
UNIVERSIDADE NOVA DE LISBOA
itqb nova

NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

NOVA
MEDICAL SCHOOL

NOVA
IMS
Information
Management
School

DESDE
INST
ME
UNIVER

MASTER IN STUDY PROGRAM NAME

SPECIALIZATION MULTI-OMICS FOR LIFE AND HEALTH SCIENCES

NOVA University Lisbon

September, 2025

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



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

Dedicatory lorem ipsum.

# Acknowledgements

Acknowledgments

”

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

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

# Abstract

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

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

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

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

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

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

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

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

# Resumo

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

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

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

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

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

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

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

# Contents

# List of Figures

# Glossary

**MetaBase** MetaBase from Clarivate [**metabase**] *9)*

# Acronyms

**ABC** Algorithm Benchmarking Consortium *3)*

**CBDD** Computational Biology for Drug Discovery *9)*
**CMap** Connectivity Mapping *3)*

**DEGs** Differentially Expressed Genes *3)*

**MoA** Mechanism of Action *2)*

# 1

# Introduction

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

## 1.1 Motivation and Goals

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

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

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

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

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

On the other hand, enrichment tools take perturbation signatures as query input and utilize prior knowledge, such as regulon network or collections of perturbation-induced differentially expression genes (DEGs), as a reference. The primary function of these tools is to assess whether certain regulons or gene sets (e.g., those associated with transcription factors (TFs)) are significantly enriched in the perturbed data. Several algorithms have been developed based on this approach, each producing an enrichment score. The Research and development (R&D) of new drugs is a fast-growing area that has also experienced significant growth in complexity in recent years. Due to the time-consuming, costly and multidisciplinary nature of the process, drug discovery remains a challenging domain. Over half of clinical trial failures are attributed to inefficiency, underscoring the importance of identifying and validating pharmacological targets, and highlighting the lack of knowledge of the drug's Mechanism of Action (MoA) as one of the major barriers in drug discovery. The thorough understanding of the MoA represents a critical initial step in this process, and computational methods can accelerate this by providing more efficient and cost-effective alternatives to traditional approaches. These approaches can accurately do target identification and prioritization, thereby reducing the need for lengthy experimental trials. A key to understanding a compound's MoA lies in transcriptomics data, which captures the molecular changes triggered by a perturbagen and reflects the system's changes in the gene expression profiles. While traditional RNA sequencing methods remain too costly for large-scale expression signatures, recent high-throughput technological advances, such as the L1000 assay, enable the cost-effective generation and analysis of large-scale omics datasets. Several existing databases provide public access to transcriptomic data from experiments involving diverse chemical and genetic perturbagens across different cell lines. These data can be exploited using various computational tools to establish the causes of specific gene expression changes in a biological system. Three primary approaches have emerged: causal reasoning, connectivity mapping and enrichment tools.

Causal reasoning, a topology-based method, utilizes a list of perturbation signatures and a biological interaction network to determine potential causes for the observed gene expression profile. The network is defined as signed and directed graph describing

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

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

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

## 1.2 Scope

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

## 1.3 Parallel Contributions

This study expands the state of art in causal reasoning using gene expression data and causal graphs by presenting a robust framework and methods for benchmarking various algorithms designed for this purpose. Beyond this primary focus, several parallel projects with real-world challenges and novel data were developed, and enriched the consultant

experience allowing for an expansion in the expertise of bioinformatics. The parallel projects, spanning various domains of computational biology, include:

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

**More projects ...** More projects ...

## 1.4 Structure

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

# 2

# Literature Review

*This section details the*

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

## 2.2 Causal Reasoning algorithms as an approach for target inference

# 3

# Materials and Methods

## 3.1 Gene expression data

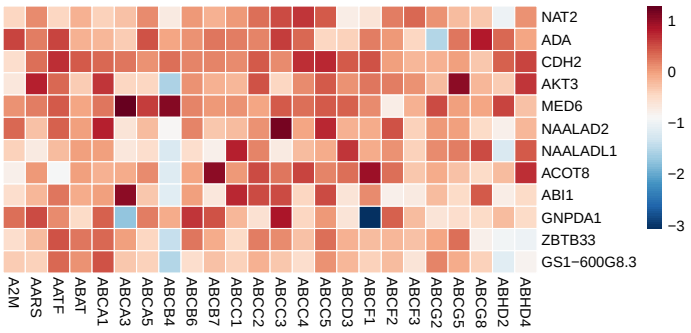


Figure 3.1: Transctipotomic signatures perturbation from LINCS shRNA, cell line a375.

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 |

## 3.2 Biological Network

## 3.3 Algortithms

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

Table 3.2: Algorithm summary.

| Tool | Algorithm | Description | Resource | Dataset | Reference |
|---|---|---|---|---|---|
| **Causal Reasoning** | | | | | |
| CARNIVAL | Description | Network | DEG/Full | [**carnival**] | |
| CausalR | Description | | | [**causalr**] | |
| ProTINA | Description | | | [**protina**] | |
| CIE | Description | | | [**cie**] | |
| NicheNet | Description | | | [**nichenet**] | |
| causalReasoning | Description | | | [**cbdd**] | |
| Signatures | Description | | | [**cbdd**] | |
| quaternaryProd | Description | | | [**cbdd**] | |
| **Causal Reasoning (baseline)** | | | | | |
| randomWalk | Description | Network | DEG/Full | [**cbdd**] | |
| networkPropagation | Description | | | [**cbdd**] | |
| overconnectivity | Description | | | [**cbdd**] | |
| hiddenNodes | Description | | | [**cbdd**] | |
| interconnectivity | Description | | | [**cbdd**] | |
| overconnectivity | Description | | | [**cbdd**] | |
| **CMap** | | | | | |
| KS | Description | Signatures | DEG/Full | [**cmap**] | |
| XCos | Description | | | [**cmap**] | |
| XSum | Description | | | [**cmap**] | |
| ZhangScore | Description | | | [**cmap**] | |
| GSEAweight0 | Description | | | [**cmap**] | |
| GSEAweight1 | Description | | | [**cmap**] | |
| **Enrichment** | | | | | |
| wmean | Description | Regulons | Signatures | [**cbdd**] | |
| fgsea | Description | | | [**fgsea**] | |
| viper | Description | | | [**viper**] | |
| ulm | Description | | | [**cbdd**] | |
| mlm | Description | | | [**cbdd**] | |
| udt | Description | | | [**cbdd**] | |

| | | |
|---|---|---|
| mdt | Description | [**cbdd**] |
| wsum | Description | [**fgsea**] |

# 4

# Adding Support to a New School (work in progress)

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

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

MetaBase

# A

# *NOVAthesis* covers showcase

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

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

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

## A.1 A section here

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

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

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

# B

# Appendix 2 Lorem Ipsum

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

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

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

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

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

# I

# Annex 1 Lorem Ipsum

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

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

UNIVERSIDADE NOVA
DE LISBOA