UNIVERSIDADE NOVA DE LISBOA

itqb nova

NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

NOVA
MEDICAL SCHOOL

NOVA
IMS
Information
Management
School

SANITATEM QUAERENS IN TROPICOS

DESDE
INSTI
MEI
UNIVER

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
SPECIALIZATION MULTI-OMICS FOR LIFE AND HEALTH SCIENCES

NOVA University Lisbon
September, 2025

Benchmarking Computational Algorithms for Enhanced Drug Discovery
Insights from Clarivate's pre-competitive Algorithm Benchmarking Consortium

Copyright © Maria Inês Nunes Vilar Gomes, NOVA School of Science and Technology, NOVA University Lisbon.
The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

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

Dedicatory lorem ipsum.

Acknowledgements

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

> *“You cannot teach a man anything; you can only help him discover it in himself.”*
>
> — **Galileo**, Somewhere in a book or speach
> (Astronomer, physicist and engineer)

Abstract

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

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

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

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

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

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

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

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

Resumo

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

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

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

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

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

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

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

Contents

List of Figures

Acronyms

ABC	Algorithm Benchmarking Consortium *3, 4)*
CARNIVAL	CAusal Reasoning for Network identification using Integer VALue programming *2)*
CBDD	Computational Biology for Drug Discovery *2, 32)*
CDS-DB	Cancer Drug-induced Gene Expression Signature Database *9, 10)*
CIE	Causal Inference Engine *2)*
CMap	Connectivity Mapping *2, 8)*
CRC	Colorectal Cancer *6)*
DD	Drug Discovery *1, 5, 6, 8)*
DEGs	Differentially Expressed Genes *2)*
DNA	Deoxyribonucleic Acid *8)*
DR	Drug Repositioning *5)*
FDA	Food and Drug Administration *6)*
GEO	Gene Expression Omnibus *9, 11, 12, 15)*
HTS	High-Throughput Screening *1)*
LINCS	Library of Integrated Network-Based Cellular Signatures *8, 10, 12)*
MoA	Mechanism of Action *1, 6–9)*
OE	Overexpression *9)*
R	R Programming Language *2, 4, 22)*
RCSM	Recommended Connectivity-map Scoring Methods *3)*

shRNA	short hairpin RNA *9)*
TF	Transcription Factor *2, 8)*

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 [0, 0, 0]. As thorough understanding of the 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 [0].

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 [0], 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 [0].

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 [0] 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 [0, 0]. 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 [0].

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 [0] , 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 [0].

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 [0]. 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 [0, 0, 0]. 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 [0]. 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 [0, 0]. 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 [0], 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 [0]. 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 [0, 0, 0]. Another classic example is Thalidomide, originally used for sedation and morning sickness, and afterward repurposed for multiple myeloma, leprosy [0, 0], and to minimize the hippocampal neuronal loss [0]. 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 [0, 0]. Noteworthy cases include the schizophrenia drug Spiperone, which has been studied for its ability to induce apoptosis in Colorectal Cancer (CRC) cells [0], and Raloxifene, indicated for osteoporosis, which proved to be effective in reducing breast cancer risk in postmenopausal women [0, 0].

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 [0, 0]. 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 [0, 0]. 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) [0, 0].

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 [0], forming a complex network of interconnected and distinct nodes. The impact of a certain compound in the complex cell signaling cascade can be defined

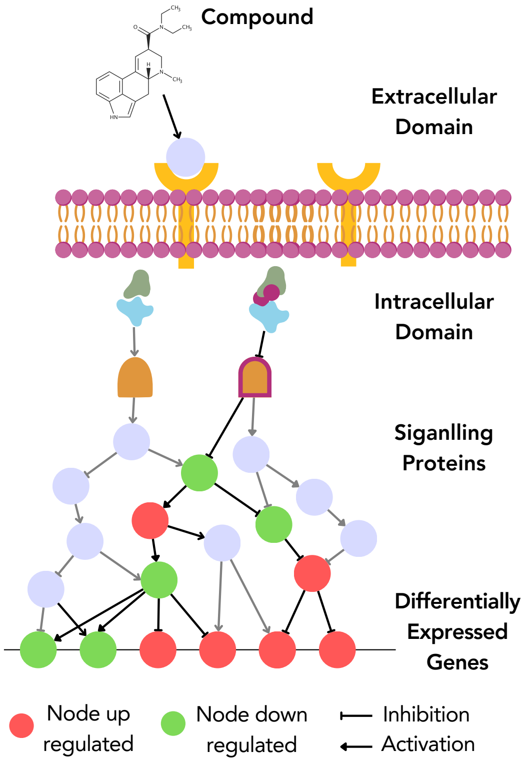

Figure 2.1: Schematic representation of a compound-induced cellular signaling cascade, where binding of the perturbing compound to its extracellular receptor domain triggers downstream intracellular signal transduction events through various signaling proteins, ultimately altering gene expression levels 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.

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 [0]. 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 Transcriptomics 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 [0]. 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 [0, 0, 0]. 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 [0, 0, 0]. 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 [0]. 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 [0]. 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 [0].

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) [0]. 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 [0]. 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 [0]. In a preliminary phase, LINCS released over 6000 signatures from around 1300 small molecules, many of which were FDA-approved [0]. 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 [0]. 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 [0]. 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) [0] 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 [0]. 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 [0].

ChemPert [0], 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 [0]. 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 [0]. Although traditional single-cell RNA sequencing (scRNA-seq) is essential for analyzing tissue heterogeneity, its high cost remains a barrier. Sci-Plex [0] 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 [0, 0]. 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 transcriptomics 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 [0] and Open TG-GATES [0].	[0]
CDS-DB	Cancer Drug-induced gene expression Signature DataBase (cdsdb) 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 [0] and ArrayExpress [0] and supports data browsing, searching, and single signature analysis.	[0]
CEBS	Chemical Effects in Biological Systems is a publicly accessible repository for toxicogenomic data (cebs), 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 [0].	[0]
ChemPert	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.	[0]

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
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 [0] (ilincs.org) provides an interactive platform to explore these datasets and the relationships between compounds, gene expression changes, and disease.	[9, 40]
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 [0]. 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.	[0]
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.	[36, 37]

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
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).	[52, 53]
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.	[0]
GWPS	Genome-Wide Perturb-Seq (gwps.wi.mit.edu) 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.	[0]

Continued on next page

Table 2.1 – *Continued from previous page*

Database	Description	Ref.
Open TG-GATEs	Toxicogenomic Project - Genomics Assisted Toxicity Evaluation System (open-tggates) 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.	[0]
PertOrg	PertOrg 1.0 (inbirg.com/pertorg) [51] 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.	[0]
PerturBase	PerturBase (perturbase.cn) 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	[0]

Continued on next page

Table 2.1 – *Continued from previous page*

Database	**Description**	**Ref.**
PerturbAtlas	PerturbAtlas (perturbatlas.kratoss) 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.	[0]
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).	[0]
STARGEO	The Search Tag Analyze Resource for GEO (stargeo.org) is an open resource constructed using GEO's publicly available functional genomics data [48]. STARGEO platform provides 3031859 reliable and annotated samples of gene signatures from humans, mice, and rats.	[0]
ToxicoDB	ToxicoDB (toxicodb.ca) integrates and harmonizes diverse in vitro toxicogenomic datasets from three sources (Open TG-GATEs [0], DrugMatrix [46], and ArrayExpress [47]). The aim is to easily perform queries and to summarize the relationships between gene expression and toxicant effects [59]. 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.	[0]

2.3 Prior knowledge Network

Understanding the physical molecular interactions within a biological system is crucial for contextualizing experimental data. Computational methods for elucidating the mechanisms of action allow the integration of omics data with prior knowledge of the interactions between biological entities [28]. These interactions can be represented with more or less complexity and can be included in the analysis as supplementary data sources. A prior knowledge network (PKN) is a collection of interactions where nodes represent molecular entities (such as proteins, genes, or metabolites) and edges illustrate their relationships. Understanding causal graphs is key to modeling and interpreting these networks, as they depict cause-and-effect relationships. In such graphs, nodes represent variables, while directed edges represent causal influences, indicating that a change in one variable affects another [60]. Furthermore, edges can be signed, indicating whether a causal node employs a positive or negative effect on the second variable, and weighted, to show the connection strength [60]. In causal graphs that model biological networks, multi-edge connections are common, with two or more edges linked to the same node. Networks can be classified based on interaction types and node characteristics. Protein-protein interaction (PPI) networks show direct interactions between proteins. Gene Regulatory Networks (GRN) Figure **??** illustrate how transcription factors influence gene expression [1]. Signal transduction networks describe how cells process external signals. Metabolic networks display relationships between enzymes and metabolites. Furthermore, networks are not always composed of molecular entities, as is the case with the disease network model, which links diseases using genes and mutations as connections. These networks fit experimental data to predictions from causal graphs describing the system. The choice of the PKN should match the data type. For instance, for transcriptomic data, integrating a Protein-Gene Regulatory Network (PGN) may be beneficial, while for metabolomic data, metabolic networks are more suitable. Researchers have made significant efforts to construct regulatory networks. A primordial example was the functional characterization of yeast genes through PPI analysis. This study aimed to show the guilt-by-association principle, by inferring an unknown protein's function by looking at its interactions with nearby entities [60, 61]. The guilt-by-association principle is a foundational concept in biological networks. It suggests that genes with similar functions often interact with the same proteins or have similar expression patterns [62]. This principle also applies to drugs that cause similar transcriptional responses and may have comparable mechanisms of action (MoA) [18]. Biological interactions can be described with different levels of complexity. A network is an intricate representation of the global interactome, linking all entities in the system. These interactions are primarily established in published experimental research and can vary in the amount of associated information. Based on supporting studies, each interaction may include details about direction, signal, and confidence level. This helps filter data, creating a network with more reliable relationships. Still, networks can be noisy and incomplete, with high rates of false positives and negatives and a tendency for well-researched entities to become overrepresented [28, 63, 64]. In contrast, a pathway is a simpler version of a network. It illustrates a series of molecular interactions that begin with one entity and follow a specific signaling cascade. This arrangement helps classify entities by their common biological roles. Yet, pathways often miss crosstalk between other pathways and provide a static view of a dynamic process [28]. The entities' overrepresentation issue also applies to pathways. Another way to show interactions is, for example, through regulons. Regulons are groups of co-regulated genes controlled by a common transcription factor and are usually represented as GRN (Figure 2). The choice of network type must always be appropriate to the scientific question and the type of data with which the network is used. The interactions between biological entities and the complexity of these interactions should be considered. For example, if pathways are used instead of a full network, they might miss some interactions and changes over time. If the study targets a specific cell type or tissue, it's important to use tissue-specific networks. Databases like TissueNet [65] can provide molecular interactions specific to a particular cellular context. The use of large-scale causal graphs for gene expression data interpretation was first introduced by Pollard et al. [64, 66]. This study aimed to infer the molecular causes of the changes in oxidative phosphorylation gene expression in skeletal muscle from type 2 diabetes (DM2) patients. For this purpose, the gene expression data were integrated with a large-scale model created from over 210,000 molecular relationships based on the DM2 literature. Computer-aided causal reasoning on these complementary data identified that the observed changes are linked to decreased glucose transport, impaired insulin signaling, and increased risk of post-transplant diabetes [66]. Given the good results obtained from supplementing the studies with PKN, the identification of interactions began to receive more attention. The development of high-throughput screening techniques such as yeast two-hybrid screening and DNA microarray [67] allowed the detection of PPI. Those interactions began to be deposited in databases that provide molecular interaction data. Nowadays, there are several public and commercial network and pathway resources. Table 2 summarizes some of the main resources of biological pathways and networks. Two resources, public and commercial, that provide composite networks are OmniPath and MetaBase™, respectively. OmniPath [68] is a freely available resource of prior knowledge in molecular biology. It combines data from over 100 resources and builds five integrated databases with different types of data: Interactions (several molecular interactions organized into sub-networks), Post-Translational Modifications (enzyme-substrate reactions), Complexes (35,000+ protein complexes), Annotations (proteins and complexes annotations, such as the function, localization, tissue, etc.) and Intercell (inter-cellular signaling roles, such as, if a protein is a ligand, a receptor, an extracellular matrix component, etc.) [68]. The interactions database is a composite signaling network that offers several manually curated subnetworks, encompassing a total of 282,504 unique interactions. Each subnetwork has different types of interactions, including post-translational interactions, transcriptional interactions, post-transcriptional interactions, and other interactions involving small molecules. The number of interactions per subnetwork is described in Figure 3. One of the GRNS that is provided by this database is the CollecTRI-derived regulons [1] Figure **??**. This collection contains high-confidence signed transcription factor (TF) - target gene interactions. These interactions were compiled from 12 resources, including information inferred from text mining, manual curations, and several publicly available databases.

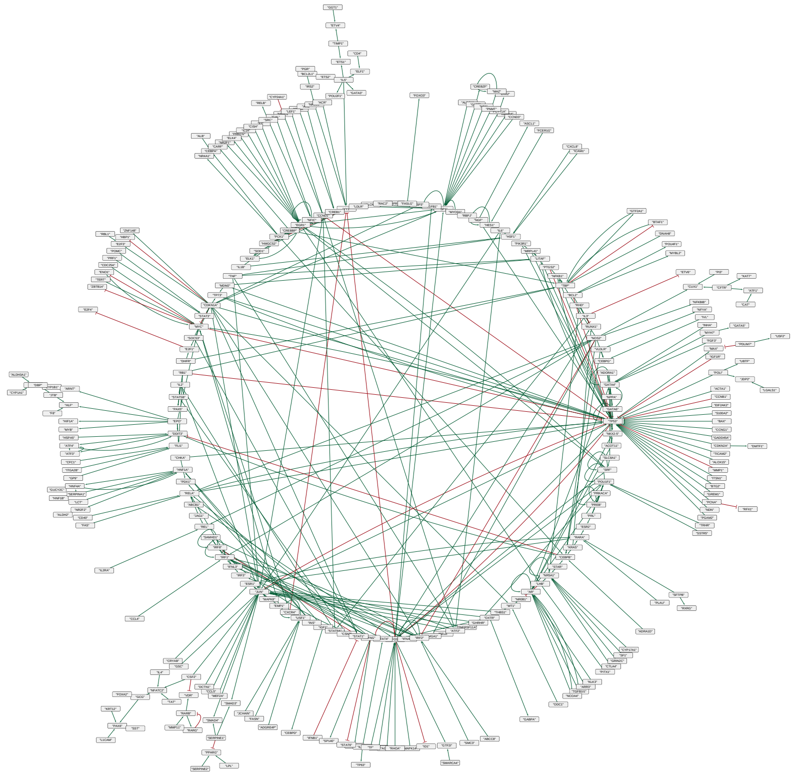

Figure 2.2: Gene regulatory network based on regulon representation of human transcriptional interactions. CollecTRI-derived regulons were extracted from the decoupleR (v. 2.12.0) package, which provided 43,178 interactions. CollecTRI collection [1] is a comprehensive, curated resource of transcription factors (TFs) and their target genes, expanding on DoRothEA. This figure illustrates a subset of those 1,000 interactions, with edge color indicating the mode of regulation (green for activation and red for repression) using RCy3 (v. 2.26.0) R package.

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

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

2.4 Computational methods for MoA inference

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

2.4.1 Similarity-based methods for comparative analysis

2.4.2 Enrichment-based tools for downstream analysis for downstream analysis

2.4.3 Topology-based methods for upstream analysis

2.5 Benchmarking of computational methods for MoA inference

The use of computational methods to elucidate mechanisms of action is becoming increasingly indispensable for integrating and interpreting the multitude of available data. Given the plethora of existing computational tools, choosing the appropriate data and methods to answer specific scientific questions can be challenging. When a new tool is developed and published, it is usually compared with popular existing methods. For those with less experience, distinguishing the benefits of a novel tool from others that may be equally advantageous but better suited for different applications or data types can be difficult. A comprehensive benchmarking study of the tools is crucial for evaluating available methods in a standardized way, providing sufficient information to accurately choose the best tools and data for a given study [88]. A key component of a benchmarking study is the use of gold-standard datasets, against which the results obtained from a method are compared. By comparing these results with the ground truth, it is possible to evaluate performance metrics and statistical analyses, consistently distinguishing different computational algorithms based on their behavior with certain types of data. It is widely acknowledged that evaluating the vast available methods is important for obtaining more accurate results. Therefore, following certain good practices when conducting a benchmarking study is essential. These studies can be carried out by the authors who implemented the tool, independent groups, or as organized challenges, such as those organized by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) Consortium [89]. When the authors do the evaluation, the aim is usually to demonstrate the advantages and performance improvements over other techniques. In other contexts, it is extremely important to define the scope and purpose of the benchmark. The selection of the methods should reflect the relevance of the study's objective and include publicly available implementations to ensure accessibility. Parameter optimization can significantly affect a tool's behavior, including runtime, yet finding the optimal values is not always straightforward. Thus, balancing default settings with computational efficiency is important. Regarding datasets, in the context of studying the compound's mechanism of action, it is crucial to include diverse data sources and generation methods to ensure representativeness and a credible assessment of performance. For instance, transcriptomic data, if making sense in the scope, should ideally include both bulk RNA-seq and single-cell RNA-seq data to broaden the options and use two widely used types of data. Since there are no perfect, fully curated datasets, it is necessary to ensure quality to avoid biasing the results and performance of the tools [89]. The same applies to the gold standard datasets, which serve as the ground truth and are fundamental for statistical analysis and assessing metrics, defining the essence of a benchmarking study. Some benchmarking studies arise from an effort to contextualize gene expression data with several computational algorithms. Hosseini-Gerami et al. [27] evaluated the performance of different causal reasoning algorithms to recover direct compound targets of small molecules and associated signaling pathways using gene expression data. The study compared four causal reasoning algorithms against networks from two different sources and transcriptomics data from one database. Hill et al. [60] conducted a study that provided a more comprehensive framework by analyzing a diverse range of algorithms, networks, and datasets to assess how well network-based algorithms prioritize and connect gene lists derived from transcriptomics data. This study integrated 17 algorithms, categorized into three main groups: (1) Node Prioritization Algorithms, which rank network nodes based on connectivity, (2) Causal Regulator Algorithms, which identify upstream regulators of gene expression changes, and (3) Subnetwork Identification Algorithms, which extract subnetworks linking input genes. The algorithms were applied to three PPI networks, each with different structures and levels of curation, using hundreds of datasets from four sources to cover scenarios where certain data types might be unavailable. The first network combined data from various sources, resulting in a mix of signed/unsigned and directed/undirected interactions. The second network included only signed, directed, and high-confidence interactions, while the third was a large-scale, undirected PPI network. This study exemplifies good practices while doing a comparative analysis, by including and integrating different resources, although exploring the parameter landscape for each algorithm was beyond the scope of this work, so it was not included. Typically, a final benchmarking analysis ranks the algorithms in terms of the most appropriate use for distinct applications, and so the choice of algorithm(s) may depend on the specific use case [60]. By providing a robust assessment of the capabilities of existing algorithms, these studies leverage knowledge and provide guidelines for researchers to choose which resources should be used in certain situations [88].

3

Materials and Methods

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

3.1 Benchmarking architecture setup

Several tools and algorithms are available for most research tasks in computational biology, and new algorithms and tools are published every week. Systematic benchmarking of tools is a time- and resource-consuming endeavor, while a lack of benchmarking carries several potential risks. Finding the right computational tool for a given research question is essential. Researchers usually carry out published benchmarking to demonstrate that their tool performs better than others. ABC is a consortium created in 2021 that aims to help members reduce 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].

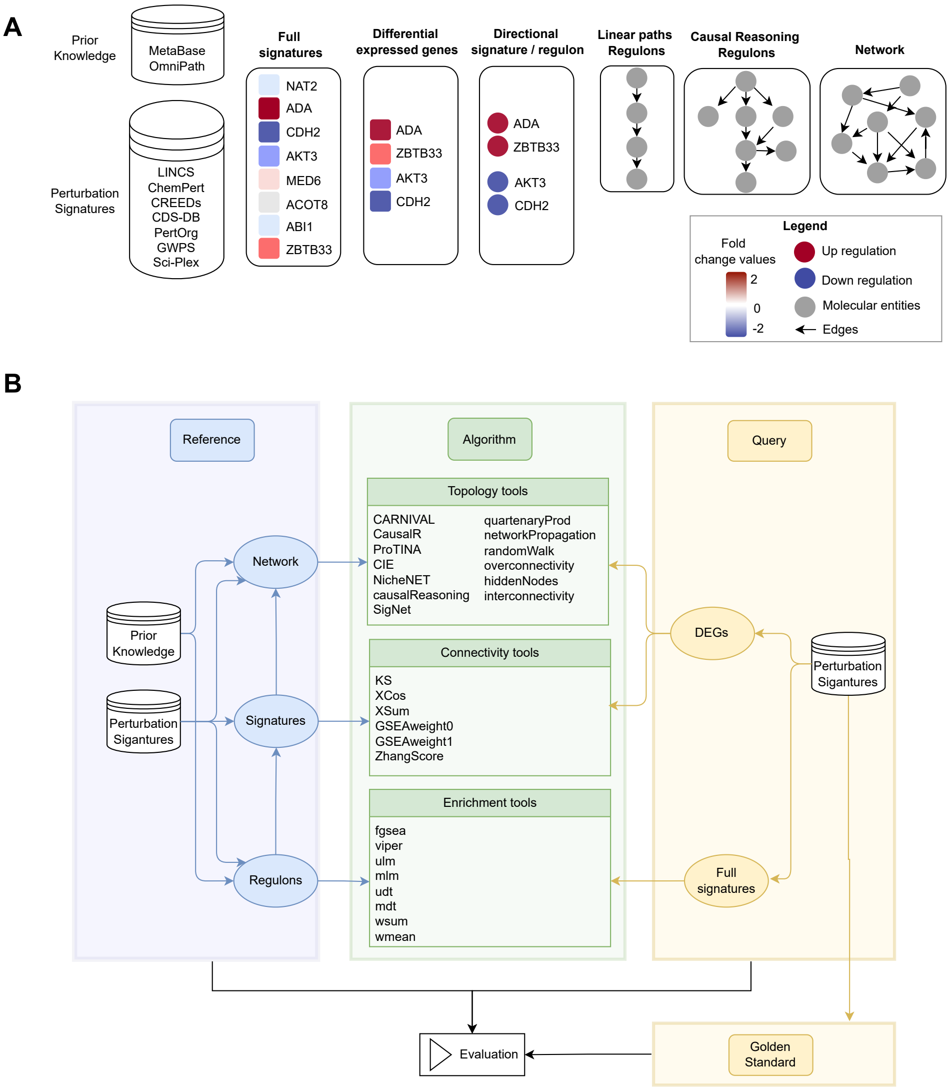

Figure 3.1: Schematic of the study architecture. A. Perturbation signatures collected from seven public sources are used in the benchmarking framework either as reference, query, and gold standard (known targets) datasets. Prior knowledge networks, used as reference, were derived from two sources: OmniPath (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 transcriptomics datasets from 10 different public sources, summarized in Table 3. Chemical and genetic perturbagens were included, spanning bulk microarray, bulk RNA seq, and single-cell RNA seq assays. Each dataset contains more than hundreds of perturbation signatures. For each collection, the perturbagen type, the total number of unique perturbagens profiled, and the subset for which a gold standard target annotation is available were recorded. The gold standard is necessary for the evaluation. It is a set of known targets, whether drug-protein interactions or genes deliberately perturbed which were used to assess the ability of the algorithms to recover true upstream regulators from observed expression changes. The LINCS expands upon the original CMap by leveraging the cost-effective L1000 platform, which directly measures 978 "landmark" transcripts and imputes the remaining transcriptome to reconstruct genome-wide expression profiles. LINCS comprises several distinct collections of perturbations in human cell lines: over 30,000 unique small-molecule treatments, CRISPR knockouts targeting 5,156 genes, cDNA overexpression of 3,780 genes, and shRNA knockdowns of 4,854 genes. The level 5 data were retrieved from the CLUE platform (available at: `https://clue.io/data/CMap2020#LINCS2020`). This level already contains the differential expression signatures with z-scores aggregated across biological replicates without p-values. Since each perturbagen usually appears under multiple conditions (different doses, time points, and cell lines), these were condensed into a single consensus signature per perturbagen by extracting every available gene's z-score and then using the median value across signatures. For the gold standard, directional effects were assigned as follows: for chemical perturbations, CDDI (Cortellis Drug Discovery Intelligence) database annotations (i.e. if the drug is annotated as the antagonist of target X, that target would be assigned negative effect in golden standard); for CRISPR and shRNA datasets, the target genes were assigned with inhibition effect, and for OE, each target was assigned with activation effect. ChemPert is a manually curated compendium of 82,256 transcriptional signatures derived from non-cancer cell compound perturbation experiments. Most signatures originate from bulk expression studies in various cell lines, and each is represented as a list of DEGs indicating only up- or down-regulation (no fold-change values or p-values).

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

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

Table 3.1: Datasets summary.

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

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

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

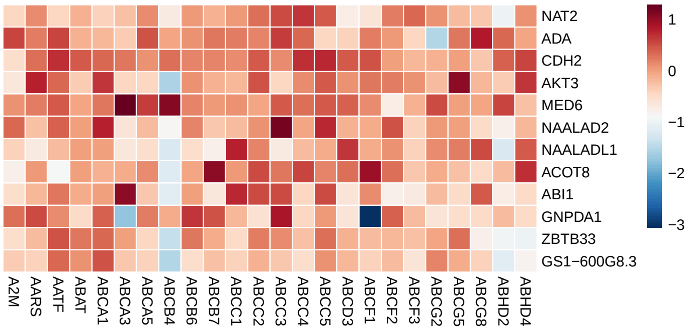

Figure 3.2: Transctipotomic signatures perturbation from LINCS shRNA, cell line a375.

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	3,3927	1,5229	1,7693	81,866	61,214	101,752	657,746
	Regulons	1,1739	1,0476	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

3.2.2 Prior Knowledge: Interaction Networks

One type of input that can serve as a reference is prior knowledge data for contextualizing gene expression signatures. The benchmarking framework depends on three complementary types of this data: PKN (global interaction networks), regulons (regulator-target gene sets), and pathway-derived linear maps (Table **??**). Although these resources vary in their coverage, they are related, as illustrated in Figure 4. Including sources of different sizes and densities is particularly important for understanding how topology-based algorithms manage that in terms of performance. Additionally, an increase in network size can introduce noise that may disturb the rationale.

3.3 Algortithms

To carry out a systematic and robust comparative evaluation of inference algorithms, wrapper functions were developed to standardize the input data and output formatting in different computational approaches. Each algorithm has specific data requirements and processing methods, requiring adaptation to ensure compatibility in the common framework. A wrapper is nothing more than a function that serves as an intermediary layer. These are important to handle data type conversions, parameter standardization, and result formatting, allowing diverse algorithms to be executed consistently regardless of their underlying implementation differences. This approach addresses the inherent complexity of having algorithms coming from different approaches. Here there are two types of wrappers: shared and individual. The shared wrapper architecture incorporates an already established package that bundles several algorithms inside, unlike the individual ones that incorporate single algorithms. The connectivity mapping approaches from the RCSM package, enrichment methods from decoupleR, and topology-based algorithms from CBDD were implemented in shared wrappers. On the other hand, causal reasoning CARNIVAL, CausalR, ProTINA, CIE, and NicheNet were incorporated in individual wrappers. Table 5 provides a complete list of algorithms together with their annotations.

Some supporting helper functions were also implemented to facilitate essential data conversions across all wrappers. Those functions include mapping identifiers between transcriptomic datasets and network nodes to ensure the same IDs and converting the input data when necessary. For the query input data, the tool may need a full signature or DEGs. When DEGs are required, the full signature can be filtered using a fold change and p-value threshold or by simply taking the top threshold for differentially expressed genes by fold change magnitude. Regarding reference, the workflow can start with PKN or full signatures, and for topology, enrichment, and CMap tools, require networks, regulons/gene sets, and full signatures, respectively. To use this large variety of input data and tools, some conversions are required to run them uniformly. All the conversions are indicated by the arrows in Figure 4 B. The parameters selected for each algorithm can be found in Supplementary table 2. As the input data, the output should also have the same format, so it is possible to evaluate the performance of each algorithm. For that reason, at the end of each run, all algorithm wrappers return a table with all prioritized regulators identified without any significance filtering applied. The output contains a score column, and the larger score reflects greater confidence in this regulator being causal for observed differential expression patterns. Score may be signed if the tools can predict directionality of perturbation. In that case, regulators are ranked by absolute value of score, and activation/repression status is stored in separate column effect (-1/1 values).

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

Table 3.3: Algorithm summary.

Tool	Algorithm	Description	Resource	Dataset	Reference
Causal Reasoning					
CARNIVAL	Description	Network	DEG/Full	[0]	
CausalR	Description			[0]	
ProTINA	Description			[0]	
CIE	Description			[0]	
NicheNet	Description			[0]	
causalReasoning	Description			[0]	
Signatures	Description			[0]	
quaternaryProd	Description			[0]	
Causal Reasoning (baseline)					
randomWalk	Description	Network	DEG/Full	[0]	
networkPropagation	Description			[0]	
overconnectivity	Description			[0]	
hiddenNodes	Description			[0]	
interconnectivity	Description			[0]	

overconnectivity	Description			[0]
CMap				
KS	Description	Signatures	DEG/Full	[0]
XCos	Description			[0]
XSum	Description			[0]
ZhangScore	Description			[0]
GSEAweight0	Description			[0]
GSEAweight1	Description			[0]
Enrichment				
wmean	Description	Regulons	Signatures	[0]
fgsea	Description			[0]
viper	Description			[0]
ulm	Description			[0]
mlm	Description			[0]
udt	Description			[0]
mdt	Description			[0]
wsum	Description			[0]

3.3.1 Connectivity Mapping

The **??** represents the wrapper function framework for running connectivity mapping algorithms from the RCSM package [87]. This package provides uniform implementations of several CMap scoring methods including Kolmogorov-Smirnov (KS), and GSEA-based approaches. The function is designed to accept filtered DEG lists as query input and full perturbation signatures as reference data. If full signatures are used as the query, they are converted to DEGs using the filtering parameters (Supplementary table 2), as well as the additional parameters. RCSM R package includes a variety of algorithms already implemented, each designed to quantify the similarity or dissimilarity between query and reference perturbation signatures. Those algorithms include the Kolmogorov-Smirnov (KS) statistic which was used in the original Connectivity Map [9]; Xcos, a cosine similarity metric between query and reference fold-changes; Xsum connectivity map statistic based on the sum of reference fold-change values of query genes; GSEAweight0 is a GSEA weighted KS ES with parameter p = 0, meaning that the fold-change magnitude is not taken into account; GSEAweight1 with parameter p = 1, fold-change magnitude contributes linearly; Zhang, a connectivity mapping score first suggested in [105]. The function handles the different algorithm requirements by preparing either separate up- and down-regulated gene lists for most methods or simple gene vectors for XCos. The function also includes optional regulator filtering for TF mode. The output is formatted to return regulator rankings with similarity scores, directional effects, and optional statistical significance measures. The results are sorted by absolute score magnitude to prioritize the most relevant regulatory relationships regardless of similarity direction. For these algorithms, the regulator score measures the similarity of the query versus the reference perturbation signature.

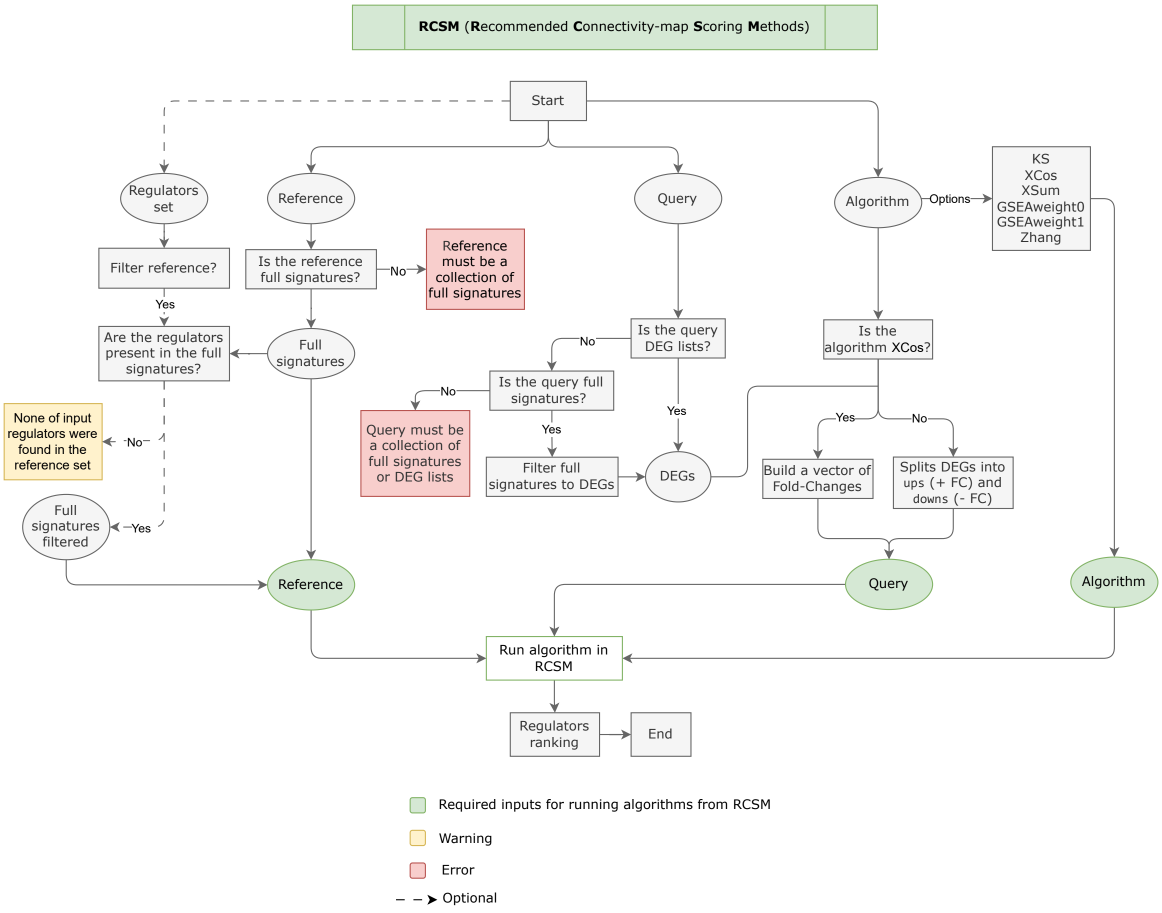
RCSM (Recommended Connectivity-map Scoring Methods)
Start
Regulators set
Reference
Query
Algorithm
Options
KS
XCos
XSum
GSEAweight0
GSEAweight1
Zhang
Filter reference?
Is the reference full signatures?
No
Reference must be a collection of full signatures
Yes
Are the regulators present in the full signatures?
Full signatures
Is the query DEG lists?
No
Is the algorithm XCos?
Is the query full signatures?
No
Query must be a collection of full signatures or DEG lists
Yes
Filter full signatures to DEGs
Yes
DEGs
Yes
No
Build a vector of Fold-Changes
Splits DEGs into ups (+ FC) and downs (- FC)
None of input regulators were found in the reference set
No
Full signatures filtered
Yes
Reference
Query
Algorithm
Run algorithm in RCSM
Regulators ranking
End
🟩 Required inputs for running algorithms from RCSM
🟨 Warning
🟥 Error
- - ➤ Optional

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

3.3.2 Pathway Enrichment

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

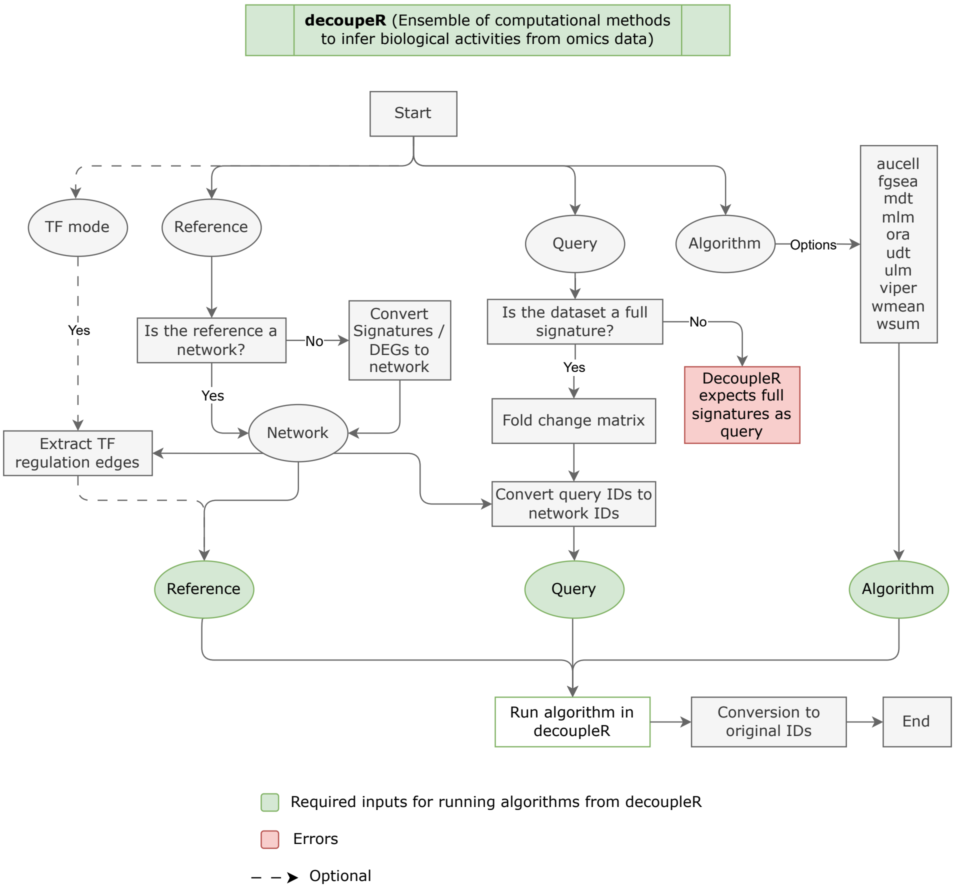
decoupeR (Ensemble of computational methods to infer biological activities from omics data)
Start
TF mode
Reference
Query
Algorithm
Options
aucell
fgsea
mdt
mlm
ora
udt
ulm
viper
wmean
wsum
Yes
Is the reference a network?
No
Convert Signatures / DEGs to network
Yes
Network
Extract TF regulation edges
Is the dataset a full signature?
No
Yes
DecoupleR expects full signatures as query
Fold change matrix
Convert query IDs to network IDs
Reference
Query
Algorithm
Run algorithm in decoupleR
Conversion to original IDs
End
Required inputs for running algorithms from decoupleR
Errors
Optional

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

4

Results

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

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

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, including the activity of transcription factors and thus the expression of the genes. All these changes that happen in the system will ultimately produce the desired (or undesired) drug effect. As with many topics in the scientific community, there are different views on the real need to know the exact molecular target or MoA of a drug's function [106]. Some defend and prioritize the practical results of drugs, given the effective number of drugs on the market for which the MoA is 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. Given the benefits, advances in this area are moving towards finding methods that accelerate this time-consuming process while maintaining reliable results. For example, in 2020, the Connectivity Mapping Institute created a challenge on the Kaggle platform [107] to develop an algorithm that would 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 evaluate and compare it with others [95]. The aim is to demonstrate that the algorithm outperforms others already existing. However, these studies have some limitations. Typically, simulated data, such as the LINCS dataset, is used instead of real experimental data. Results from this simulated data may be unreliable because the variability and complexity of biological systems are not fully assessed. Benchmarking studies are therefore essential because they aim to test various components in an unbiased and realistic way to evaluate the performance 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 of around 5 seconds per signature. These results are consistent with other studies. In Hill et al. [60], node prioritization methods, including randomWalk followed by networkPropagation and GeneMANIA, were also identified as the top performers. Our study demonstrated that performance depends on the choice of reference network, with the MetaBase network providing the best results, probably due to its greater molecular coverage. Hill et al. used a combination of interactions from STRING, MetaBase, and BioPlex. This variation in topological networks while maintaining the same performance suggests that randomWalk is a robust and reliable algorithm. The overconnectivity and networkPropagation algorithms also performed well in our study. However, each had some limitations in terms of accuracy, reliability, or run time. For example, networkPropagation obtained results like randomWalk in 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 context that previous studies did not characterize. 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. The selection of methods goes beyond simple performance rankings. When 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 [87] and others [108] identified ZhangScore as the best algorithm among connectivity mapping methods. However, in our study, this algorithm performed poorly among all. 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 through the worst computational efficiency, by having the highest runtime. This huge difference in the results can be interpreted as the actual 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 better at pathway-level enrichment compared with precise target identification. For example, causalReasoning achieved perfect reliability, but modest AUC, aligning with the observation that these methods struggle to pinpoint exact drug targets directly. 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 with causal reasoning algorithms in regard with other studies, was scalability. With the complete failure of the most sophisticated algorithms, such as, CARNIVAL. Especially, when these algorithms performed well. In previous benchmarking studies SigNet and CARNIVAL with OmniPath network as reference data had the best performance [27]. However, while 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 failing to complete 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. 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. The quality of the datasets can be a determining factor in the performance of these tools. The truth is that, 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 them [51]. The data is only checked on a technical level, if the samples really come from GEO, and if the samples are labeled correctly. The attempt at biological evaluation, by looking for specific patterns of signatures, if two signatures come from perturbing the same gene, showed inconsistent results. Moreover, there are certain datasets in which the authors' validation was practically non-existent [39, 56], but the filters used in this study to prepare the dataset increased the quality of the data, as was the case for 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 [54]. Sci-Plex, with drug perturbations, validated its data through some complementary statistical analysis [42]. The data from LINCS are more controversial. First, L1000 measures around 978 landmark genes and computationally infers the rest (12,000). And although each perturbation is tested in triplicate, and z-scores are adjusted to minimize replicate inconsistencies [40], the reliability of these data is questionable. Recent alternatives, such as DRUG-seq, demonstrate advantages. One study compared L1000 data with DRUG-Seq data [52, 53]. DRUG-seq directly measures more than 10,000 genes without relying on inference, including several not covered by L1000, 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 [54], offers another promising alternative with greater transcriptome coverage at a reduced cost [109]. Also, the cancer cell line composition limits LINCS's applicability to other contexts. This is being addressed by programs such as NeuroLINCS [110], which create signatures using patient-derived induced pluripotent stem cells [109]. 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 [28]. Tools for constraining networks based on tissue-specific expression data from resources like the Human Protein Atlas [111] or specialized databases like TissueNet [65] could improve performance [28]. Our results showing MetaBase Linear Path Regulons' excellent pathway enrichment despite poor coverage support this tissue-specific approach. The focus on transcriptomics-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 [85] and COSMOS [112] 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, such as known clear cell renal cell carcinoma drug targets, illustrates how the integration of multiple data types can generate novel biomarker hypotheses. Data capturing changes in cell morphology after perturbation, and the chemical structure of compounds [113] can become a valuable complement to expression data in DD studies [28]. Some considerations and future work that can be explored have to do with certain adjustments or paths that could be taken, but which, due to the size of the project for this phase, were not possible to include. One aspect that has not been considered in this study, but which could effectively influence the behavior and therefore the performance of the algorithms, is the optimization of the algorithm parameters. Algorithm behavior and performance can change dramatically with parameter tuning, and it is suggested as good practice for benchmarking studies [94]. Yet comprehensive optimization across all algorithm-dataset combinations proved computationally unreasonable for this study. The same goes for the methods for filtering datasets when the algorithms required DEG, and the input query data was a full signature. Other filtering methods can be tested. The parameters of the algorithms and the preparation of the input data are defined and prepared in the wrapper functions to be changed if needed. As this is a benchmarking study, the entire workflow is set up precisely to test multiple options, even though not all of them have been evaluated in this work. Regarding evaluation, 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 can 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 [28]. 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, use databases such as TissueNet [65], which provides 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. And following the reasoning, it could be customized to suit the context and even keep only highly reliable interactions. If we do not just focus on this type of algorithm, there are more software and interesting tools with a similar purpose, but with different reasoning, that may be worth exploring too, always depending on the question we want to answer. One example is Drug2ways [114], 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 [115]. In summary, based on the results obtained and the intersection of previous studies, 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 considered. In addition, other algorithm optimization parameters should be considered, rather than sticking to the default settings. Finally, with the evolution of all these areas and the avalanche of increasingly available data, the integration of other levels of OMICS data and even the use of machine learning methods should be considered.

References

[0] E. R. G. (ERG). *Drug Development - Final Report*. Report. 2024. URL: https://aspe.hhs.gov/reports/drug-development.

[0] B. Alexander-Dann et al. "Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data". In: *Mol Omics* 14.4 (2018), pp. 218–236. DOI: 10.1039/c8mo00042e.

[0] A. Antona et al. "Dissecting the Mechanism of Action of Spiperone-A Candidate for Drug Repurposing for Colorectal Cancer". In: *Cancers (Basel)* 14.3 (2022). DOI: 10.3390/cancers14030776.

[0] J. Avorn. "The 2.6$BillionPill^{-}MethodologicandPolicyConsiderations$". In: *New England Journal of Medicine* 372.20 (2015), pp. 1877–1879. DOI: doi:10.1056/NEJMp1500848. URL: https://www.nejm.org/doi/full/10.1056/NEJMp1500848.

[0] P. Badia-i-Mompel et al. "decoupleR: ensemble of computational methods to infer biological activities from omics data". In: *Bioinformatics Advances* 2.1 (2022). ISSN: 2635-0041. DOI: 10.1093/bioadv/vbac016. URL: https://doi.org/10.1093/bioadv/vbac016.

[0] I. Bezprozvanny. "The rise and fall of Dimebon". In: *Drug News Perspect* 23.8 (2010), pp. 518–23. DOI: 10.1358/dnp.2010.23.8.1500435.

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

[0] A. Brazma et al. "ArrayExpress—a public repository for microarray gene expression data at the EBI". In: *Nucleic Acids Research* 31.1 (2003), pp. 68–71. ISSN: 0305-1048. DOI: 10.1093/nar/gkg091. URL: https://doi.org/10.1093/nar/gkg091.

[0] R. Browaeys, W. Saelens, and Y. Saeys. "NicheNet: modeling intercellular communication by linking ligands to target genes". In: *Nature Methods* 17.2 (2020), pp. 159–162. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0667-5. URL: https://doi.org/10.1038/s41592-019-0667-5.

[0] J. Cao et al. "Comprehensive single-cell transcriptional profiling of a multicellular organism". In: *Science* 357.6352 (2017), pp. 661–667. DOI: doi:10.1126/science.aam8940. URL: https://www.science.org/doi/abs/10.1126/science.aam8940.

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

[0] S. Chakraborti, G. Ramakrishnan, and N. Srinivasan. "Repurposing Drugs Based on Evolutionary Relationships Between Targets of Approved Drugs and Proteins of Interest". In: *Computational Methods for Drug Repurposing*. Ed. by Q. Vanhaelen. New York, NY: Springer New York, 2019, pp. 45–59. ISBN: 978-1-4939-8955-3. DOI: 10.1007/978-1-4939-8955-3_3. URL: https://doi.org/10.1007/978-1-4939-8955-3_3.

[0] D. Cook et al. "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework". In: *Nature Reviews Drug Discovery* 13.6 (2014), pp. 419–431. ISSN: 1474-1784. DOI: 10.1038/nrd4309. URL: https://doi.org/10.1038/nrd4309.

[0] R. Edgar, M. Domrachev, and A. E. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Research* 30.1 (2002), pp. 207–210. ISSN: 0305-1048. DOI: 10.1093/nar/30.1.207. URL: https://doi.org/10.1093/nar/30.1.207.

[0] A. Engelberg. "Iconix Pharmaceuticals, Inc.–removing barriers to efficient drug discovery through chemogenomics". In: *Pharmacogenomics* 5.6 (2004), pp. 741–744. ISSN: 1462-2416.

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

[0] B. Ganter et al. "Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database". In: *Pharmacogenomics* 7.7 (2006), pp. 1025–44. DOI: 10.2217/14622416.7.7.1025.

[0] B. Ganter et al. "Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action". In: *Journal of biotechnology* 119.3 (2005), pp. 219–244. ISSN: 0168-1656.

[0] L. Gennari et al. "Raloxifene in breast cancer prevention". In: *Expert Opin Drug Saf* 7.3 (2008), pp. 259–70. DOI: 10.1517/14740338.7.3.259.

[0] D. Hadley et al. "Precision annotation of digital samples in NCBI's gene expression omnibus". In: *Scientific data* 4.1 (2017), pp. 1–11. ISSN: 2052-4463.

[0] L. Hosseini-Gerami et al. "Benchmarking causal reasoning algorithms for gene expression-based compound mechanism of action analysis". In: *BMC Bioinformatics* 24.1 (2023), p. 154. ISSN: 1471-2105. DOI: 10.1186/s12859-023-05277-1. URL: https://doi.org/10.1186/s12859-023-05277-1.

[0] T. R. Hughes et al. "Functional discovery via a compendium of expression profiles". In: *Cell* 102.1 (2000), pp. 109–126. ISSN: 0092-8674.

[0] Y. Igarashi et al. "Open TG-GATEs: a large-scale toxicogenomics database". In: *Nucleic acids research* 43.D1 (2015), pp. D921–D927. ISSN: 1362-4962.

[0] M. Iwata et al. "Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics". In: *Scientific Reports* 7.1 (2017), p. 40164. ISSN: 2045-2322. DOI: 10.1038/srep40164. URL: https://doi.org/10.1038/srep40164.

[0] X. Ji, J. M. Freudenberg, and P. Agarwal. "Integrating Biological Networks for Drug Target Prediction and Prioritization". In: *Computational Methods for Drug Repurposing*. Ed. by Q. Vanhaelen. New York, NY: Springer New York, 2019, pp. 203–218. ISBN: 978-1-4939-8955-3. DOI: 10.1007/978-1-4939-8955-3_12. URL: https://doi.org/10.1007/978-1-4939-8955-3_12.

[0] Å. Johansson et al. "Precision medicine in complex diseases-Molecular subgrouping for improved prediction and treatment stratification". In: *J Intern Med* 294.4 (2023), pp. 378–396. ISSN: 0954-6820 (Print) 0954-6820. DOI: 10.1111/joim.13640.

[0] B. B. Kakoti, R. Bezbaruah, and N. Ahmed. "Therapeutic drug repositioning with special emphasis on neurodegenerative diseases: Threats and issues". In: *Front Pharmacol* 13 (2022), p. 1007315. ISSN: 1663-9812 (Print) 1663-9812. DOI: 10.3389/fphar.2022.1007315.

[0] S. Kato et al. "Challenges and perspective of drug repurposing strategies in early phase clinical trials". In: *Oncoscience* 2.6 (2015), pp. 576–80. DOI: 10.18632/oncoscience.173.

[0] A. B. Keenan et al. "Connectivity Mapping: Methods and Applications". In: *Annual Review of Biomedical Data Science* 2.Volume 2, 2019 (2019), pp. 69–92. DOI: https://doi.org/10.1146/annurev-biodatasci-072018-021211. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-biodatasci-072018-021211.

[0] A. E. Kel. "Search for Master Regulators in Walking Cancer Pathways". In: *Biological Networks and Pathway Analysis*. Ed. by T. V. Tatarinova and Y. Nikolsky. New York, NY: Springer New York, 2017, pp. 161–191. ISBN: 978-1-4939-7027-8. DOI: 10.1007/978-1-4939-7027-8_8. URL: https://doi.org/10.1007/978-1-4939-7027-8_8.

[0] J. Lamb et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". In: *Science* 313.5795 (2006), pp. 1929–35. DOI: 10.1126/science.1132939.

[0] K. Lin et al. "A comprehensive evaluation of connectivity methods for L1000 data". In: *Brief Bioinform* 21.6 (2020), pp. 2194–2205. DOI: 10.1093/bib/bbz129.

[0] A. Liu et al. “From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL”. In: *npj Systems Biology and Applications* 5.1 (2019), p. 40. ISSN: 2056-7189. DOI: 10.1038/s41540-019-0118-z. URL: https://doi.org/10.1038/s41540-019-0118-z.

[0] Z. Liu et al. “CDS-DB, an omnibus for patient-derived gene expression signatures induced by cancer treatment”. In: *Nucleic Acids Res* 52.D1 (2024), pp. D1163–d1179. DOI: 10.1093/nar/gkad888.

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

[0] B. K. Martin et al. “Optimized single-nucleus transcriptional profiling by combinatorial indexing”. In: *Nature Protocols* 18.1 (2023), pp. 188–207. ISSN: 1750-2799. DOI: 10.1038/s41596-022-00752-0. URL: https://doi.org/10.1038/s41596-022-00752-0.

[0] C. Martini et al. “CEBS update: curated toxicology database with enhanced tools for data integration”. In: *Nucleic Acids Res* 50.D1 (2022), pp. D1156–d1163. DOI: 10.1093/nar/gkab981.

[0] “Mechanism matters”. In: *Nature Medicine* 16.4 (2010), pp. 347–347. ISSN: 1546-170X. DOI: 10.1038/nm0410-347. URL: https://doi.org/10.1038/nm0410-347.

[0] S. Müller-Dott et al. “Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities”. In: *Nucleic Acids Research* 51.20 (2023), pp. 10934–10949. ISSN: 0305-1048. DOI: 10.1093/nar/gkad841. URL: https://doi.org/10.1093/nar/gkad841.

[0] S. K. Nair et al. “ToxicoDB: an integrated database to mine and visualize large-scale toxicogenomic datasets”. In: *Nucleic Acids Research* 48.W1 (2020), W455–W462. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa390. URL: https://doi.org/10.1093/nar/gkaa390.

[0] H. Noh, J. E. Shoemaker, and R. Gunawan. “Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection”. In: *Nucleic Acids Res* 46.6 (2018), e34. DOI: 10.1093/nar/gkx1314.

[0] K. Park. “A review of computational drug repurposing”. In: *Transl Clin Pharmacol* 27.2 (2019), pp. 59–63. DOI: 10.12793/tcp.2019.27.2.59.

[0] M. Pilarczyk et al. “Connecting omics signatures and revealing biological mechanisms with iLINCS”. In: *Nature Communications* 13.1 (2022), p. 4678. ISSN: 2041-1723. DOI: 10.1038/s41467-022-32205-3. URL: https://doi.org/10.1038/s41467-022-32205-3.

[0] J. K. Ryu and J. G. McLarnon. "Thalidomide inhibition of perturbed vasculature and glial-derived tumor necrosis factor-alpha in an animal model of inflamed Alzheimer's disease brain". In: *Neurobiol Dis* 29.2 (2008), pp. 254–66. DOI: 10.1016/j.nbd.2007.08.019.

[0] F. Sohraby, M. Bagheri, and H. Aryapour. "Performing an In Silico Repurposing of Existing Drugs by Combining Virtual Screening and Molecular Dynamics Simulation". In: *Computational Methods for Drug Repurposing*. Ed. by Q. Vanhaelen. New York, NY: Springer New York, 2019, pp. 23–43. ISBN: 978-1-4939-8955-3. DOI: 10.1007/978-1-4939-8955-3_2. URL: https://doi.org/10.1007/978-1-4939-8955-3_2.

[0] S. R. Srivatsan et al. "Massively multiplex chemical transcriptomics at single-cell resolution". In: *Science* 367.6473 (2020), pp. 45–51. DOI: 10.1126/science.aax6234.

[0] R. B. Stoughton and S. H. Friend. "How molecular profiling could revolutionize drug discovery". In: *Nature Reviews Drug Discovery* 4.4 (2005), pp. 345–350. ISSN: 1474-1776.

[0] A. Subramanian et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles". In: *Cell* 171.6 (2017), 1437–1452.e17. DOI: 10.1016/j.cell.2017.10.049.

[0] M.-A. Trapotsi, L. Hosseini-Gerami, and A. Bender. "Computational analyses of mechanism of action (MoA): data, methods and integration". In: *RSC Chemical Biology* 3.2 (2022), pp. 170–200. DOI: 10.1039/D1CB00069A. URL: http://dx.doi.org/10.1039/D1CB00069A.

[0] M. Wang et al. "TREAP: A New Topological Approach to Drug Target Inference". In: *Biophysical Journal* 119.11 (2020). doi: 10.1016/j.bpj.2020.10.021, pp. 2290–2298. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2020.10.021. URL: https://doi.org/10.1016/j.bpj.2020.10.021.

[0] Y. Wang, J. Yella, and A. G. Jegga. "Transcriptomic Data Mining and Repurposing for Computational Drug Discovery". In: *Computational Methods for Drug Repurposing*. Ed. by Q. Vanhaelen. New York, NY: Springer New York, 2019, pp. 73–95. ISBN: 978-1-4939-8955-3. DOI: 10.1007/978-1-4939-8955-3_5. URL: https://doi.org/10.1007/978-1-4939-8955-3_5.

[0] Z. Wang et al. "Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd". In: *Nat Commun* 7 (2016), p. 12846. ISSN: 2041-1723. DOI: 10.1038/ncomms12846.

[0] M. Waters et al. "CEBS–Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data". In: *Nucleic Acids Res* 36.Database issue (2008), pp. D892–900. DOI: 10.1093/nar/gkm755.

[0] Z. Wei et al. “PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization”. In: *Nucleic Acids Research* 53.D1 (2024), pp. D1099–D1111. ISSN: 1362-4962. DOI: 10.1093/nar/gkae858. URL: https://doi.org/10.1093/nar/gkae858.

[0] Z. Wu, Y. Wang, and L. Chen. “Network-based drug repositioning”. In: *Molecular BioSystems* 9.6 (2013), pp. 1268–1281. ISSN: 1742-206X. DOI: 10.1039/C3MB25382A. URL: http://dx.doi.org/10.1039/C3MB25382A.

[0] Z. Zhai et al. “PertOrg 1.0: a comprehensive resource of multilevel alterations induced in model organisms by in vivo genetic perturbation”. In: *Nucleic Acids Res* 51.D1 (2023), pp. D1094–d1101. DOI: 10.1093/nar/gkac872.

[0] Y. Zhang et al. “PerturbAtlas: a comprehensive atlas of public genetic perturbation bulk RNA-seq datasets”. In: *Nucleic Acids Research* 53.D1 (2024), pp. D1112–D1119. ISSN: 1362-4962. DOI: 10.1093/nar/gkae851. URL: https://doi.org/10.1093/nar/gkae851.

[0] K. Zhao and H.-C. So. “Using Drug Expression Profiles and Machine Learning Approach for Drug Repurposing”. In: *Computational Methods for Drug Repurposing*. Ed. by Q. Vanhaelen. New York, NY: Springer New York, 2019, pp. 219–237. ISBN: 978-1-4939-8955-3. DOI: 10.1007/978-1-4939-8955-3_13. URL: https://doi.org/10.1007/978-1-4939-8955-3_13.

[0] M. Zheng et al. “ChemPert: mapping between chemical perturbation and transcriptional response for non-cancer cells”. In: *Nucleic Acids Res* 51.D1 (2023), pp. D877–d889. DOI: 10.1093/nar/gkac862.

A

NOVAthesis covers showcase

Text
Text

A.1 A section here

Text Text

B

Appendix 2 Lorem Ipsum

Text

Text
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC Text

I

Annex 1 Lorem Ipsum

Text

UNIVERSIDADE NOVA
DE LISBOA

2025

BENCHMARKING COMPUTATIONAL ALGORITHMS FOR ENHANCED DRUG DISCOVERY

Maria Inês Nunes Vilar Gomes

MASTER IN **Computational Biology and Bioinformatics**

SPECIALIZATION Multi-Omics for Life and Health Sciences

NOVA
UNIVERSIDADE NOVA
DE LISBOA