studies have demonstrated the success of establishing new drug-disease relationships [**RN48**]. 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 [**RN54**, **RN64**, **RN94**]. Another classic example is Thalidomide, originally used for sedation and morning sickness, and afterward repurposed for multiple myeloma, leprosy [**RN54**, **RN64**], and to minimize the hippocampal neuronal loss [**RN50**]. Moreover, the low success rate (5%) for phase I clinical studies of cancer treatments led to increased attention in DR for oncology, resulting in several promising findings [**RN64**, **RN63**]. Noteworthy cases include the schizophrenia drug Spiperone, which has been studied for its ability to induce apoptosis in Colorectal Cancer (CRC) cells [**RN51**], and Raloxifene, indicated for osteoporosis, which proved to be effective in reducing breast cancer risk in postmenopausal women [**RN64**, **RN61**].

Understanding how cellular signaling (Figure **??**) is modulated upon a stimulus is essential for identifying potential drug targets and finding new indications for an existing drug. When a drug enters a biological system, it typically interacts directly or indirectly with cellular targets, regulating the activity of signaling networks and pathways. This is commonly referred as the MoA [**RN52**, **RN53**]. These interactions are relevant across the whole DD process, from initial investigation to clinical trials.

A deep understanding of a drug MoA allows to uncover drug-exposure 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 sufficient safety and efficacy [**RN38**, **RN68**]. Yet, not knowing the mechanisms of the compounds can be extremely disadvantageous. This is demonstrated by the case of Dimebon. Originally developed as an antihistamine drug, it later entered clinical trials (with the MoA still unknown) for treatment for Alzheimer's disease, failing to show meaningful clinical efficacy in phase 3 studies. Later, it was clarified that it was the activity on the histamine and serotonin receptors that caused the initial observed cognitive efficacy, instead of the stabilization of mitochondria (as first hypothesized) [**RN38**, **RN69**].

Although we refer to the target(s) of a compound as "direct", this is often not the case. From a chronological point of view, there are a series of interactions that result in modulation of biological processes, 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 the chain of biochemical reactions forming part of the cell signaling cascade. This process is characterized by the signaling pathways leading to a certain cellular response. These pathways can also interact with each other through crosstalk [**RN94**], forming a complex network of interconnected and distinct

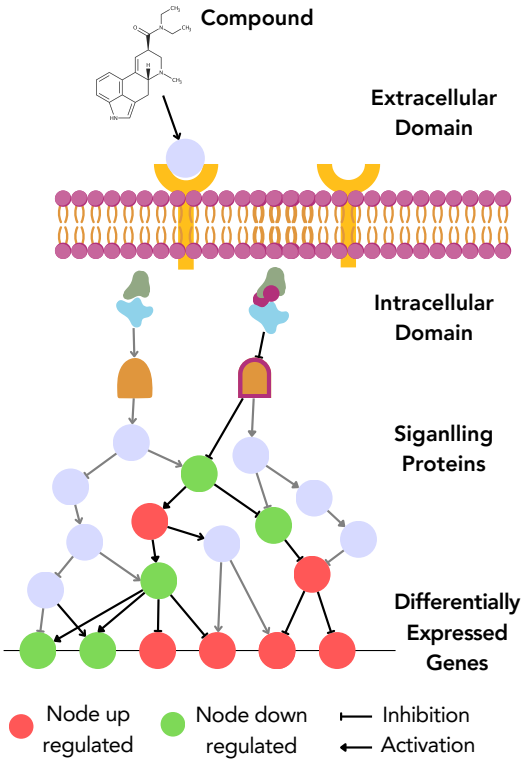

Figure 2.1: Schematic representation of a compound-induced cellular signaling cascade, where binding of the perturbing compound to its extracellular receptor domain triggers downstream intracellular signal transduction events through various signaling proteins, ultimately altering gene expression levels in the nucleus. The red nodes represent upregulated genes, the green nodes represent downregulated genes, arrows denote activation events, and T-bar edges indicate inhibitory interactions.

nodes. The impact of a certain compound in the complex cell signaling cascade can be defined and observed on a system level by high throughput approaches such as genomics, Transcriptomics, proteomics, metabolomics, and even phenomics. Each of them provides a different perspective on the compound's activity. Despite these technological progresses, experimental identification of targets, signaling proteins, and biological pathways (MoA) modulated by an uncharacterized compound can be extremely difficult. To facilitate this process high throughput *in silico* methods have become an attractive option, given the affordability, speed and the amount of data provided. These methods act as a screening process, helping to identify and produce mechanistic hypotheses for further experimental validation [**RN38**]. 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 for each situation is not always simple, requiring precise identification of the scientific question that needs to be addressed. By doing this, researchers can choose the most appropriate data type and bioinformatics tools to efficiently study a compound's MoA.

2.2 Transcriptomic data

Transcriptomic data provides a comprehensive view of gene expression changes in response to a compound. Following the compound's perturbation, this data will reflect the differential mRNA expression, by capturing modulated signaling and TF activity changes triggered by the perturbagen. For this reason, these data are crucial for understanding a compound's MoA. After a cellular disturbance, changes in gene expression levels give rise to a perturbated gene expression signature [**RN114**]. In 2000, the first reference database was built from a compilation of *Saccharomyces cerevisiae* gene expression signatures derived from pharmacological and genetic perturbations. During this effort, the same authors hypothesized that, to generate a wider collection of reference signatures, less expensive gene expression experiments would be required [**RN115**, **RN116**, **RN117**]. Since then, driven by growing interest in drug discovery optimization, several databases have emerged to aggregate and publicly provide perturbagen transcriptomic signatures. These databases allow to extract information of gene expression changes in response to certain manipulations and treatments (Table **??**).

DrugMatrix was the first larger molecular toxicology database, created in 2006 by Iconix Pharmaceuticals (later acquired by the National Institute of Environmental Health Sciences) and publicly available since 2011 [**RN115**, **RN118**, **RN119**]. The database comprises the gene expression responses, detected via microarray analyses, 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 specific types of toxicity [**RN121**]. However, studies *in vivo* limit the number of disturbances that can be studied, due to excessive costs that make it impractical to generate data on a large scale, as well as all the associated ethical implications [**RN34**]. In addition, transcriptional changes are usually cell-specific, so a precise analysis of transcriptomic changes should be carried out considering the cell type, further increasing the complexity of this effort [**RN86**].

CMap is a resource that systematically establishes connections between the MoA of chemical compounds, diseases, and biological processes through pattern matching between signatures derived from cell lines exposed to different perturbagens. The first version of CMap (CMap 1.0) generated 453 signatures derived from 164 distinct small molecules applied at two-time points (over a range of concentrations), to four human cell lines (MCF7, PC3, HL60, and SKMEL5) [**RN34**]. Although the goal could be achieved using signatures from various -omic layers, this resource focused only on mRNA expression data through Deoxyribonucleic Acid (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 applications in the DD area. The experimental conditions tested were few and undiversified in terms of perturbagens and biological systems. The recent advances in high-throughput technology allowed large data acquisition, as in 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 (representatives of various biological processes), in cells exposed to a different set of stimuli [**RN30**]. Using the Luminex L1000 platform, expression of these landmark genes is directly measured, and computational methods and *in silico* imputation then infer the expression of 11350 additional genes, to provide wider reconstruction of the genomic profiles [**RN30**]. In a preliminary phase, LINCS released over 6000 signatures from around 1300 small molecules, many of which were FDA-approved [**RN115**]. As of today, LINCS 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 pharmacologic (small molecules) and genetic modulators, such as gene knockdowns using short hairpin RNA (shRNA) and/or CRISPR and induced Overexpression (OE). LINCS L1000 provides data on 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 [**RN30**]. The data provided by LINCS L1000 has increased the understanding of the association between changes in gene expression and certain disorders, facilitating drug repositioning and contributing to the generation of testable hypotheses about the MoA of less characterized compounds. However, the presence of gene expression changes imputed and not measured directly can lead to some inaccuracies. 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 with protein, expression due to post-translational modifications [**RN38**]. Despite these limitations, several computational methods have been developed to apply these data to facilitate the drug discovery process. These methods will be described later in this chapter.

The Cancer Drug-induced Gene Expression Signature Database (CDS-DB) [**RN84**] is an interactive, user-friendly resource, released in September 2023, aiming to provide data on gene expression in patient samples, following exposure to anti-cancer therapies. It compiles gene expression profiles from 78 patient-derived paired pre- and post-treatment datasets, (from Gene Expression Omnibus (GEO) and ArrayExpress databases), with manually curated clinical information. These sources have been organized into 219 CDS-DB datasets, composed of paired pre- and post-treatment gene expression profiles from multiple human samples. Pairing has been performed considering a wide range of factors (therapeutic regimen, administration dosage, cancer subtype, sampling location, time, and drug response status). From those, datasets containing at least two patients have been used to generate differential expression analyses, resulting in 181 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 a valuable resource for MoA elucidation studies as it provides well-curated gene expression data for various cancer types and treatments, including pre- and post-treatment samples obtained from both grouped and individual patients [**RN84**]. Nonetheless, CDS-DB and LINCS are databased focused exclusively on cancer cells (respectively patient-derived and cell lines), so they are not ideal for addressing the challenges related to transcriptional responses in non-transformed cells [**RN86**].

ChemPert [**RN86**], emerged as a manually curated resource that maps the relationships between chemical perturbations, their protein targets, and downstream transcriptional signatures in non-transformed 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 untreated non-transformed cells, extracted from GEO, ArrayExpress, and LINCS L1000 (Level 3 data); (3) Transcriptional responses after perturbation, categorized as upregulated or downregulated. ChemPert database encompasses over 82000 transcriptional signatures from the exposure to 2566 chemical compounds across 167 different non-transformed cells 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 in the biological responses, such as the existence of rare cell subpopulations surviving chemotherapy [**RN88**]. 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 [**RN97**]. Although traditional single-cell RNA sequencing (scRNA-seq) is essential for analyzing tissue heterogeneity, its high cost remains a barrier. Sci-Plex [**RN88**] was introduced to overcome this limitation, by combining two techniques: nuclear hashing and combinatorial indexing-based RNA sequencing (sci-RNA-seq), to assess global transcriptional responses at single-cell resolution [**RN125**, **RN126**]. 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 tagged transcripts from individual cells, enabling high-throughput scRNA-seq at a much lower cost.

Table 2.1: List of resources that provide transcriptomic data driven from perturbagen.

Database	Description	Ref.
ArrayExpress	ArrayExpress is a curated repository of functional genomics experiments (not involving toxic compounds), including microarray and HTS data from various perturbation studies. Established in 2003, it maintains high-quality standards through manual curation, offering structured metadata and accessioned datasets for diverse biological conditions, including compound treatments and diseases. It provides access to data from other sources, such as DrugMatrix [**RN102**] and Open TG-GATES [**RN121**].	[**RN122**]
CDS-DB	Cancer Drug-induced gene expression Signature DataBase (CDS-DB) is a patient-derived cancer drug response database that compiles pre- and post-treatment microarray and RNA-seq data. It provides data for a better understanding of the treatment effects, drug response, and resistance mechanisms. It includes curated metadata from GEO [**RN98**] and ArrayExpress [**RN122**] and supports data browsing, searching, and single signature analysis.	[**RN84**]
CEBS	Chemical Effects in Biological Systems (CEBS) is a publicly accessible repository for toxicogenomic data, established in 2008. It integrates multiple types of experimental data, including detailed study designs and timelines, clinical chemistry profiles, histopathological findings, as well as microarray and proteomics data, collected from studies investigating the effects of both exposure to specific compounds and genetic alterations [**RN124**].	[**RN123**]

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
ChemPert	ChemPert is particularly useful for comprehending the molecular impacts of chemicals in non-cancer cellular contexts. It provides 82270 manually curated transcriptional signatures, derived from 167 non-cancer cell types incubated with 2566 perturbagens (such as small molecules, drugs, cytokines, and growth factors). It also includes the protein targets for 57818 chemical compounds, and the effects (activation, inhibition, or unknown) of those on the targets.	**[RN86]**
CMap / LINCS	Connectivity mapping compiled gene expression responses from human cell lines both of epithelial (MCF7 -breast cancer- and PC3 -prostate cancer) and non-epithelial origin (HL60 -leukemia- and SKMEL5 -melanoma), treated with 164 small molecules prior gene expression profiling using Affymetrix GeneChip. Building on this, the Library of Integrated Network-Based Cellular Signatures clue.io uses Luminex bead arrays to measure 978 reference genes and infer the expression of 11,350 additional transcripts, resulting in a database with over one million gene expression profiles, from 20000+ chemical and genetic perturbagens. LINCS data are available in Phase 1 (GEO access: GSE92742) and Phase 2 (lincsportal; GEO access: GSE70138). iLINCS **[RN114]** provides an interactive platform to explore these datasets and the relationships between compounds, gene expression changes, and disease.	**[RN34, RN30]**
CREEDS	Crowd Extracted Expression of Differential Signatures CREEDS is a manually validated collection that annotated and extracted gene expression signatures from GEO, assembled from crowdsourcing **[RN98]**. The manual curation resulted in 2176 genetics, 875 chemicals, and 828 disease perturbation signatures, from both human and mice. It was later expanded with 8620 genetic, 4295 chemical, and 1430 disease perturbation signatures automatically extracted from 2543 GEO studies.	**[RN87]**

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
DrugMatrix	Public toxicogenomic database (GEO Access: GSE59927; drug matrix) with microarray-based gene expression profiles from rat tissues exposed to 600+ compounds and with 5000+ signatures available. The database includes repeated and single-dose studies at 6 h, 24 h, 3 days, and 5 days. Transcriptomic profiling was performed using the Affymetrix GeneChip Rat Genome 230 2.0 and GE Codelink™ 10,000 gene rat array platforms.	**[RN118, RN119]**
DRUG-seq	Digital RNA with pertUrbation of Genes is a high-throughput RNA sequencing platform designed for cost-effective transcriptomic profiling, reducing costs to just 1/100th of standard RNA-seq. It is suitable for testing multiple treatments in parallel, since it uses in-well cell lysis in 384-well plates, enabling large-scale screening of compounds. In an initial study, DRUG-seq was applied to osteosarcoma cells treated with 433 drugs at 8 dosages. A follow-up study further validated the platform with extensive testing, and introduced an open-source analysis pipeline. Novartis has deposited two datasets in GEO using this technique. The first contains bulk RNA-seq data from U-2 OS osteosarcoma cells treated with 7 small molecules at 3 concentrations for 12 hours. The second treats the same cell line with 14 compounds, each tested across an 8-point dose-response range with 3 replicates per dose. (GEO access: (1) GSE120222; (2) GSE176150; Data analysis pipeline: DRUG-seq).	**[RN105, RN130]**
GEO	Gene Expression Omnibus (GEO) is a public repository for user-submitted transcriptomics data spanning a wide range of perturbation studies, diseases, and experimental conditions across various organisms and platforms. GEO is an appropriate resource for data mining and large-scale transcriptome analysis since it is updated frequently with a vast collection of datasets. GEO provides tools for depositing, querying, and retrieving gene expression and molecular abundance data. It serves as a foundation for more specialized databases.	**[RN98]**

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
GWPS	Genome-Wide Perturb-Seq (GWPS) is a public resource that provides single-cell genetic perturbation data generated by CRISPR interference (CRISPRi) in over 2.5 million human cells (cell lines: K562, chronic myeloid leukemia and RPE1, retinal pigment epithelial). It includes 1946 signatures, each representing a loss-of-function perturbation that triggers a strong transcriptional response.	**[RN89]**
Open TG-GATEs	Toxicogenomic Project - Genomics Assisted Toxicity Evaluation System (Open TG-GATEs) is a public resource that contains 1483 unique signatures from microarray-based gene expression profiles of human and rat liver tissues exposed to 170 compounds. It is built around repeated concentration time-course studies, enabling to evaluate the long-term consequences of exposure to toxic compounds. In addition to transcriptomic profiles, the database also contains histopathology, biochemistry, and hematology data.	**[RN120]**
PertOrg	PertOrg 1.0 **[RN87]** is a public database with curated gene expression data from genetically modified organisms. It curates non-human high-throughput gene expression and phenotypic data from in vivo genetic perturbation experiments in eight model organisms. The perturbation includes gene knockdown, knockout, and overexpression. It currently aggregates 58707 transcriptome profiles and 10116 comparison datasets, including 122 single-cell RNA-seq datasets, with a total of over 8.6 million differentially expressed Gene signatures, retrieved from GEO and ArrayExpress. This tool not only enables the retrieval of the curated data, but it also provides a platform to search and browse various genetic perturbations and to compare gene lists against their signatures, linking perturbations to pathways, cell types, and phenotypic outcomes.	**[RN85]**

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
PerturBase	PerturBase is a public database for single-cell perturbation data. It curated 122 datasets from 46 publicly available studies, covering 24254 genetic and 230 chemical perturbations from approximately 5 million cells, across 31 human and murine cell types. PerturBase is organized into two main modules: the Dataset and the Perturbation modules. The former is for exploring individual studies, whereas the latter for comparing perturbation effects. This resource enables detailed quality control, denoising, differential expression and functional analysis across various cellular contexts, with a direct download option	**[RN97]**
PerturbAtlas	PerturbAtlas is a resource that re-analyzes publicly available RNA-seq libraries to provide detailed, quantitative insights into gene expression, transcript profiles, and alternative splicing after genetic perturbation, including knockdown, knockout, knock in, over-expression, mutations, and multi-condition experiments. Currently, it provides a vast curated collection of 122801 RNA-seq libraries from 7778 studies across 13 species, sourced from ENCODE, GEO, ArrayExpress, and SRA.	**[RN129]**
Sci-Plex	Sci-Plex is a method that combines nuclear hashing with combinatorial indexing-based RNA sequencing (sci-RNA-seq), to quantify global transcriptional responses to thousands of independent perturbations at single-cell resolution and in a single experiment. Sci-Plex screened 3 cancer cell lines (A549, K562, MCF7), exposed to 188 compounds (GEO Access: GSE139944).	**[RN88]**
STARGEO	The Search Tag Analyze Resource for GEO (STARGEO) is an open resource constructed using GEO's publicly available functional genomics data **[RN98]**. STARGEO platform provides 3031859 reliable and annotated samples of gene signatures from humans, mice, and rats.	**[RN127]**

Continued on next page

Table 2.1 – *Continued from previous page*

Database	**Description**	**Ref.**
ToxicoDB	ToxicoDB integrates and harmonizes diverse in vitro toxicogenomic datasets from three sources (Open TG-GATEs [**RN120**], DrugMatrix [**RN102**], and ArrayExpress [**RN122**]). The aim is to easily perform queries and to summarize the relationships between gene expression and toxicant effects [**RN128**]. Currently, ToxicoDB encompasses curated datasets from liver tissue and three cell lines (Hep-G2, HepaRG, and Hepatocytes) in humans and rats, covering a total of 234 compounds.	[**RN128**]

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 [**RN88**]. The integration of the two techniques allowed to simultaneously profile thousands of single-cell transcriptomes across nearly 5000 independent samples. Genome-Wide Perturb-Seq (GWPS) is a large-scale database describing 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 [**RN89**]. With this approach, every expressed gene was silenced, to capture detailed transcriptional responses. This massive dataset allowed 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 [**RN89**]. Additionally, the rich single-cell data enabled a deep exploration of complex biological process, 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. 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.

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 [**RN89**]. The vast array of high-throughput experimental data stored frequently serves as the foundation for more specialized databases, making it an essential building block for several other bioinformatic databases. Additionally, querying GEO often requires manual confirmation of sample identity, since this database does not have an obligation to publish the metadata accurately [**RN115**]. This results in a proportion of the publicly available data not usable, because of lack of associated metadata. To overcome this, databases using data originally from GEO often perform curation and quality checks to ensure that all the metadata are correctly provided. 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 [**RN87**]. 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 [**RN87**]. 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 [**RN85**] is another database built on extracted data from GEO [**RN98**]

and ArrayExpress [**RN122**], to construct a comprehensive tool to analyze and download curated gene expression and phenotypic data from genetically modified organisms. This extensive database contains 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 covers 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 Ribonucleic Acid (RNA)i or similar techniques), gene over-expression (increased expression of a target gene through vector-based methods), 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 via a hypergeometric test, and dataset enrichment analysis for perturbation datasets where the user-given gene set is over-represented [**RN85**]. The huge collection of curated non-human data and all the functionalities provided 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 gene expression changes. However, a better understanding of the effects of a given perturbagen can be achieved integrating transcriptomic data with prior knowledge networks. This allows the mapping of gene regulatory networks, to provide additional information behind those from the snapshot captured by a perturbation signature.

2.3 Prior knowledge Network

Understanding molecular interactions within a biological system is crucial for contextualizing experimental data. Computational methods allow the integration of omics data with prior knowledge of the interactions between biological entities [**RN38**], to better understand the MoA of specific perturbagens. 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 [**RN37**]. 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 [**RN37**]. 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 TF influence gene expression [**RN145**]. 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 PKN should match the data type. For instance, for transcriptomic data, integrating a PKN 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 [**RN37**, **RN103**]. 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 [**RN133**]. This principle also applies to drugs that cause similar transcriptional responses and may have comparable MoA [**RN64**].

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 initially characterized in published experimental research with a varying 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 false negatives and a tendency for well-researched entities to become overrepresented [**RN131**, **RN38**, **RN136**]. 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 do not capture crosstalk among them, providing a static view of a dynamic process [**RN38**]. The entities' overrepresentation issue also applies. Another way to show interactions is, for example, through regulons. Regulons are groups of co-regulated genes controlled by a common TF and are usually represented as GRN (Figure **??**). The choice of network type should be tailored on 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. On the other hand, if a study targets a specific cell type or tissue, it's important to use tissue-specific networks.

Databases like TissueNet [**RN137**] 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.* [**RN131**, **RN135**]. This study

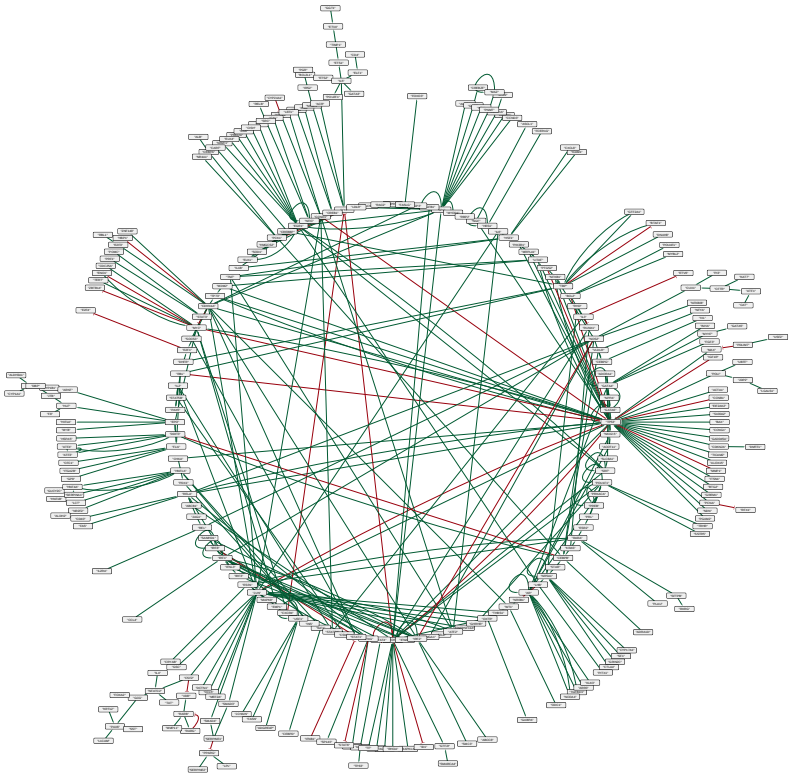

Figure 2.2: Gene regulatory network based on regulon representation of human transcriptional interactions. CollecTRI-derived regulons were extracted from the decoupleR (v. 2.12.0) package, which provided 43,178 interactions. CollecTRI collection [**RN145**] is a comprehensive, curated resource of TFs 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.

aimed to infer the molecular causes of the changes in oxidative phosphorylation gene expression in skeletal muscle from type Diabetes Mellitus Type 2 (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 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 [**RN131**].

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 [**RN138**] 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 [**RN91**] 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.) [**RN91**]. 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 **??**. One of the GRNs that is provided by this database is the CollecTRI-derived regulons [**RN145**] (Figure **??**). This collection contains high-confidence signed 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 [**RN33**] 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 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 [**RN131**]. 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.

Figure 2.3: OmniPath molecular interactions database. This figure presents a hierarchical visualization of human molecular interactions curated by OmniPath, which compiles data from diverse resources. The sunburst graph displays the numbers of distinct interaction types, including post-translational interactions (physical PPI), transcriptional interactions (gene regulatory interactions), post-transcriptional interactions (such as miRNA-mRNA, TF-miRNA, and lncRNA-mRNA interactions), and other interactions involving small molecules (encompassing drug-target, ligand-receptor, enzyme-metabolite, among others).

Table 2.2: List of resources that provide prior knowledge networks and pathways.

Database	**Type of data**	**Description**	**Ref.**
BioSNAP	Network	The Biological Stanford Network Analysis Project (BioSNAP) includes multiple biological networks encompassing different entities and relationships, such as a disease-drug association (1,334,088 edges); side effects-drug relationships (4,649,441 edges); tissue-specific protein-protein interactions (70,338 edges), physical protein-protein interactions identified in human (342,353 edges) and many more. BioSNAP also provides a tool, Mambo, for the construction, representation, and analysis of large multimodal networks.	[**RN140**]
IntAct	Network	IntAct is a public molecular interaction database with over 1,600,000 interactions, derived from literature curation (23,000+ publications) and user submissions (75,000+ experiments) for multiple species.	[**RN139**]
KEGG	Pathways	Kyoto Encyclopedia of Genes and Genomes (KEGG) primarily features metabolic pathways but also includes signal transduction and disease-specific data. KEGG Pathway consists of manually curated pathway maps, illustrating molecular interactions, reactions, and networks across various domains, such as metabolism, cellular processes, human diseases, drug development, etc. KEGG covers 578 human pathways, 12,629 drugs (including 2,499 drug groups), and 2,894 human diseases.	[**RN141**]

Continued on next page

Table 2.2 – *Continued from previous page*

Database	**Type of data**	**Description**	**Ref.**
MetaBase™	Network	MetaBase™ (MetaBase) provides commercially available manually curated networks, larger (higher number of nodes) and denser (higher number of connections) than other publicly available databases. It contains 4.2 M+ molecular interactions with directionality, mechanism, and effect. Although this is primarily described as a network resource, MetaBase also includes curated pathway maps. It covers human, rat, and mouse genes. metabaseR is a software package that facilitates access to MetaBase content in the statistical programming language R.	[**RN33**]
OmniPath	Network	OmniPath is a freely accessible resource that integrates molecular biology data into five main databases: Interactions, Post-Translational Modifications, Complexes, Annotations, and Intercellular. The interactions database contains high-confidence data from different sources encompassing 282,504 unique interactions, organized into various subnetworks. Software tools are also provided for R (OmnipathR) and Python (pypath), together with a plug-in for network visualization (OmniPath Cytoscape) [**RN146**, **RN92**].	[**RN91**]
Pathway commons	Pathways	Pathway Commons is a public repository of pathway and interaction data from 23 databases formatted in the BioPAX standard. It currently includes 6,692 pathways and 3,579,336 interactions from sources like KEGG, IntAct, BioGRID, Reactome, etc. It provides an R package [**RN151**], and the CyPath2 plugin for pathway visualization in Cytoscape.	[**RN142**]

Continued on next page

Table 2.2 – *Continued from previous page*

Database	**Type of data**	**Description**	**Ref.**
Reactome	Pathways	Reactome is a free, open-source, and peer-reviewed pathway database that is manually curated. It includes 2,769 human pathways, 23,911 interactions, 11,574 proteins, and 1,057 drugs, with over 40,000 literature references. The pathways are organized under 29 categories, such as cell-cell communication, signal transduction, etc. Reactome offers a Pathway Browser for visualizing and interacting with the data, as well as an integrated Analysis Tool [**RN150**] for pathway identifier mapping, over-representation, and expression analysis. It also provides two R packages, ReactomePA [**RN148**] for pathway analysis and ReactomeGSA [**RN149**] for Multi-Omics Comparative Pathway Analysis, and the ReactomeFIPlugIn for pathway visualization in Cytoscape.	[**RN143**]
STRING	Network	STRING is a public database that includes known and predicted PPI, derived from genomic context predictions, high-throughput lab experiments, co-expression data, automated text mining, and existing databases. STRING currently covers over 59.3 million proteins from 12,535 organisms, providing a comprehensive PPI network with confidence scores based on various data sources.	[**RN72**]
WikiPathways	Pathways	WikiPathways iis another open-source curated pathways database. It provides pathways for 27 organisms and includes 959 curated human pathways. The platform supports collaborative annotation and refinement of pathways. It also provides several software tools and workflows in R (rWikiPathways) and Python (pywikipathways), along with Cytoscape and PathVisio plug-ins for pathway visualization, among others [**RN147**]	[**RN144**]

2.4 Computational methods for MoA inference

Due to technological advances, large-scale transcriptomics datasets can now be generated affordably for many perturbations. However, extracting biological insights from this data can be complicated, requiring dedicated computational tools for the analysis. 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 [**RN57**, **RN56**]. 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 [**RN38**]. For example, in Hill *et al.*'s study, [**RN37**] 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 generating mechanistic hypotheses about the cellular targets and pathways affected by a perturbagen, in an easier and more accurate manner [**RN100**].

2.4.1 Connectivity-based tools for comparative analysis

Similarity-based approaches for CMap focus on matching gene expression signatures from query samples to reference profiles. The approach emerged from the need to connect changes caused by perturbagens, with those observed in diseased or other biological states. As stated by Lamb *et al.* [**RN34**] the resource was names CMap, due to its potential to connect drugs, genes, and diseases, with the foundational idea that if a compound induces a transcriptional signature similar to a known profile, it likely shares the same MoA or therapeutic effect. The three main components of this approach are: a query signature, a ranked list of genes up- or downregulated in a condition of interest and a reference database. Using a pattern-matching algorithm, the query signature is systematically compared against a reference database of perturbation-induced expression profiles [**RN155**]. The seminal study that proposed the first Connectivity-based approach also introduced the first CMap database, described in Chapter 2.2. As a strategy to score the similarity between data, the authors employed a nonparametric, rank-based Kolmogorov-Smirnov (KS) test [**RN34**, **RN79**]. This approach has since been adapted and extended by several methods, including those based on weighted KS statistics, such as the variations of the Gene Set Enrichment Analysis (GSEA), and alternative metrics such as the eXtreme Sum (XSum) score and signed rank-based methods like the ZhangScore [**RN79**].

For each reference profile, the algorithm evaluates whether query upregulated genes cluster near the top of the ranked list and downregulated genes near the bottom, indicating a positive connection, or a negative connection. This yields a connectivity score between

1 (strong concordance) and -1 (strong anticorrelation), with scores near zero indicating no significant association [**RN34**]. In this way, these methods not only generate direction of connectivity but also provide information on the strength of the link. The similarity scores can be directly linked to biological interpretations. A high positive score may indicate that a compound could enhance a biological condition, while a high negative score suggests a potential inhibitory effect [**RN102**].

CMap has become a powerful tool for drug repurposing and mechanistic studies because it leverages large-scale gene expression data, to connect drugs, diseases, and gene perturbations. By comparing a query signature with a vast database of compound-induced profiles, candidate drugs that may reverse or mimic disease-specific transcriptional changes can be identified [**RN102**]. This approach not only facilitated the creation of several drug-induced molecular perturbation signature databases [**RN84**], but also supported studies applying the CMap concept in drug repositioning and MoA elucidation [**RN86**]. Consequently, this resulted in prediction and experimental validation of novel applications for existing drugs

2.4.2 Enrichment-based tools for downstream analysis

As the number of gene expression studies increases, it becomes harder to extract relevant information. Pathway analysis helps to frame large changes in gene expression that, if isolated, lack biological context. These methods convert gene lists into a meaningful and interpretable biological process, by linking expression data to specific biological pathways.

Biological pathways are identified by characterizing the cascade of interactions occurring in the system in response to a given perturbation. These interactions can be observed in transcriptomic data and translated into significantly enriched pathways, to capture how changes in gene expression are preferentially affecting some pathways more than others. However, pathway analysis does not trace the causal paths leading to observed gene expression, but rather it infers which pathways might be affected by changes in gene expression. One commonly used method to detect these relevant pathways is GSEA (Gene Set Enrichment Analysis). The input for GSEA is a ranked list of genes, based on metrics like fold-change and a predefined gene set. The algorithm then checks if each one of a pre-defined list of gene sets cluster at the ends of the ranked list. By calculating an enrichment score (ES), GSEA shows how genes are distributed at the top or bottom of the list. A Kolmogorov– Smirnov test is then applied to evaluate the statistical significance of the enrichment [**RN152**]. Since it is a rank-based approach, using a full range of gene expression changes avoids arbitrary gene-level thresholds, as there is no need for a strict cutoff for DEGs. This reduces bias and increases sensitivity to pathways where individual gene changes are small but consistent. However, this thoroughness comes with a high computational cost.

Over-Representation Analysis (ORA), on the other hand, is a simpler and faster method

for pathway enrichment. It's ideal for large-scale screening and initial data exploration. This analysis uses a hypergeometric test or Fisher's exact test, a background gene set and list of genes of interest, meeting a specific fold-change or statistical threshold (such as DEGs). The statistical test compares the overlap between the input gene list and each pathway for each gene in the list. Since ORA only needs a non-ranked gene list, it can be more straightforward than GSEA. However, by relying on a cutoff for the input list the pathway detection is affected by a certain degree of subjectivity. Additionally, ORA does not consider the magnitude or direction of gene expression changes and, by assuming independence among genes and pathways, it may overlook complex interactions.

Overall, enrichment-based methods help provide a broader and better view of biological processes. However, they lack the context of mechanisms behind observed changes. One way to account for the cross-interaction between different genes within the system is through topology or network-based algorithms. These algorithms can combine transcriptomic data with network information to predict drug or disease targets.

2.4.3 Topology-based methods for upstream analysis

While pathway enrichment methods compare gene expression data with the respective encoded proteins and the signaling pathways, on the other hand, topology-based methods treat a gene expression pattern as the result of a specific perturbation. Several algorithms fall into the category of topology-based methods, as they take advantage of a topology network as prior knowledge. A subset of them is categorized as causal reasoning algorithms. Causal reasoning can be described as the process of looking at what happened, the effects, and trying to infer the upstream causes. In this context, change in a process in response to specific stimulus is considered as a proxy for causality, in line with the following axiom "A causes B if a change in A leads to a difference in B, assuming everything else stays the same." [**RN157**]. This thinking has deep philosophical roots, and, in systems biology, it has a more testable and specific meaning. In early medical applications of causal reasoning, "A" represented the treatment, and "B" was the outcome.

In DD studies, this concept can be applied by overlaying high-throughput measurements, such as gene expression data, onto a network. The algorithm itself is a sophisticated causal node prioritization tool that makes predictions based on network topology. It requires direct interactions, meaning directed cause-effect relationships, connecting each pair of biological entities. Additionally, it uses edge effects, to know (at least for some edges) whether it represents activation or inhibition. For DEGs, it doesn't necessarily require the exact fold-change values; instead, it is sufficient to specify which genes are up- and down-regulated. At its core, a causal reasoning algorithm works backward. It starts with a node in the network and follows it downstream for a few steps, predicting what would happen if this node were activated (or repressed) in the biological system of interest. For example, by analyzing the network structure, it can determine if the activation of a specific regulatory protein will also activate a certain TF, directly affecting

the expression of target genes in a positive or negative way (respectively by activating or repressing gene expression). The algorithm makes these predictions and then compares them with what is observed in the data. This helps identify which upstream regulators best explain the downstream changes that are seen [**RN32**]. However, not every edge in a prior network is active during a specific experiment, and not all predicted downstream effects occur. This is the reason why statistical scoring is crucial. In causal-network inference, all evaluation methods compare each regulator predicted downstream effects against the observed gene expression changes, but the exact scoring metric depends on the type of network used [**RN81**]. The simplest approach merely counts how many predictions are correct and incorrect. A more refined enrichment score (ES) uses Fisher's exact test to evaluate whether the targets of a regulator are enriched among the differentially expressed genes while ignoring whether each edge is activated or inhibited. To capture that extra layer of information, the Quaternary Scoring (QS) computes a z-score to quantify how well the predicted up/down directions match the observed up/down changes in the data [**RN156**]. Early work by Pollard *et al.* [**RN135**] combined the two methods, computing an overlap and concordance p-values for regulators in type-2 diabetes expression data [**RN156**]. However, Fisher's test cannot leverage signed edges even when those signs are known. To address this, Chindelevitch *et al.* [**RN73**, **RN131**] developed a ternary scoring (TS) method that models the signed network and observations as a dot-product distribution, allowing exact p-values for both the activation and inhibition hypotheses of each regulator. This scoring method requires a fully annotated causal graph, i.e., every edge must have a known direction and effect information. More recently, Fakhry *et al.* [**RN158**] introduced a QS method that extends TS to networks containing both signed and unsigned interactions, preserving directional inference wherever possible while still accommodating unannotated edges [**RN81**]. Early methods used simpler calculations, while newer methods used more precise statistical tests to ensure accuracy. However, simpler z-score and enrichment-based methods are still popular because they are faster to compute and easier to use and understand. Causal reasoning algorithms are widely used but have many nuances depending on the network's structure and how far one can trace from regulators to downstream genes. Improving its accuracy in predicting regulators is a key focus. Also, they are complex and computationally demanding, especially with the increase in input data.

There are evidences that topology-based methods may be more useful in DD than pathway enrichment methods. Topology-based causal reasoning takes advantage of directed and signed molecular networks to infer the most plausible upstream regulators from transcriptomic endpoints. By inferring key signaling nodes whose activity can directly explain the changes observed in the experimental data, these methods pinpoint candidate drug targets and generate mechanistic hypotheses for further validation. On the other hand, pathway enrichment methods only infer which biological processes are affected by changes in gene expression. These methods do not consider that after transcription, events such as translation and post-translational modifications will also affect protein

activity [**RN53**].

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 multidimensional biological 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 benchmarked against popular existing methods. Without a deep expertise in the area, distinguishing the benefits of a novel tools can be difficult.

A comprehensive benchmarking study 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 [**RN108**]. A key component of a benchmarking study is the use of gold-standard datasets, against which results are compared. By this process, often formalized via well-defined good practises, it is possible to evaluate performance metrics and statistical analyses and to compare different algorithms on specific types of data.

Benchmarking studies can be carried out by the authors that implemented the tool, independent groups, or as organized challenges, such as those by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) Consortium [**RN109**]. If the evaluation is performed by the authors, the aim is usually to demonstrate the advantages and performance improvements over other techniques. In other contexts, it is important to define the benchmark's scope and purpose. 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 mechanism of action studies, it is crucial to include diverse data sources and generation methods, to ensure representativity and a credible assessment of performance. For instance, transcriptomic data 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 [**RN109**]. The same applies to the gold standard datasets, which serve as the ground truth, and they represent the essence of a benchmarking study.

Other similar studies arise from an effort to contextualize gene expression data with several computational algorithms. Hosseini-Gerami *et al.* [**RN53**] evaluated the performance of different causal reasoning algorithms to recover direct targets of small molecules and associated signaling pathways using gene expression data. The study compared four causal reasoning algorithms against networks from two diverse sources and transcriptomics data from one database. Hill *et al.* [**RN37**] 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 when comparatively analyzing different algorithms and integrating different resources, without providing a complete exploration of the parameter landscape (considered beyond the scope of the work).

Typically, a 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 [**RN37**]. By providing a robust assessment of the capabilities of existing algorithms, these studies provide guidelines for researchers to choose the most appropriate tool for the scientific question of interest [**RN108**].

3

Materials and Methods

The following section describes the workflow of the benchmarking study. It begins by describing the input data used, both transcriptomic 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 R&D 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 R&D 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 Statistical Software (v4.4.1) [96].

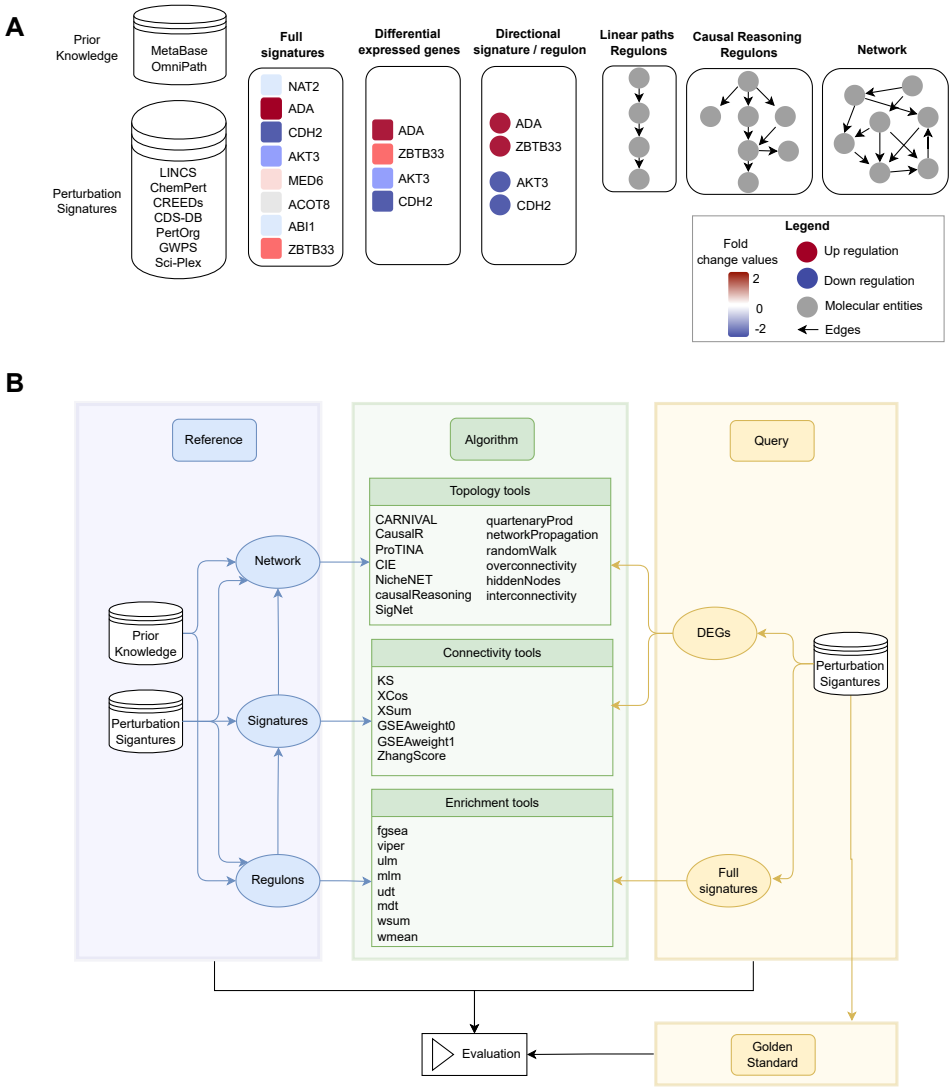

Figure 3.1: Schematic of the study architecture. A. Perturbation signatures collected from seven public sources are used in the benchmarking framework either as reference, query, and gold standard (known targets) datasets. Prior knowledge networks, used as reference, were derived from two sources: OmniPath (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 transcriptomic datasets from 10 different public sources, summarized in Table **??**. Chemical and genetic perturbagens were included, analyzed by 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 and it consists of a set of known targets (drug-protein interactions or genes deliberately perturbed), 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 CLUE). 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 annotations were used to identify molecular targets inhibited by specific compounds; 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 by microarray or RNA-Seq (including 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. For the golden standard, perturbation effects were considered as activation in the case of knock-in, overexpression and activation, and inhibition for all 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 concept of causal inference can be described as the ability of algorithms to find the target candidates of a perturbation, based on gene expression data generated from a specific experimental study. 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

Table 3.1: Summary of gene expression datasets used in this study. Each dataset includes transcriptomic signatures derived from chemical or genetic perturbations. The "Perturbagen" column specifies the type of compound or gene perturbation applied, while the "Type" column labels it as either chemical or genetic. The number of perturbagens refers to the unique compounds or genes perturbed in the dataset. The signatures with the golden standard column indicate how many signatures have associated targets that can be used for benchmarking. The signature type describes the format and content of the signature, such as full profiles or DEG lists.

Data set	Perturbagen	Type	Number of perturbagens	Signatures with golden standard	Signature type	Ref.
LINCS compounds	Compound (small molecules)	Chemical	33627	3540	Full	[**RN30**]
ChemPert	Compound (small molecules, ligands, drugs)	Chemical	2508	1304	DEGs (up/down gene sets)	[**RN86**]
CDS-DB	Compound (small molecules) Patient-derived	Chemical	181	181	Full	[**RN84**]
Sci-Plex	Compound (Single cell; Different doses)	Chemical	189	405	Full (scRNA-seq)	[**RN88**]
CREEDs	Disease, small molecules, single gene perturbations	Chemical Genetic	3051 (875 drugs, 2176 genes)	2642	DEGs (up/down gene sets)	[**RN87**]
LINCS CRISPR	CRISPR KO	Genetic	5156	5156	Full	[**RN30**]
LINCS OE	cDNA over-expression	Genetic	3780	3780	Full	[**RN30**]
LINCS shRNA	shRNA interference	Genetic	4854	4854	Full	[**RN30**]
PertOrg	shRNA interference; CRISPR knockdown; Over-expression	Genetic	7398	951	DEGs	[**RN85**]
GWPS	CRISPR interference	Genetic	1946	1946	Full (scRNA-seq)	[**RN89**]

targets were identified. When using signatures derived from drug perturbation, it can be hard to identify the exact compound used only from gene expression. Instead, it's easier and more meaningful to infer the target(s) of the compound (i.e. the biologically active protein that the drug binds to). Although MetaBase also contains compound information, most networks do not, but they do include gene or protein targets that can be used as proxy instead. 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, including all target gene symbols. (2) Small molecules were mapped against the drugs in the CDDI database, then, depending on the annotation level, one or more of the following information were included for downstream analyses: target drug annotation, names, synonyms or structural information. (3) The target lists provided by the authors and the one from Cortellis Drug Discovery Intelligence (CDDI) were then merged to form the final set of targets for each therapeutic agents.

3.2.2 Prior Knowledge: Interaction Networks

One of inputs that can serve as a reference is the prior knowledge data, required 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 **??**). Although these resources vary in their coverage, they are interconnected, as illustrated in Figure **??**. Including sources of different sizes and densities is particularly important for understanding how the performance of topology-based algorithms is affected by the type of the input. Additionally, an increase in network size can also introduce noise that may disturb the extraction of biologically relevant information.

The interactions are obtained from two databases, OmniPath [**RN91**] and MetaBase [**RN33**]. OmniPath is a public database with protein-protein, transcriptional, and RNA-related interactions. MetaBase™ is a manually curated systems-biology database, provided by Clarivate, containing over 4 million directional molecular interactions, such as protein-protein, protein-RNA, compound-protein, etc. From each of these two sources, PKN and regulons were obtained as used as input in the benchmarking process. Canonical linear pathways were extracted only from MetaBase and annotated according to four main concepts: 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, coded as 0) from transcriptional regulation edges (transcription factors with their target genes, coded as 1). Finally, weight determines interaction confidence based, among the others, on literature support. Regardless of the database source used, the mandatory annotation for each interaction is directionality information, whereas any other information will 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 and with assigned mechanism = 0, whereas transcriptional regulatory interactions imported using import _ transcriptional _ interactions were annotated as mechanism = 1. These were combined into a single network, and interactions with effect = 0 were kept only if mechanism = 0. Nodes in OmniPath are proteins or protein complexes (UniProt IDs), with the corresponding gene symbol(s). For using MetaBase as another source of PKN, the global network was extracted using the networkFromMetabase function, via the metabaser (v. 5.1.0) and CBDD (version 20.0.3) R packages. Unlike OmniPath, it already includes both signaling interactions (mechanism = 0) and transcription regulation interactions (mechanism = 1). Only high-confidence interactions with defined effect (activation or inhibition) were kept. Originally, the network contained specific MetaBase network objects, that were processed to add only the corresponding gene symbols to the network (using the function metabaser::annotate.nwobj2gene).

Table 3.2: Summary table of OmniPath and MetaBase prior knowledge resources. Number of nodes and edges are displayed for each resource. Network refers to the full interaction network, while regulons and linear path regulons are downstream-derived, regulators subsets. The regulator and target columns correspond to the number of source and corresponding target nodes, respectively. Gene space counts how many of those nodes correspond specifically to genes (not proteins or others). The three edge-type columns indicate the number of activation, inhibition, or transcriptional regulation interactions. The total number of interactions for each resource is under total column.

		Nodes			**Edges**			
Resource		Regulator	Target	Gene				
space	Activation	Inhibition	Transcriptional regulation	Total				
OmniPath	Network	6,166	6,723	7,809	119,113	13,680	64,367	145,896
	Regulons	4,442	6,723	5,622	5,842,390	4,270,032	10,112,422	10,112,422
Metabase	Network	33,927	15,229	17,693	81,866	61,214	101,752	657,746
	Regulons	11,739	10,476	9,988	23,844,526	21,469,352	45,313,878	45,313,878
	Linear Path Regulons	2,922	9,465	3,185	3,493,007	1,361,149	4,854,156	4,854,156

Another way of representing interactions that can be used as reference data for both topology-based and enrichment-based tools are the regulons. For both the resources regulons were extracted using the causal reasoning algorithm. through the CBDD::hypothesisGeneration function, providing as parameters the downstream depth for the search (in this case, 4 steps) and the position in the pathway where transcription regulation links may appear (set to anywhere in the path). Bearing in mind that both networks used as input contain the directionality of the signal, this function will then predict which targets are influenced (by activation or inhibition) by each specific regulator. Finally, all possible interactions were filtered to retain only those where a node and all downstream activated or repressed genes are present. The final number of nodes and interactions of the regulons is also detailed in Table **??**. In addition to the network and the regulons, canonical linear pathwayss, available from MetaBase were also included. Those are linear sequences of biological entities and interactions between them. They are automatically generated from pathway maps and represent highly curated canonical signaling paths. They start from an important signaling molecule and end, usually with a transcription regulation event or another downstream molecular response.

3.3 Algorithms: implementation and wrapper function's architecture

To carry out a systematic and robust comparative evaluation of inference algorithms, wrapper functions were developed to build a common framework and to standardize the input data and output, so to ensure compatibility between each algorithm data requirements and processing methods. A wrapper is a function that serves as an intermediary layer. These are required 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 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 **??** 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 matching IDs and converting the input data, if 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. As reference datasets, the workflow can start with PKN or full signatures, whereas for topology, enrichment, and CMap tools respectively require networks, regulons/gene sets, and full signatures. To use this large variety of input data and tools, some conversions are required. All the conversions are indicated by the arrows in Figure **??** B. The parameters selected for each algorithm can be found in Supplementary table 2.

As for the input data, the output should also have a similar shared format, so to make it 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, with the larger score reflects greater confidence in this regulator being causal for the observed differential expression patterns. Score may also 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 (coded as -1 or 1 respectively).

3.3.1 Connectivity Mapping

Figure **??** represents the wrapper function framework for running CMap algorithms from the RCSM package [**RN79**]. This package provides uniform implementations of several CMap scoring methods including KS, and GSEA-based approaches. The function is designed to receive filtered DEGs lists as query input and full perturbation signatures as reference data. If full signatures are used as the query, they are internally 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, designed to quantify the similarity between query and reference perturbation signatures.

Table 3.3: Summary of computational methods evaluated in the benchmarking study. A total of 27 algorithms categorized into the following methodological approaches: 1) Enrichment, 2) Connectivity Mapping, 3) Topology with Causal Reasoning, and 4) Topology with Node Prioritization. For each algorithm, the corresponding R package implementation and version used are reported. The "Reference" and "Query" columns indicate the required input data types to run the algorithm. The "Output" column specifies whether algorithms produce node rankings (prioritized lists of potential regulators) or subnetworks (connected components representing regulatory cascades involved in the MoA).

Method	Algorithm(s)	R Package	Reference	Query	Output	Ref.
Enrichment	fgsea viper ulm mlm udt mdt wsum wmean	decoupleR (v. 2.12.0)	Regulons	Full signatures	Node ranking	**[RN35]**
Connectivity Mapping	KS XCos XSum GSEAweight0 GSEAweight1 ZhangScore	RCSM (v. 0.3.0)	Full Signatures	DEGs	Node Ranking	**[RN79]**
Topology (Causal Reasoning)	CARNIVAL	CARNIVAL (v. 2.16.0)	Network	DEG/Full Signatures	Subnetwork	**[RN41]**
	CausalR	CausalR (v. 1.38.0)				**[RN32]**
	ProTINA	Protina (v. 0.1.0)			Node ranking	**[RN80]**
	CIE	CIE (v. 1.0.0)			Node ranking	**[RN81]**
	NicheNet	Nichenetr (v. 2.2.0)			Node ranking	**[RN42]**
	causalReasoning	CBDD (v. 21.0.0)			Node ranking	**[RN73]**
	SigNet				Node ranking	**[RN74]**
	quaternaryProd					
Topology (Node Prioritization)	networkPropagation	CBDD (v. 21.0.0)	Network	DEG/Full Signatures	Node ranking	**[RN75]**
	randomWalk					**[RN76]**
	overconnectivity					**[RN77]**
	hiddenNodes					
	interconnectivity					**[RN78]**

Those algorithms include the KS statistic, used in the original CMap [**RN34**]; 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, which ignore the fold-change magnitude for computation; GSEAweight1 with parameter p = 1, where fold-change magnitude contributes linearly to the final score; and Zhang, a CMap score first suggested in [**RN161**]. The wrapper function handles the different algorithm requirements by preparing either separate up- and down-regulated gene lists or simple gene vectors for XCos, also including 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 [**RN35**] was used. The package was initially used to benchmark approaches for TF activity inference. 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). Of all the algorithms already implemented in decoupleR, only GSEA and the others that respect directionality were considered for this benchmarking effort. As for CMap algorithms, a shared wrapper function (Figure **??**) was built to prepare the input and output data, 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 FC matrix and ID space conversions can also performed, if required to match network node identifiers. For algorithms that support directional information, the wrapper uses the edge type from the network. Extra parameters can be supplied to the argument list (Supplementary table 2). The output of the wrapper function consists of regulator rankings with scores, effects, and p-values (if returned by the algorithm) organized by signature name and sorted by absolute score magnitude

3.3.3 Topology-based methods

CARNIVAL [**RN41**] integrates different sources of PKN, including signed and directed PPI, TF targets, and pathway signatures, to yield a causal subnetwork explaining the MoA behind the observed omics data. This algorithm expects query DEGs as input and a

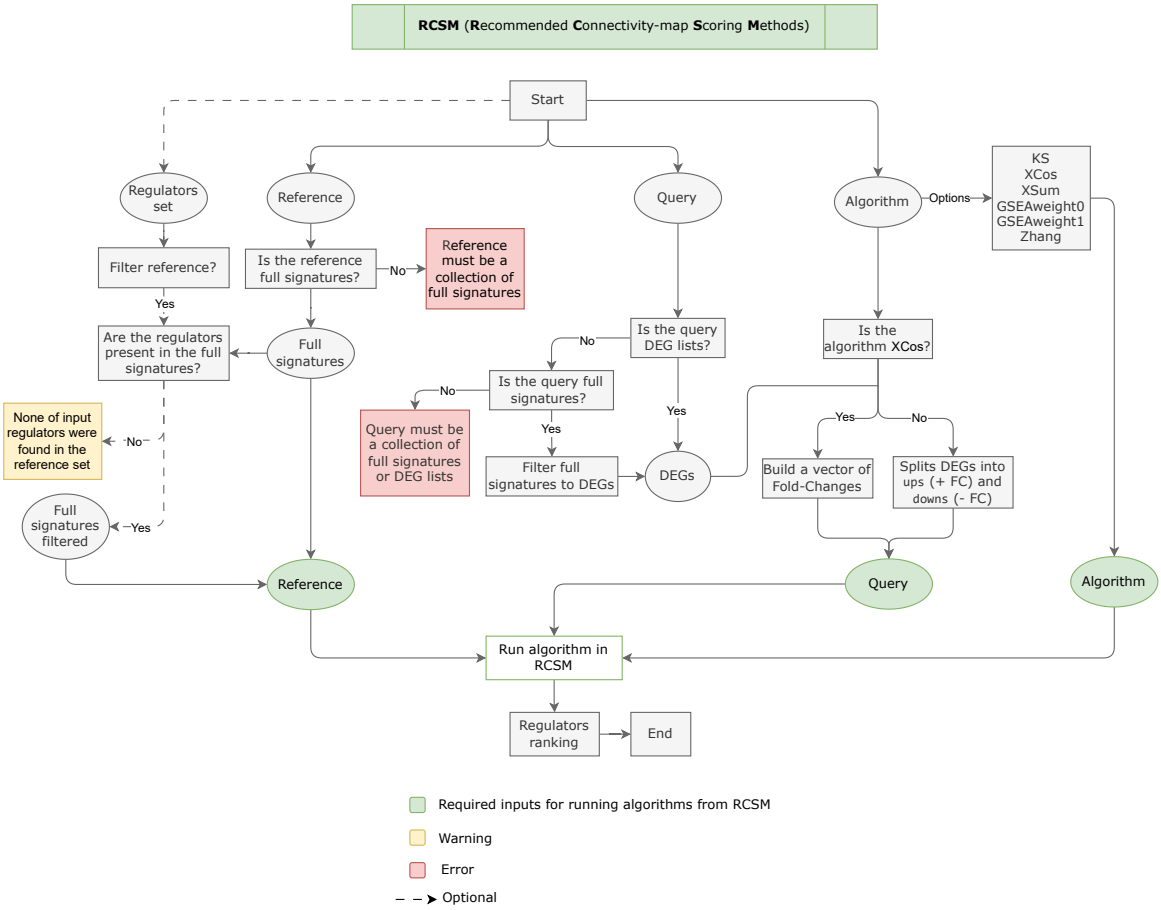

Figure 3.2: Flowchart representing the main steps for implementing connectivity mapping algorithms pre-built in the RCSM R package. The general computational pipeline for executing connectivity-based methods, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicate required inputs, while red highlight potential errors.

network as reference. CARNIVAL wrapper (Figure 7) begins by validating essential inputs, including the CBC solver path required for the optimization engine. It determines the execution mode based on whether perturbation targets and pathway weights are provided, enabling either the standard CARNIVAL or the inverse CARNIVAL algorithms. For reference network preparation, the wrapper handles multiple input formats by converting if needed signature collections or DEG lists into networks (with source-interaction-target). The full signatures are also converted into the matrix format. CARNIVAL performs TF activity inference with either a network-derived approach, if effect and mechanism are present, or using DoRothEA regulons from decoupleR. As an option, depending on the execution mode selected, a pathway activity score can be calculated using PROGENy from decoupleR. The network is filtered to contain a subset of the interactions, keeping only nodes reachable from relevant TF. When the CARNIVAL algorithm is executed, the

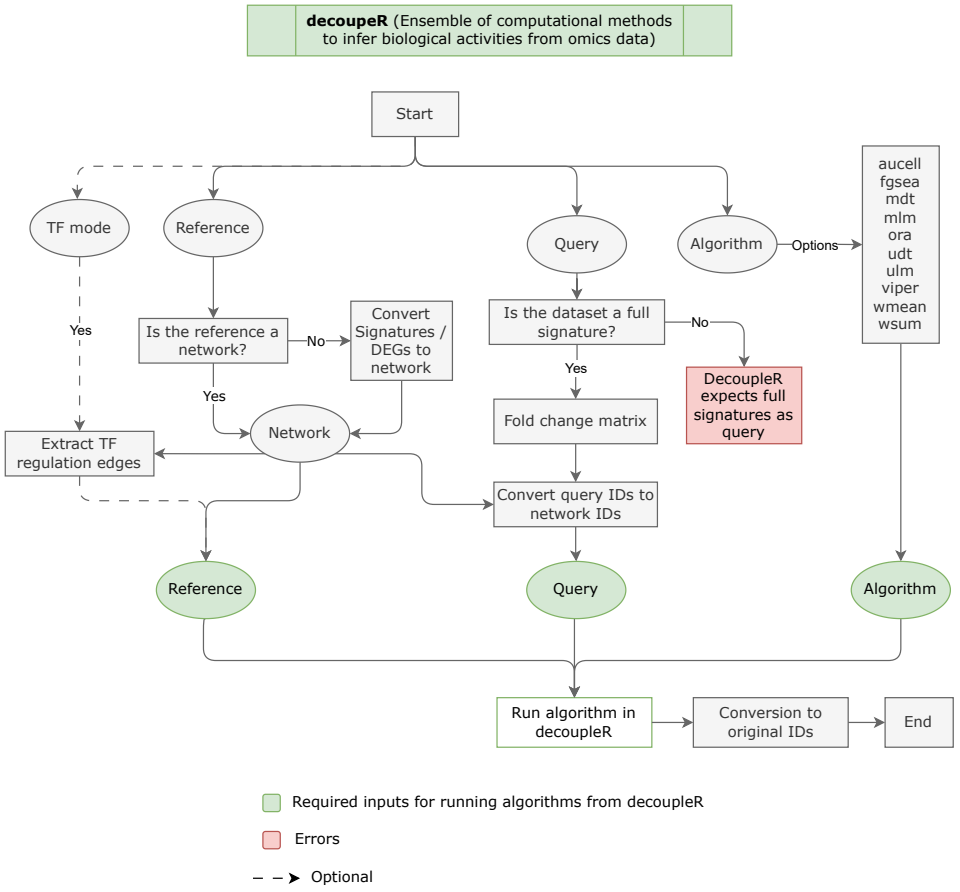

Figure 3.3: Flowchart representing the main steps for implementing enrichment algorithms pre-built in the decoupleR package. The general computational pipeline for executing enrichment-based methods, showing the main input requirements, data preprocessing. Green indicates required inputs, while red highlights potential errors.

results are provided with regulator scores representing the proportion of solutions where each node appears in the causal subnetwork.

Protein Target Inference by Network Analysis (ProTINA) [**RN80**] algorithm generates protein activity scores based on gene expression changes through network perturbation analysis. The wrapper (Figure **??**) starts by checking if the query is a collection of full signatures and if the reference data is a network. If the reference is a signatures or a DEGs list, it is then converted to a network. Then, query data is filtered to remove entire signatures without any DEGs, as ProTINA cannot process full profiles with only zero fold-change expression. After that, the full signatures are converted to a matrix of fold changes, to remove genes with zero standard deviation across all signatures. ProTINA was originally designed to handle experiments with multiple replicates, time points, or dosages for each perturbation. Since none of the data used for benchmarking has measurements in different conditions, the function only creates a vector with consecutive integers starting from 1 to the number of signatures, representing the number of groups. Reference data

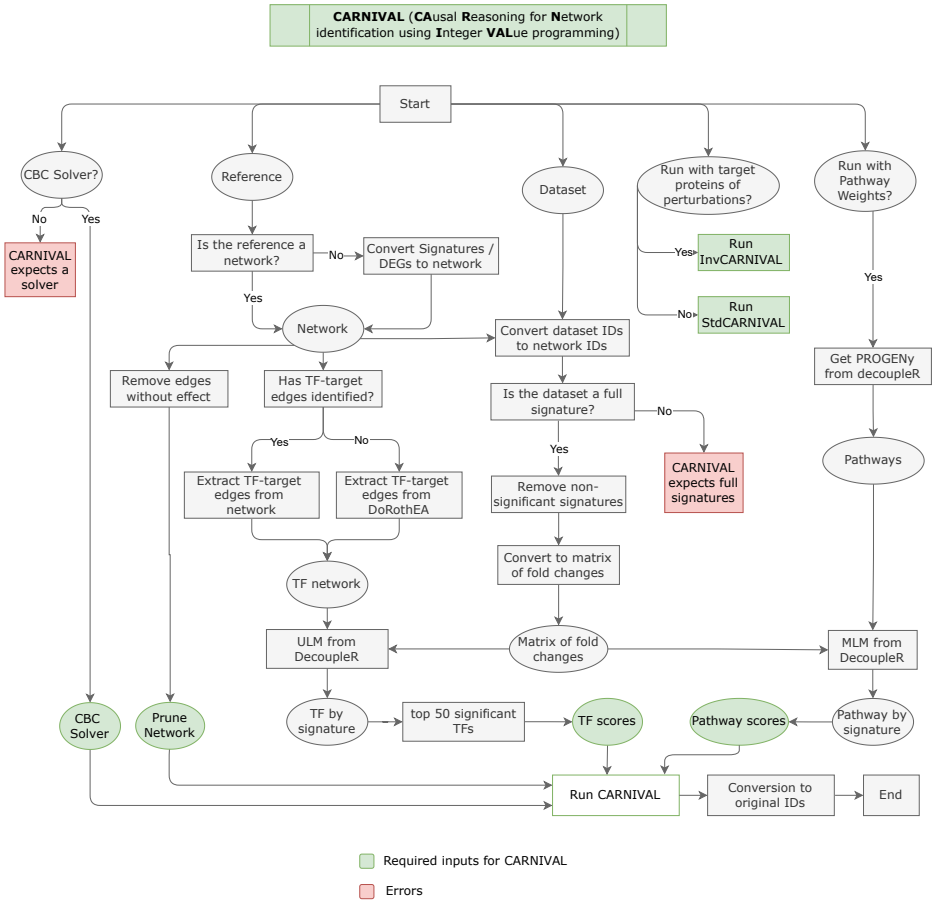

Figure 3.4: Flowchart representing the main steps for implementing the CARNIVAL R package. The general computational pipeline for executing this topology-based method, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicates required inputs, while red highlights potential errors.

is processed to construct a PGN, which combines PPI from the input network with TF-target gene relationships. From the PGN it is calculated an adjacency matrix where rows represent proteins, columns represent genes, and matrix elements indicate the regulatory relationships. Finally, the algorithm returns a matrix of protein (regulator) activity scores (signed Z-scores) for each perturbation group. The wrapper converts the results back to gene symbols and ranks regulators by their activity scores, providing both magnitude and directionality of predicted protein activities.

To identify DEGs, the CausalR **[RN32]** wrapper (Figure **??**) starts by processing the query signatures using the common filtering parameters described above. The resulting data is then converted to the required CausalR format where genes are assigned with regulation status values (1 for upregulation, -1 for downregulation, 0 for unchanged). Non-DEGs within the signature are marked as unchanged, rather than excluded. The final format for the query dataset is a matrix with two columns for each experiment, reporting gene name and regulation status (1, -1, or 0) instead of fold-change values. As

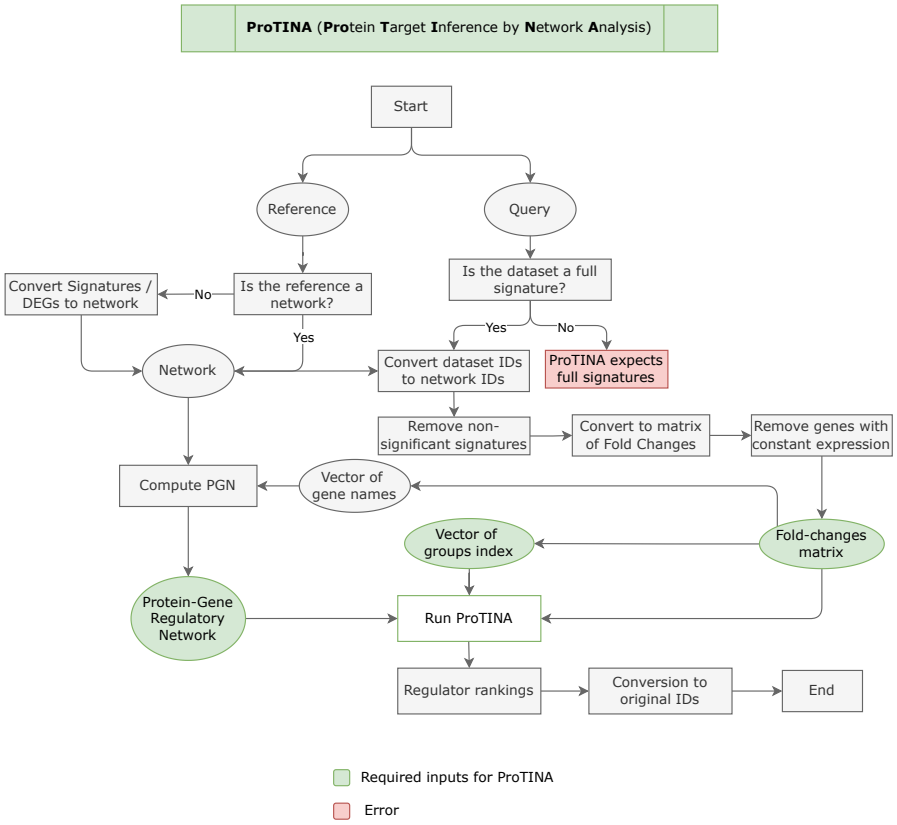

Figure 3.5: Flowchart representing the main steps of ProTINA algorithm implementation. The general computational pipeline for executing this topology-based method, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicates required inputs, while red highlights potential errors.

for ProTINA, signatures that do not have any DEGs are removed. If reference data are not already provided in the correct format, the function can convert signature collections or DEG lists into directed causal networks. Edges connect the regulators (signature names) to their targets (DEGs) with edge effects corresponding to DEG fold-change signs. After having a network as reference, CausalR requires the construction of a Computational Causal Graph (CCG), which contains twice the number of vertices and edges as each regulators is reported both as up and down/regulated. CausalR allows two options of algorithms, RankTheHypotheses and runSCANR. RankTheHypotheses algorithms use the configurable path-length parameter delta, to control how many network edges can be traversed from regulator hypotheses to the observed gene expression data, enabling from a direct transcriptional regulation (delta = 1) analysis to a multi-step causal cascade (delta > 1). Finally, the algorithm returns, per each signature, a regulator ranking with the regulator name, and the corresponding score (difference between correctly and incorrectly predicted DEGs), p-value, and the predicted regulatory effect.

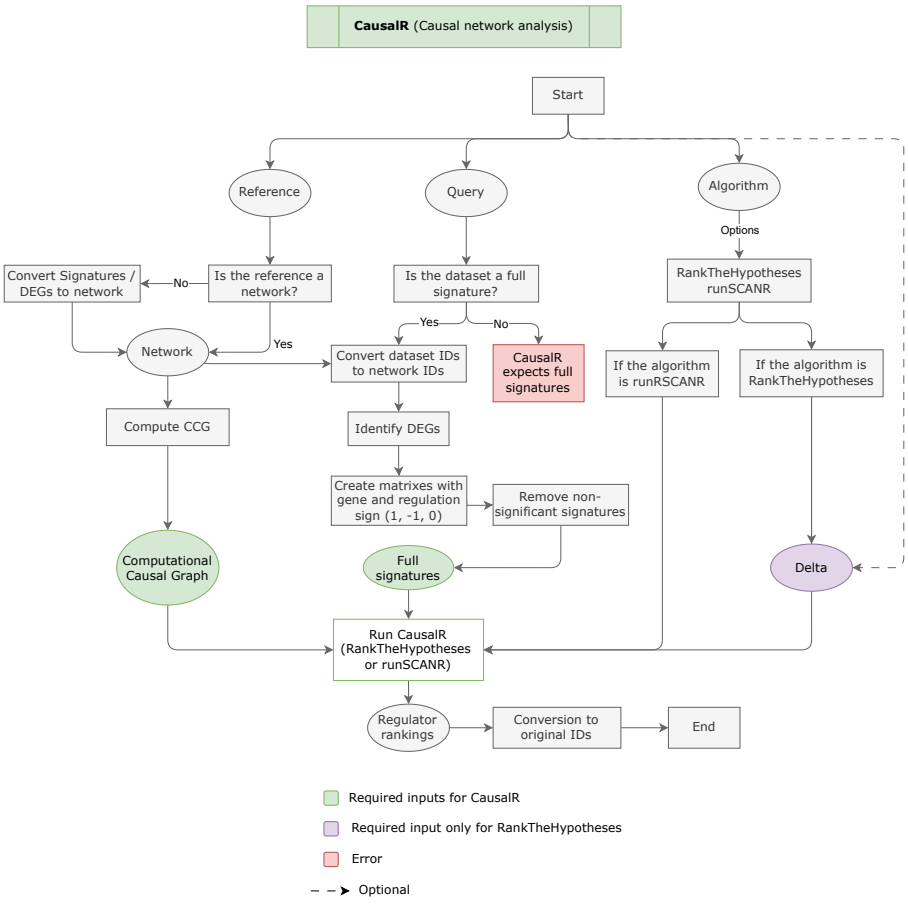

Figure 3.6: Flowchart representing the main steps of CausalR algorithm implementation. The general computational pipeline for executing this topology-based method, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicates required inputs, while red highlights potential errors.

The CIE **[RN81]** wrapper (Figure **??**) starts by converting (if needed) full signatures to DEGs, then transforming reference data into a network format. A key aspect of this implementation involves preparing the network data structure to create two essential inputs: network entities (nodes) and network relations (edges). The entities data frame contains unique gene identifiers classified as "mRNA" or "Protein" based on the targets of the transcriptional regulation edges (mechanism = 1). The relations data frame links source and target nodes using their distinct identifiers and maps the network edges with their regulatory effects. The function supports three CIE algorithms: Fisher for unsigned networks, Ternary for completely signed networks, and Quaternary for partially signed networks (default). The results from CIE are the regulators and the corresponding causal reasoning scores, regulatory effects, and a p-value ranking.

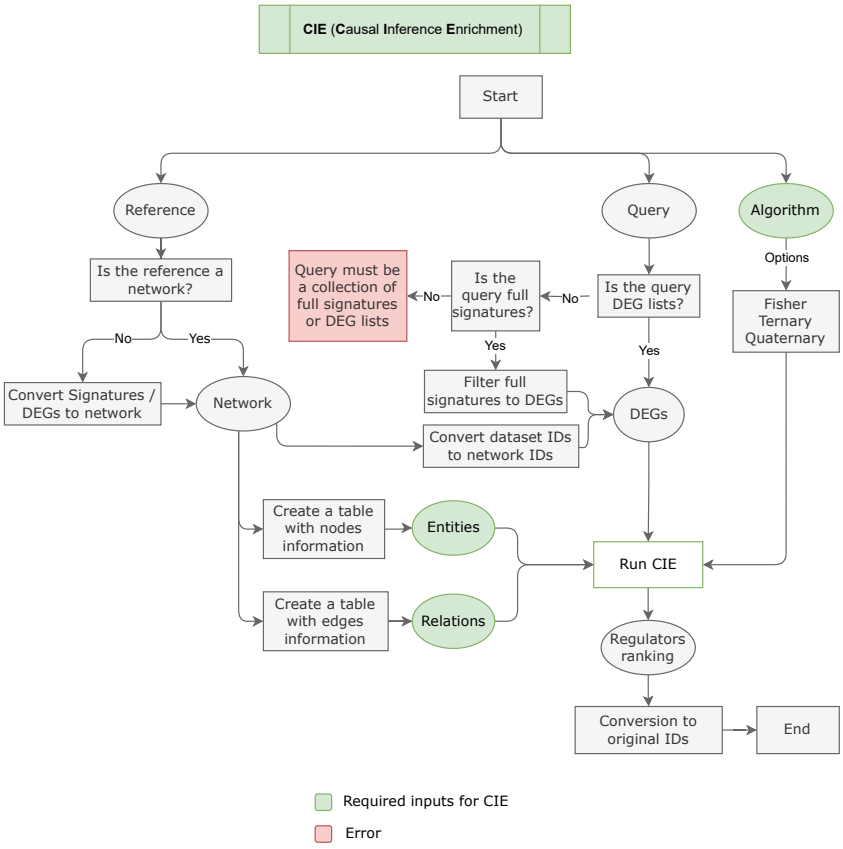

Figure 3.7: Flowchart representing the main steps of CIE algorithm implementation. The general computational pipeline for executing this topology-based method, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicates required inputs, while red highlights potential errors.

As the other wrappers, also the one implemented for NicheNet **[RN42]** (Figure **??**) initially ensure that the input formats are correct by converting the input signatures to DEGs and reference data to networks. From the reference network, three sub-networks

are created based on the regulatory mechanism: one ligand-receptor network and one signaling network from the signaling interactions (mechanism = 0) and one transcriptional network from gene regulatory interactions (mechanism = 1). Ligand-receptor and signaling networks are created based on an optional regulator's vector, providing a list of source nodes to be used for filtering signaling interactions. If this the filtering vector is not provided, only unique source nodes are used to create the set of regulators. The ligand-receptor network includes only edges where the source node is a regulator, whereas the remaining edges are presented in the signaling network, with the downstream signaling cascades down to the target genes. In the next step, a weighted network is constructed (construct _ weighted _ networks function from nichenetr R package [**RN42**]), by merging the three subnetworks into one graph structure and assigning each edge a baseline weight (1 for both gene regulatory and ligand-receptor network edges). To down-weight the influence of highly connected nodes, hub correction is applied to the network (apply _ hub _ corrections). With the weighted network, a ligand-target matrix is built (construct _ ligand _ target _ matrix) containing target genes as rows, ligands as columns and scores for each entry (inferred signaling strength), from each ligand to each gene. At this point, the inputs required to run NicheNet core function (predict _ ligand _ activities) are ready: a vector of DEG, the full set of genes present in the ligand-target matrix (used as the background), the ligand-target matrix of regulatory scores and the list of ligands (regulators) to prioritize. The output from NicheNet is ranked by corrected Area Under the Precision-Recall Curve (AUPR) scores and identifiers are converted back if needed.

The implementation of the CBDD wrapper functions (CBDD baseline and CBDD causal reasoning) follows an identical structure, with both functions sharing the same core preprocessing pipeline and output handling. For this reason, they are both represented in a single schema, with the differences highlighted (Figure **??**). Both starts with validation and preprocessing of input data, converting query datasets to DEG lists, handling reference networks as described previously and performing ID conversion to map query genes/regulators to network-specific identifiers. The wrapper functions then execute their respective algorithms, and the results are returned after ID conversion. Despite the similarities, there are differences between the implementations that reflect their distinct computational requirements. The CBDD baseline function supports randomWalk, overconnectivity, interconnectivity, networkPropagation, and hiddenNodes algorithms, while CBDD causal reasoning runs causalReasoning, SigNet, and quaternaryProd. The causal reasoning wrapper requires network attributes such as the effect and mechanism, reflecting the need for directed, signed edges for causal inference. Moreover, CBDD baseline converts networks to igraph objects, while CBDD causal reasoning creates a causal graph. Essentially, both wrappers return regulator scores, but since CBDD causal reasoning considers directionality, the results contain additional information about the effect of the relationships (predicted activation or inhibition).

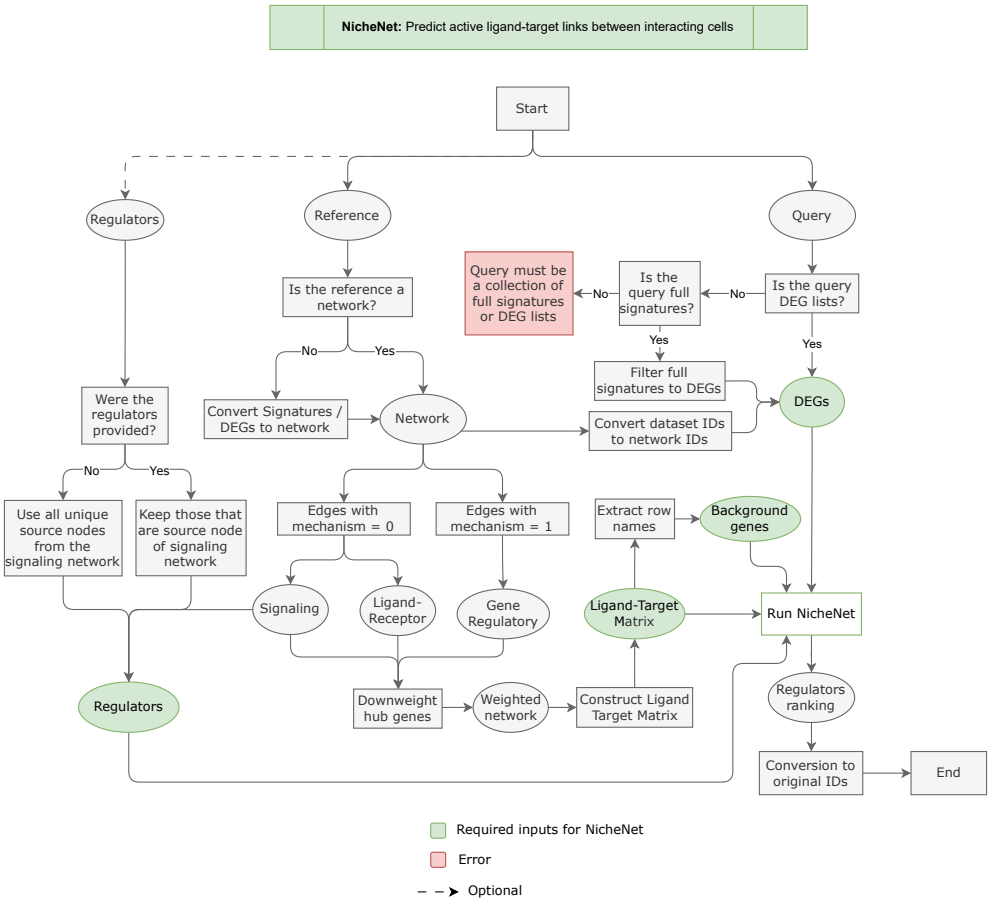

Figure 3.8: Flowchart representing the main steps of NicheNet algorithm implementation. The general computational pipeline for executing this topology-based method, showing the main input requirements, data preprocessing steps, algorithm execution, and output generation. Green indicates required inputs, while red highlights potential errors.

3.4 Evaluation

The final part of the study pipeline involves evaluating the algorithms by comparing the outputs from the algorithms' runs with the golden standard data. This evaluation process involves three main steps: configuration, running the algorithm, and evaluating performance against the standards. The configuration step includes selecting the input data (both the query and reference), as well as the standard with the verified list of the targets. Then, the execution requires defining the algorithm's parameters. Finally, metrics and methods are defined to assess performance of each algorithm, considering the known targets provided by the golden standard data.

The evaluation was performed at the level of each signature, aggregating results to obtain an average assessment by dataset. In general, the algorithms return all the ranked and scored regulators, so the threshold for considering which regulators are significant was also defined in two ways. If an algorithm provides p-values for the regulators, the

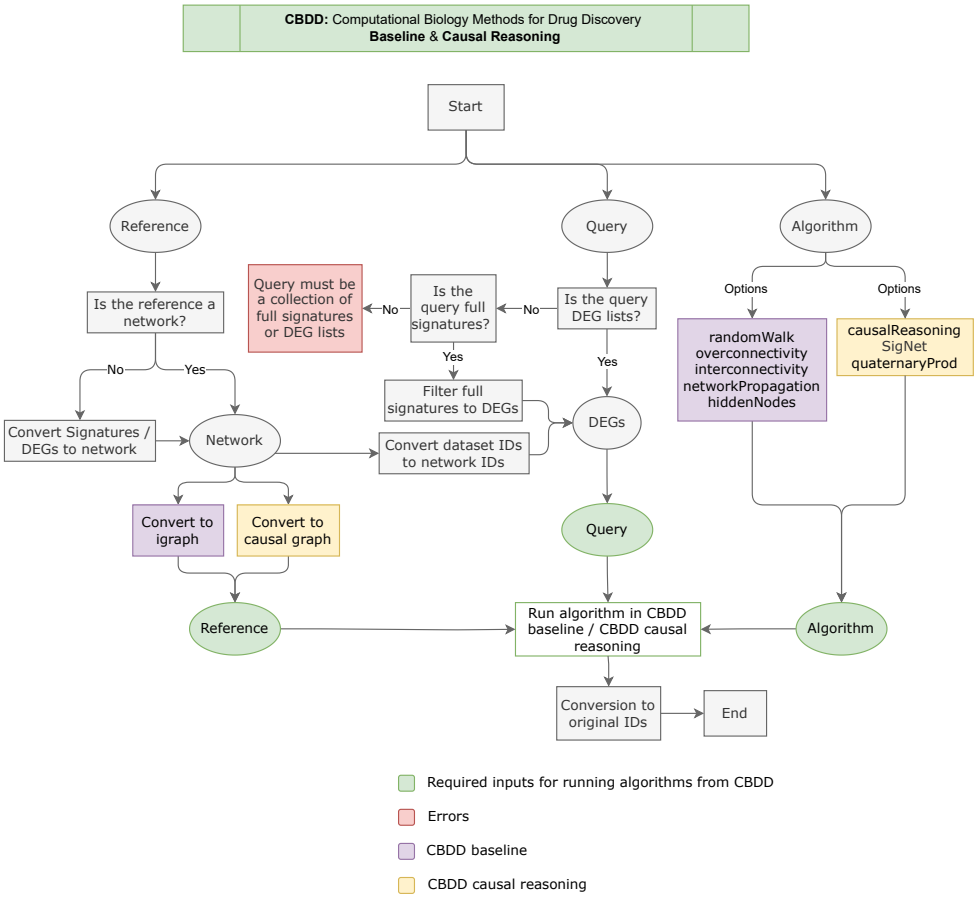

Figure 3.9: Flowchart representing the main steps of the wrapper function of both the CBDD baseline and CBDD causal reasoning algorithms implementation. General computation pipeline and color coding as previously.

threshold of 0.05 is used to consider that regulator as significant (if p-value > 0.05). If this is not the case, the top 1

The performance was measured at a target and pathway level. The former considers target recovery as the percentage of golden standard targets among all significant regulators, top 1

1. Overall recovery: average of target recovery percentages, enrichment scores, and scaled AUPR (for the top 1 %) across all signatures.

2. Recoverable signatures only: target recovery percentage calculated only on the subset of signatures whose true targets appear in the reference network.

3. Win rate: to assess each algorithm's ability to identify actual target genes, the win rate represents how frequently a given output ranking contains a true target higher than all other competing algorithms. The win percentage above expected (WPAE) is obtained deducting the expected win rate under random chance (1 divided by the

number of algorithms), as the number of competing algorithms can change between runs.

The assessment of performance at the pathway level is equally important to understand whether the algorithm not only recovers the target itself, but also the biological pathways to which these targets are connected. From the selected significant regulators, an enrichment analysis of the biological pathways was performed using the CBDD enr function, with MetaBase mapping gene sets to the pathways that are most associated with the targets predicted by the algorithms. The statistical significance of the overlap between biological pathways identified from the targets and those manually annotated via Metabase was calculated with a hypergeometric test (hh.pvalue function from CBDD). The result was recorded as a pathway enrichment score (-log10(p-value)), where a higher score indicates that the algorithm is able to identify pathways directly affected by the perturbation. For comparing the results, the mean pathway enrichment score for each algorithm was computed by averaging those signatures whose true targets appear in at least one MetaBase pathway.

4

Results

In the following section, the results of the benchmarking study are presented across three main levels: the algorithms, the reference and the query dataset. For each level, the metrics to assess the performance are analyzed to capture both computational feasibility and biological relevance. A final ranking combines all the scores, aiming to identify the best algorithm providing real-world applicability at each dimension of data.

A full evaluation was conducted with 86 runs (Table **??**), resulting from the combination of 10 query datasets and 9 reference datasets (excluding self-comparisons). Given the extensive number of individual runs, the results will be presented as performance summaries of the four key dimensions: algorithms, datasets, reference, and target identification performance. Table **??** presents the runs that were evaluated, detailing which combination of query and reference datasets was used, and some characteristics of each query dataset.

4.1 Algorithms

To evaluate the performance of the 27 algorithms, several metrics were used describe the output generated from each package. One is the practical scalability of each tool, and the other is the biological accuracy of the results. To address the first aspect, the scalability (runtime) and the reliability were measured. To evaluate the biological relevance of the predictions, the AUC, the win rate, and the pathway enrichment scores were measured.

Table 4.1: Summary of evaluated runs. Tool category: C = Connectivity mapping tools, N = Network tools, E = Enrichment tools; Query dataset characteristics: D = Drug perturbation, G = Gene perturbation, C = Cell lines, T = Tissues/in vivo models, Sc = Single-cell data. *: Selected impact signatures from 7390 signatures, **: Generated consensus signatures from 82256 signatures.

Query / Reference	Perturbation	Origin	Signatures	MetaBase	OmniPath	MetaBase (linear path)	MetaBase (regulons)	OmniPath (regulons)	LINCS CRISPR	LINCS OE	LINCS shRNA	GWPS Perturb-Seq
CDS-DB	D	T	181	N	N	NE	NE	NE	CNE	CNE	CNE	CNE
Sci-Plex	D	C (Sc)	405	N	N	NE	NE	NE	CNE	CNE	CNE	CNE
PertOrg	G	T	951*	N	N	NE	NE	NE	CNE	CNE	CNE	CNE
ChemPert	D	C	1304**	N	N	N	N	N	CN	CN	CN	CN
GWPS	G	C (Sc)	1980	N	N	NE	NE	NE	CNE	CNE	CNE	-
CREEDS	DG	T	2642	N	N	N	N	N	CN	CN	CN	CN
LINCS Compounds	D	C	3540	N	N	NE	NE	NE	CNE	CNE	CNE	CNE
LINCS OE	G	C	3780	N	N	NE	NE	NE	CNE	-	CNE	CNE
LINCS shRNA	G	C	4854	N	N	NE	NE	NE	CNE	CNE	-	CNE
LINCS CRISPR	G	C	5156	N	N	NE	NE	NE	-	CNE	CNE	CNE

Execution time and algorithm failures can prevent the use of algorithms in research settings. For this reason, evaluating those metrics is important for understanding the feasibility of each computational approach. The runtime represents the time in seconds that each tool took to execute, and it is directly correlated with query and reference datasets size. In this context, a mean runtime captures how long on average each tool takes to generate the desired output. Whilst this value can be informative, a better approach to present these data is to consider mean runtime per signature (Figure 13) or a value normalized considering the size of the reference dataset (the final runtime score in Figure 17). The final runtime score was calculated as follows. First, within each reference, each tool's per-signature times were rescaled from 0 (slower) to 1 (faster). Then, those values were averaged across all references and inverted, so that faster algorithms get higher scores. As shown in Figure 17, the majority of topology and enrichment methods topped the rankings with scores greater than 0.90, except for the enrichment methods udt and viper, which had lower scores (0.63 and 0.39, respectively). The topology methods, causal reasoning, interconnectivity, and network propagation, along with the enrichment methods ulm, wmean, and wsum, topped the rank with scores near 1, providing final output on average in 1-2s per signature. Connectivity-mapping tools stayed in the middle of the ranking with scores ranging from 0.80 to 0.90, and mean runtimes 13- 24s per signature. In contrast, methods like NicheNet and ProTINA (respectively 416s and 126s per signature) were markedly slower, thus obtaining a score close to 0.

To complement the runtime evaluation, a reliability score was also computed to assess how often each algorithm failed. The final reliability score is an average of a success rate (runs that returned results) and a run rate (runs launched) from a scale of 0 to 1, where scores close to 1 indicate more reliable algorithms. The rates can be found in Table 7, while the final reliability scores are represented in Figure 17. At the top of the rank, some of the topology and enrichment methods, with causalReasoning, randomWalk, and ulm topped scoring close to 0.99 (i.e. returning results in 98

5

Discussion

This chapter discusses the results by highlighting their implications and comparing them with the existing literature. Finally, the key findings are summarized into conclusions and suggestions for some future perspectives.

The MoA of a small molecule involves its interaction with specific molecular targets. When these primary targets are activated or inhibited, they trigger changes in downstream signaling pathways, modulating the activity of transcription factors and therefore the expression of individual genes. All these changes will ultimately produce the desired drug effect together with any additional undesired side effects. As with many topics in the scientific community, there are different views on the real need of a complete understanding of the molecular target and the MoA of a drug [**RN112**]. Some defend and prioritize clinical benefit, given the high number of drugs on the market for which the MoA is currently unknown. On the other hand, some argue that understanding the effects of compounds is essential in the early stages of DD. Knowing the MoA of a drug not only helps to guide the process but also increases predictability and can even help to understand potential side effects. While understanding these mechanisms isn't strictly necessary for successful drug development, it plays a crucial role in improving efficiency. For this reason, advances in system biology are focused on reducing the process for MoA reconstruction, keeping reliable results. For example, in 2020, the Connectivity Mapping Institute created a challenge on the Kaggle platform [**RN162**] to develop an algorithm to predict the MoA of a new drug. Along with these collective efforts, several bioinformatics tools have emerged to infer causal regulators. Authors who develop a new algorithm typically perform comparative benchmarking to highlight the novelty of their discovery [**RN109**]. However, these studies have often some limitations. The main one is that simulated data, such as the LINCS dataset, is used instead of real experimental data. This approach may fail to capture the variability and complexity of biological systems, leading to a biased evaluation of the computation tool. Benchmarking studies are therefore essential because they aim to test various components in an unbiased and realistic way, to evaluate the performance of different classes of algorithms.

This study represents the most comprehensive evaluation of MoA recovery methods

published to date. While previous studies have focused on specific categories of methods, limited network sources, or small selected datasets, this study covered 27 algorithms, including topology-based, enrichment, and connectivity mapping methods, incorporating public and commercial reference datasets, and experimental and simulated data from seven distinct sources. This extends previous benchmarking efforts and shows critical differences between theory and practical feasibility.

Topology-based methods consistently achieved the best performance across all metrics in direct target retrieval. randomWalk was almost always at the top of the ranking for AUC, reliability, and win rate, as well as run times with around 5s per signature. These results are consistent with other studies. In Hill et al. [**RN37**], node prioritization methods, including randomWalk followed by networkPropagation and GeneMANIA, were also identified as the top performers, using a combination of interactions from STRING, MetaBase, and BioPlex as reference dataset. Our study confirmed the superiority of randomWalk, together with evidences that performance depends on the choice of reference network, with the MetaBase network providing the best results, probably due to its greater molecular coverage.

The similar performance of randomWalk with different reference datasets demonstrate the robustness and reliability of the algorithm. The overconnectivity and networkPropagation algorithms also performed well in our study. However, each one had some limitations in terms of accuracy, reliability, or run time. For example, networkPropagation obtained results similar to randomWalk in terms of AUC, but with reduced reliability in large-scale executions, suggesting possible concerns with scalability.

Our study reveals an important trade-off between direct recovery and the biological contex, not previously characterized in other studies. Methods that achieved the highest pathway enrichment scores, such as wmean and wsum, also obtained negative WPAE values, indicating that they do not outperform direct target classification methods. On the other hand, although randomWalk performs well in target identification, it seems to compromise pathway enrichment performance, by obtaining half the scores of wmean and wsum. This inverse relationship suggests that algorithms designed to capture biological context end up sacrificing accuracy in identifying individual causal nodes. As a consequence, methods selection goes beyond simple performance rankings, as it should also account the need to a deeper understanding of the biological context surrounding a specific target. If understanding the broader biological context is more important than identifying single targets, these enrichment methods can provide better results, despite lower accuracy. In contrast, when experimental validation resources limit testing to a few candidates, the accuracy of randomWalk seems to be more appropriate.

A study by Lin, K., et al [**RN79**] and others [108] identified ZhangScore as the best algorithm among connectivity mapping methods. However, in our study, this algorithm performed poorly. In general, all CMap algorithm achieved low AUC values (< 0.08), win rate values (< 0.03), and pathway enrichment scores ($0.20 – 0.31$). Among the CMap

algorithms, GSEAweight1 showed the best results, followed by KS, with Zhang only distinguishing itself because of the worst computational efficiency, having the highest runtime. This huge difference compared with published results can be interpreted considering the task being evaluated. While the studies mentioned above were assessing the similarity between signatures, in this study we are evaluating the ability to identify the true target responsible for the perturbation signatures. These results suggest that similarity-based algorithm fails to capture the causal regulatory relationships underlying perturbation responses, meaning that gene expression pattern relationship does not necessarily reflects shared regulatory mechanism.

Consistent with Hill et al [27], we observed that causal reasoning methods performance is better at pathway-level enrichment, compared with precise target identification. For example, causalReasoning achieved perfect reliability, but modest AUC, in line with the observation that these methods struggle to pinpoint exact drug targets. This pattern was observed for SigNet and CausalR [27] and it can reflect that many regulatory targets are not TFs or are not directly observable in expression data. The most interesting and surprising finding from this study is the issue of scalability for causal reasoning algorithms, which resulted in complete failure of the most sophisticated among them (CARNIVAL). In previous benchmarking studies, SigNet and CARNIVAL had the best performance [27], with OmniPath network as reference data. However, despite CARNIVAL's theoretical ability for capturing complex regulatory interactions, its computational requirements make it unusable for realistic dataset sizes. This emphasizes that algorithm benchmarking must consider not just accuracy but also scalability and runtimes, that are often neglected when evaluations focus on small and curated datasets. In this large-scale evaluation several computationally intensive algorithms (e.g., NicheNet, ProTINA) are impractical for high-throughput use, with runtimes exceeding 400 seconds per signature or total failure to complete the analyses. For methods intended for integration into large-scale screening pipelines, these constraints represent a critical limitation.

With a strong influence on algorithm performance, the importance of network choice was no exception in our study. Metabase demonstrated the most consistent target recovery results, while others appear better suited for pathway enrichment. Network selection should be aligned with the research goal, with denser and high-coverage networks for precision target recovery and more structured, curated pathway resources for context mapping.

The consistent superiority of drug perturbations over genetic approaches is important to highlight. The assumption that genetic perturbations are more precise can be refuted by our results, where well-designed drug perturbations seem to provide better results. Also, the poor performance of LINCS genetic perturbations highlights potential limitations of cell line-based genetic screens.

The validation approaches used by the original dataset creators varied considerably, affecting result interpretation, as a consequence the quality of the datasets can be a determining factor in the performance of these tools. Among the 7 sources of datasets

used, based on the validation carried out by the authors, some datasets do not seem to be the most suitable for obtaining reliable results. CREEDS is one such case where the data is extracted from public databases and not generated by the researchers involved in the benchmarking effort [**RN87**]. The data is only checked on a technical level, to confirm that the samples are from GEO, and if they are labeled correctly. The attempt at biological evaluation, by looking for specific patterns of signatures (i.e. if two signatures come from perturbing the same gene), showed inconsistent results. Moreover, for some datasets validation was practically non-existent [**RN85**, **RN86**], and the filters used during the pre-processing step of this study actually increased the quality of the data (this is the case of ChemPert and PertOrg). Both sources for the single-cell datasets showed a good validation of the data. GWPS, focused on gene knockdown signatures, used strict criteria to define significant responses [**RN89**]. Sci-Plex, with drug perturbations, validated its data through some complementary statistical analysis [**RN88**]. The data from LINCS are more controversial. First, L1000 measures around 978 landmark genes and computationally infers the rest (12,000). Additionally, despite each perturbation is tested in triplicate and z-scores are adjusted to minimize replicate inconsistencies [**RN30**], the reliability of these data is still questionable. Recent alternatives, such as DRUG-seq, demonstrate to have some advantages over L1000. Direct comparison between the two datasets [**RN130**, **RN105**] showed that DRUG-seq directly measures more than 10,000 genes without relying on inference, at a lower cost. When testing the accuracy of both, DRUG-Seq proved to be more accurate in distinguishing samples between different diseases. Also, the emergence of Perturb-seq, which combines CRISPR perturbations with single-cell RNA-seq [**RN89**], offers another promising alternative with greater transcriptome coverage at a reduced cost [**RN165**]. Finally, the cancer cell line composition is another shortcoming of this dataset limiting applicability to other contexts. This is being addressed by programs such as NeuroLINCS [**RN164**], which create signatures using patient-derived induced pluripotent stem cells [**RN165**]. Our finding that tissue-derived signatures consistently outperformed cell line data supports the expansion of data derived from other sources.

The critical importance of reference network selection extends beyond simple coverage metrics. Large-scale networks may include interactions irrelevant to specific cellular contexts [**RN38**]. For this reason, tools for constraining networks based on tissue-specific expression data from Human Protein Atlas [**RN166**] or TissueNet [**RN137**] could improve performance [**RN38**]. Our results showing MetaBase Linear Path Regulons' excellent pathway enrichment, despite poor coverage, support this tissue-specific approach.

The focus on transcriptomic-based methods, in this thesis, but also other benchmarking studies, represents just one facet of MoA inference. Recent multi-OMICS integration tools like SignalingProfiler 2.0 [**RN100**] and COSMOS [**RN99**] demonstrate the value of combining transcriptomics with proteomics, metabolomics, and phosphoproteomics data. COSMOS's successful application to capture relevant crosstalks within and between multiple omics layers, resulting in identification of known clear cell renal cell carcinoma drug targets, illustrates how the integration of multiple data types can generate novel

biomarker hypotheses. Additionally, data capturing changes in cell morphology and the chemical structure of specific compounds [**RN167**] can become a valuable complement to expression data in DD studies [**RN38**].

Some considerations and future work that can be explored are related with certain adjustments or experimental directions that could have been considered, but have been left out of this study for time and resource limitations. One aspect that has not been considered, but could effectively influence the performance of the algorithms is parameter optimization. Algorithm behavior and performance can change dramatically with parameter tuning, and it is suggested as good practice for benchmarking studies [**RN108**]. Yet comprehensive optimization across all algorithm-dataset combinations proved computationally unreasonable. The same applies to the methods for filtering DEGs from full signatures. However, the parameters of the algorithms and the preparation of the input data are specified and calculated in the wrapper function and they can be changed if needed. Being this a benchmarking study, the entire workflow is set up precisely to test multiple options, even though not all of them have been included in this work. When evaluating, another metric that could be included was an ensemble score. Instead of relying on a single tool or data set, a score combining the results of several tools could also be generated to produce more robust predictions. Future benchmarking efforts can also weight results by validation confidence, given that the heterogeneity in dataset validation can partially explain some performance variations across query datasets.

The molecular networks used as a reference in this study, both MetaBase and OmniPath, are of considerable size, with hundreds of thousands of interactions. When using these networks, one of the recommendations is to ensure that the interactions present are within the context of the study, in this case, the type of cell or tissue being studied [**RN38**]. This reasoning makes sense if we consider that these large networks include all kinds of interactions, including those specific to other cells or tissues. To this end, instead of using global networks, more specific and context-relevant networks can be used, which may already be prepared by the databases themselves. Alternatively, databases such as TissueNet [**RN137**], provide interaction networks that are specific to certain tissues. Another thing that could be considered would be to integrate networks from different sources, for example, MetaBase and Network, customizing the subsets of interactions to suit the context and by keeping only those with higher reliability scores.

If we do not just focus on this type of algorithm, there are more software and tools that can be used for target identification, that leverage different computational approaches. Depending on the specific scientific questions being investigated, it could be also worth exploring them. One example is Drug2ways [**RN132**], a software implemented in Python, to identify potential drug repositioning and predict the effects of drugs. To do this, it performs causal reasoning through biological networks with three types of vertices: molecular entities, drugs, and indications. The paths between the drug and the indication are noted, as well as their direction, suggesting positive or negative regulation. The most frequent direction is taken as the predicted effect. Solutions are becoming increasingly

innovative, taking advantage of advances in computer efficiency. Such as the use of knowledge graphs and Large language models for drug repurposing **[RN163]**.

In summary, based on the results obtained and the intersection of previous research, this study recommends the randomWalk algorithm for precise target recovery, as well as wmean for pathway-level context. It is important to emphasize the importance of high-quality reference data, and for topology-based tools, to consider filtering interactions relevant to the study in question. Regarding query data, the use of experimental data and data from tissues should be preferred. In addition, other algorithm optimization parameters should be considered, rather than the default settings. Finally, given the progresses in the field of system biology and the avalanche of increasingly available data, integration of other levels of OMICS data and the use of machine learning methods promise to further improve the ability to identify target genes and perturbed pathways.

Molecular network Concordance Score (CS)

A

*NOVA*THESIS COVERS SHOWCASE

Text
Text

A.1 A section here

Text Text

B

Appendix 2 Lorem Ipsum

Text

Text

Text

I

Annex 1 Lorem Ipsum

Text

UNIVERSIDADE NOVA
DE LISBOA