PhD Program in Biomolecular Sciences
Transdisciplinary Program in Computational Biology
Department of Cellular, Computational
and Integrative Biology – CIBIO
36th Cycle

Exploring the network's world

From omics-driven machine learning workflow for drug target identification to quantification of signaling model diversity

Advisor

Prof. Enrico DOMENICI

Department of Cellular, Computational and Integrative Biology, University of Trento

Tutor

Dr. Silvia PAROLO

Fondazione The Microsoft Research-University of Trento Centre for Computational and Systems Biology (COSBI)

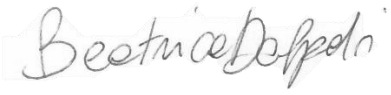
Prof. Giuseppe JURMAN

Fondazione Bruno Kessler

Ph.D. Thesis of Beatrice DALPEDRI
Department of Cellular, Computational and Integrative Biology, University of Trento
Fondazione COSBI
Academic Year 2023-2024

Declaration

I, Beatrice Dalpedri, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.



Abstract

The drug discovery process is challenging, time-consuming, and costly, with drug target identification being an essential step in developing effective therapies. Drug repurposing offers a strategy for identifying new uses for existing drugs, aiming to simplify the process. Machine learning models and network analysis methods have demonstrated promise in both drug target identification and repurposing, providing powerful tools for analyzing complex biological data. This thesis will explore the applications of neural networks and multilayer biological networks for drug repurposing opportunities and network inference problems applied to signaling pathways. A novel machine learning and network-based workflow is presented for identifying drug targets for cystinosis, a rare disease that causes progressive kidney disease, currently lacking effective therapies to prevent the kidney failure. This approach permits to recapitulate the disease mechanisms in the context of renal tubular physiology and identify candidate drug targets for further validation using a cross-species workflow and disease-relevant screening technologies. While machine learning approaches have shown promise, they often need more mechanistic understanding, which is necessary for robust drug target identification and repurposing strategies. Mechanistic models provide crucial insights into the underlying biological mechanisms, complementing machine learning techniques. However, inferring mechanistic signaling networks from omics data poses challenges due to non-identifiability, resulting in multiple valid solutions consistent with the data. After that, the focus shifts towards quantifying signaling model diversity through solver-agnostic solution sampling with CORNETO, an ongoing effort that aims to unify network inference problems via constrained optimization. Mechanistic signaling networks can be inferred from omics data and prior knowledge using combinatorial optimization and mathematical solvers to find the optimal network. However, this problem is in general, non-identifiable, and several solutions may be equally valid. Ignoring the existence of these alternative solutions leads to an incomplete picture of the hypothesis space of consistent mechanistic signaling networks. To alleviate this issue, an algorithm to explore the space of alternative solutions and to conduct sensitivity analysis on the optimal solution is implemented and presented. These algorithms are applied to data from pancreatic cancer cell lines treated with kinase inhibitors to study cellular responses to drug perturbations by inferring mechanistic signaling networks from omics data.

Index

List of Abbreviations

ANN: Artificial Neural Network
BH: Benjamini-Hochberg
BMA: Best-Match Average
BP: Biological Process
CC: Cellular Component
CKD: Chronic Kidney Disease
DAG: Directed Acyclic Graph
DL: Deep Learning
DR: Drug Repurposing
DM: Disease Molecules
FNN: Feedforward Networks
GO: Gene Ontology
GRN: Gene Regulatory Networks
HPO: Human Phenotype Ontology
ILP: Integer Linear Programming
KO: Gene knockouts
LogFC: Log Fold Change
LP: Linear Programming
MILP: Mixed Integer Linear Programming
MF: Molecular Function
MGI: Mouse Genome Informatics
ML: Machine Learning
MoA: Mechanisms Of Action
MPO: Mammalian Phenotype Ontology
PE: Pearson correlation
PG: Phenotype Gene
PH: Phi_s proportionality
PKN: Prior Knowledge Network
PPI: Protein-Protein Interaction
PT: Proximal Tubule
PU: Positive-Unlabeled
RF: Random Forest
RH: Rho proportionality
RWR: Random Walk with Restart
RT: Regularization Term
TF: Transcription Factors
WT: Wild type

1. Introduction

Drug Development and Drug Repurposing

Drug development is a complex, multi-phase process that involves discovering and bringing new pharmaceutical drugs to market. The drug discovery process is challenging, time-consuming, and extremely expensive. This process begins with identifying candidate therapeutic targets, often through basic research that uncovers the biological mechanisms underlying diseases. Then, the process moves to validating the target and identifying molecules that can modulate it. The development pipeline includes preclinical studies, where compounds are tested for efficacy and toxicity in vitro and in vivo. Once promising candidates are identified, they progress to clinical trials. Clinical trials are conducted in multiple phases to test safety, efficacy, and dosing in human subjects. The entire drug development process can take between 10 to 20 years, from discovery to market approval. The costs associated with bringing a new drug to market are substantial, often exceeding $2.6 billion (Chan et al., 2019; DiMasi et al., 2016; Pushpakom et al., 2018). The high cost and time involved in de-novo drug development have led to the emergence of drug repurposing (DR) strategies, the alternative approaches that identify new therapeutic uses for existing drugs. By leveraging the known safety profiles and pharmacokinetics of approved drugs or investigational compounds, these methodologies can significantly reduce the time and cost of drug development. It is estimated that repurposed drugs will cost approximately $300 million and could become a marketable treatment within an average period of 6.5 years (Nosengo, 2016). The success rate of repurposed drugs is five times higher than that of new drug development (Roessler et al., 2021).

Drug repurposing can be achieved through experimental or computational methods. Experimental approaches involve high-throughput screening of existing drug libraries against new biological targets to identify repurposing candidates. Phenotypic screening, where the effects of drugs are observed in disease models, can also uncover new therapeutic applications. Clinical observations occasionally lead to DR. For example, a drug initially developed for cardiovascular disease might exhibit beneficial effects in patients with another condition, such as diabetes. Understanding the underlying pathways of diseases can lead to repurposing drugs that target similar pathways in other diseases, particularly when multiple diseases share common molecular mechanisms. Computational approaches employ bioinformatics tools to predict new drug-disease relationships. Data mining, machine learning (ML), and network analysis of biological databases facilitate the identification of potential repurposing candidates (Pushpakom et al., 2018). Between the computational approaches, it is possible to classify the methodologies into five macro categories. (1) Signature matching methods involve transcriptomic and proteomic data analysis to identify drugs that are able to mimic gene and protein expression profiles associated with diseases and discover potential

repurposing opportunities. (2) Computational molecular docking methodologies are structure-based strategies. It can predict binding site complementarity between drugs and targets or molecules with analogous chemical structures. The similarity between compounds can be measured by employing a range of structural features, including 2D topological fingerprints or 3D conformations (J. Li et al., 2016). (3) Genome-wide association studies (GWAS) identify genetic variants associated with diseases and can suggest novel drug targets for repurposing. (4) Network-based approaches leverage protein-protein interaction (PPI) networks, signaling networks, gene regulatory networks (GRN) and others, to identify new drug targets within complex biological systems. These methods map relationships between genes, proteins, and diseases, facilitating the discovery of drugs that can affect multiple conditions (Misselbeck et al., 2019). (5) Retrospective clinical analysis of Electronic Heath Records (EHRs), post-marketing surveillance data, and clinical trial data can reveal repurposing opportunities. EHRs contain vast amounts of structured and unstructured data that, when analyzed, can identify new drug-disease associations (Pushpakom et al., 2018). Each of these computational approaches leverages different data types and methodologies, contributing to a more efficient and effective drug-repurposing process.

Thesis aims

The objectives of this thesis can be classified into two categories: methodological and biological.

First, from the methodological perspective, the aim is to examine and analyze three incremental applications of mathematics to models biological problem, ranging from the most deterministic to the most stochastic. Firstly, we undertake an examination of deterministic models, which are represented by standard network analysis algorithms. We then explore models that incorporate both deterministic and stochastic components, as well as optimization elements. These are exemplified by sensitivity analysis of integer linear programming problems and random walk models. Finally, the last category of models are those based entirely on statistical processes, such as deep neural networks.

Second, the aforementioned models are applied to two distinct biological questions: the identification of drug targets for rare diseases, specifically for cystinosis, and the exploration of signaling pathways in the context of Bafetinib-treated pancreatic cancer.

Chapter 2 of this thesis presents a computational workflow based on network analysis and machine learning approaches for drug repurposing applied in a rare disease context, the cystinosis. By developing a phenotype-aware computational framework, the goal is to identify new therapeutic targets and repurposing opportunities, to overcome the limitations of current cystinosis treatments, which do not prevent kidney failure.

Chapter 3 explores the challenge of non-identifiability of a single solution in the context of signaling pathway inference problems resolved with ILP model, with a particular focus

on applications in the field of pancreatic cancer. The algorithm for exploring alternative solutions space is presented, along with three methods to visualize and interpret the results. Subsequently, alternative solutions are analyzed with the aim of quantifying the robustness of the optimal solution and contextualizing the most alternative solution. These methodologies are applied to the real-world context of signaling pathway inference in the case of pancreatic cancer cells treated with Bafetinib.

In synthesis, this thesis presents methodologies that provide a wide-ranging overview of the diverse applications of mathematical solutions in the field of computational biology, particularly across different drug discovery phases.

Network analysis

Overview

Network analysis is a computational approach to study in silico complex systems involving numerous players. A network, also known as a graph, is a data structure consisting of a set of nodes (or vertices) connected by a set of edges (or links). In the case of biological networks, the nodes can represent molecular entities, such as genes, proteins, metabolites, or broader concepts, such as diseases or pathways. Instead, the edges encode information about the type (e.g., physical or functional) and interaction strength. They can be directed or undirected; a network is considered directed if all of its edges are directed; it is instead undirected if all its edges are undirected (Porras Millan et al., 2022). The following section will describe the main types of biological networks.

Types of Biological Networks

Protein-protein interaction networks

Protein-protein interaction (PPI) networks represent the physical interactions among proteins forming complexes. Given their nature, the edges of the PPI network are nondirected, as it cannot be determined which protein binds the other (Vidal et al., 2011). In the last decades, several efforts have been made to reconstruct the PPI networks of humans and other model organisms (Snider et al., 2015). The yeast two-hybrid (Y2H) system has been widely used to detect binary protein interactions on a genome-wide scale. In this in vivo genetic screening, the interaction between two proteins, usually called bait and prey, activates reporter genes that enable growth on specific media or a color reaction (Brückner et al., 2009). Complementary to the Y2H system, mass spectrometry (MS)-based approaches can identify PPI in an unbiased way (Richards et al., 2021). In affinity purification–mass spectrometry (AP-MS), cells are transfected with a protein carrying an affinity tag, anti-tag antibodies are then used to pull the protein down from cell lysates, and MS identifies interacting proteins. More recently, co-fractionation mass spectrometry (CF-MS) emerged as a powerful, high-throughput method for interactome mapping without requiring genetic manipulation

(Lund-Johansen et al., 2021; Skinnider & Foster, 2021). These technologies have led to the availability of increasingly larger PPI datasets. In 2014, a network map of 14,000 high-quality binary human protein-protein interactions was published (Rolland et al., 2014). Moreover, several efforts have been made to build reference maps of human binary PPIs by collecting interactions from multiple experiments. Among them, the Human Protein Reference Database (HPRD) is a database started in 2003 of curated PPIs supported by several lines of experimental evidence, including Y2H interactions, derived from the literature (Keshava Prasad et al., 2009). More recently, the HuRI network, including approximately 53,000 protein-protein interactions, was assembled (Luck et al., 2020).

Signaling networks

In addition to physical interactions, which are usually undirected, proteins can be connected to each other through functional interactions, such as phosphorylation. Signaling networks usually consist of interconnected signaling pathways represented by signed-directed graphs (Csabai et al., 2021). Each pathway is a series of molecular events initiated by a signal, propagated through a cascade of intracellular molecules that end in a specific cellular response. Understanding signaling networks is crucial for elucidating cellular behavior's complex mechanisms and identifying potential therapeutic targets in various diseases.

At the core of signaling networks are signaling molecules such as proteins, lipids, and ions that transmit information from the cell surface to the nucleus or other cellular compartments. These molecules interact through a series of events involving activation, inhibition, and feedback loops. The primary signaling network components include receptors, second messengers, kinases, phosphatases, transcription factors (TFs), and other regulatory proteins. The receptors are proteins located on the cell surface or within the cell that recognize and bind to specific signaling molecules. The binding of a ligand to its receptor often triggers a conformational change in the receptor, initiating the signaling cascade. The second messengers are small molecules that propagate the signal within the cell. These molecules amplify the signal received by the receptor and activate downstream signaling proteins. The kinases and phosphatases are enzymes that add or remove phosphate groups from proteins, respectively. These modifications alter the target proteins' activity, localization, and interactions, thus modulating the signaling pathway. Finally, the TFs are the proteins that regulate the expression of specific genes by binding to DNA sequences in the gene's promoter or enhancer regions. The activation or repression of TFs is a common endpoint of signaling pathways, leading to changes in gene expression that drive cellular responses.

A typical signaling pathway begins with the activation of a receptor by an extracellular signal. This activation triggers a series of intracellular events, often involving the generation of second messengers and the activation of protein kinases and other signaling molecules. The pathway concludes with the modulation of target proteins or gene expression, resulting in a specific cellular response. In the network representation

of signaling pathways, the nodes represent the signaling molecules, such as receptors, kinases, and TFs, while the directed edges represent the interactions between these molecules, including activation, inhibition, and binding events.

Numerous databases collect information on signaling networks, offering valuable resources for researchers. These databases provide curated data on signaling pathways, interactions, and regulatory mechanisms, facilitating the study and modeling of signaling networks. Key databases include KEGG (Kanehisa et al., 2020), SIGNOR (Lo Surdo et al., 2023), NetPath (Kandasamy et al., 2010), Reactome (Fabregat et al., 2018), and Panther (Mi et al., 2021).

Co-expression networks

The co-expression network is a type of network created to study the correlation patterns between genes across various biological conditions. These networks are crucial in identifying groups of genes that show similar expression patterns, thereby suggesting that these genes may be co-regulated or involved in analogous biological processes. Co-expression networks are utilized in genomics to explain gene function, regulatory mechanisms, and the underlying structure of gene interactions.

Co-expression networks are inferred from transcriptomic data where the edges among nodes (genes) indicate that the genes' expression levels are correlated across the samples. The analysis pipeline usually followed to build and analyze a co-expression begins with the collection of gene expression data, typically obtained through high-throughput techniques such as microarray or RNA-seq. Following data collection, the raw gene expression data were processed to eliminate low read counts, remove the batch effect, and normalize and transform the expression levels. Common normalization techniques include TMM (M. D. Robinson & Oshlack, 2010), DeSeq (Anders & Huber, 2010), and SCnorm (Bacher et al., 2017). Transformation methods include log transformation, variance-stabilizing transformation (VST), rank-based transformation, and Box-Cox transformation.

Subsequently, pairwise correlations between all pairs of genes are calculated, resulting in a correlation matrix where each entry represents the strength of the association between a pair of genes. Standard measures for this calculation include Pearson correlation, Spearman correlation, and mutual information. To construct the network, a threshold is applied to the correlation matrix to determine which gene pairs are significantly correlated. Afterward, the matrix is converted to a network where nodes represent genes, and edges represent the correlations between genes.

Once the network is constructed, various approaches for co-expression analysis are used to identify clusters or modules of co-expressed genes, i.e., hierarchical clustering, k-means, or more advanced community detection algorithms. Weighted gene co-expression network analysis (WGCNA) is a widely used tool for co-expression network inference and subsequent analyses, such as cluster identification. Compared to the

approaches that encode the gene co-expression using binary information (connected=1, unconnected=0), WGCNA applies a soft thresholding that assigns a connection weight to each gene pair (Langfelder & Horvath, 2008; B. Zhang & Horvath, 2005). These gene modules are often analyzed for functional enrichment using databases like Gene Ontology (GO) to understand the biological processes they may be involved in and to obtain biological insights (Chowdhury et al., 2019).

Transcriptional gene regulatory networks

Transcriptional gene regulatory networks describe the interactions between TFs and their target genes during the regulation of cellular gene transcription in response to intracellular and extracellular signals. A key element of gene expression regulation is the binding of TFs to the DNA binding sites in a target gene's promotor or regulatory regions (e.g., enhancers), to positively or negatively affect the transcriptional rate. The set of target genes regulated by a TF is usually defined as a regulon. Four different strategies to assemble GRNs can be defined: (1) curation of literature results based on reported experimental findings; (2) high-throughput measurements of TF-DNA binding such as chromatin immunoprecipitation (ChIP); (3) computational prediction of TF–target interactions using models of TF binding sites (TFBSs); (4) in silico inference from transcriptomic datasets. Garcia-Alonso et al. compared these methods for TF-gene inference and assembled DoRothEA, a comprehensive collection of TF-target interactions, each with a confidence level based on the number of supporting evidence (Garcia-Alonso et al., 2019). Recently, this resource was expanded using the information on TF–gene interactions from the CollecTRI (Collection of Transcription Regulation Interactions) meta-resource, and a GRN of 43175 signed TF-gene interactions for 1186 TFs was established (Müller-Dott et al., 2023).

The growing availability of single-cell transcriptomic datasets allows the inference of TF-gene at the single-cell level (Badia-i-Mompel et al., 2023). An increasing number of tools for GRN inference from single-cell expression data are being released. These methods are based on diverse mathematical and statistical methodologies (Kim et al., 2023). Pratapa et al. (2020) published BEELINE, a framework for benchmarking algorithms that infer GRNs from single-cell gene expression data. The authors compared 12 methods on over 400 simulated datasets and five experimental human or mouse single-cell RNA-seq datasets. Overall, they observed high heterogeneity in the results. The benchmarking analysis was limited to algorithms that did not require external knowledge for the GRN inference, such as SCENIC, a highly used method, which includes TF binding motif enrichment on gene promoter regions as a downstream pruning of the gene-gene interactions inferred from the data. Recently, tools that include multiple data modalities in the prediction have been developed. Among them, SCENIC+ (Bravo González-Blas et al., 2023), an extension of SCENIC, combines single-cell chromatin-accessibility from single-cell ATAC-seq (scATAC-seq) data with gene expression for the GRN inference.

Single-cell multi-omics data are also used by CellOracle, which, in addition to GRN inference, allows the performance on silico TF perturbations (Kamimoto et al., 2023).

Metabolic networks

A metabolic network encompasses the complete set of biochemical reactions occurring within a specific organism. These reactions are vital for sustaining life, facilitating processes such as energy production, biosynthesis of essential compounds, and waste elimination. These reactions are interconnected, forming complex pathways that convert nutrients into cellular building blocks and energy, supporting the organism's growth, development, and maintenance. Metabolic networks can be systematically encoded in a genome-scale metabolic model (GEM). GEMs are mathematical representations of an organism's metabolism, mapping the relationships between genes, enzymes, reactions, and metabolites. GEMs provide gene–reaction–metabolite connections via two primary matrices. The metabolite-reaction matrix links metabolites to the biochemical reactions they participate in, detailing the substrates and products involved in each reaction, and the reaction-enzyme-gene matrix connects biochemical reactions to the enzymes that catalyze them and the genes encoding these enzymes. This linkage is crucial for understanding how genetic variations can influence metabolic pathways. In GEMs, the stoichiometry of each metabolic reaction is defined. This includes specifying the precise quantities of reactants and products, the usage of cofactors (molecules that assist enzymes), and the dependency on particular enzymes. These models have been developed for various organisms, including microorganisms, plants, and mammalian systems like mice and humans. (Nielsen, 2017; H. Wang et al., 2021)

In recent years, the reconstruction of human metabolic networks has seen significant progress. These efforts aim to include increased metabolic reactions, thus providing a more comprehensive representation of human metabolism. This expanded coverage helps understand the complexity of metabolic processes in human health and disease (Brunk et al., 2018; J. L. Robinson et al., 2020; Thiele et al., 2013). Generic GEMs can be tailored to specific tissues, organs, and cell types by integrating gene expression data. This customization enhances the accuracy of the models, making them more relevant for studying physiological and pathological conditions (Mardinoglu et al., 2014).

One of the primary methods for studying metabolic networks is flux balance analysis (FBA). FBA is a computational approach used to predict the flow of metabolites through a metabolic network by optimizing a specific objective function, such as maximizing biomass production or ATP generation. By simulating different conditions and genetic modifications, FBA can provide insights into metabolic capabilities and constraints, helping identify potential therapeutic intervention targets (Orth et al., 2010). Metabolic networks and their representation through GEMs are essential for understanding the biochemical foundations of life. They provide a framework for exploring genetic and environmental influences on metabolism, offering insights for both basic research and applied sciences such as biotechnology and medicine.

Human Network Collections

In recent years, extensive efforts have been made to build repositories of biological networks merging interactions from several sources. OmniPath, a database of molecular biology prior knowledge, integrates more than 100 resources combining signaling pathways, protein-protein interactions, and transcriptional regulation networks (Türei et al., 2016). It provides a unified network framework of pathways and annotations for studying cellular processes. Similarly, the HIPPIE (Human Integrated Protein-Protein Interaction rEference) database combines multiple experimental PPI datasets with a normalized scoring scheme that reflects the amount and quality of evidence for every PPI (Alanis-Lobato et al., 2017). BioGRID (Biological General Repository for Interaction Datasets) offers a vast repository of curated interactions from multiple species, supporting biological and medical research (Oughtred et al., 2021). STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database integrates known and predicted protein-protein interactions from various sources, facilitating the exploration of functional associations between proteins (Szklarczyk et al., 2021). Those are the four main repositories, and, combined with all the others, they provide robust platforms for the systematic analysis and interpretation of biological networks, driving forward the fields of systems biology and network medicine (Dimitrakopoulos et al., 2022).

Mathematical and Computational Methods for Networks Analysis

The field of network analysis is characterized by the mathematical and statistical computation of network properties, which provide insights into the structural and behavioral characteristics of the network (A.-L. Barabási, 2016). Among the network properties, centrality measures play a pivotal role in determining the most influential nodes within a network (Pavlopoulos et al., 2011). One of the most straightforward measures of centrality is the node degree, which is defined as the number of interacting partners of a node. Other well-known centrality measures are betweenness centrality, closeness centrality, and eigenvector centrality (Ashtiani et al., 2018; Koschützki & Schreiber, 2008; Zitnik et al., 2024). A key challenge in network analysis is identifying highly interconnected nodes, typically referred to as modules or communities. A network module represents a locally dense neighborhood within a network, where nodes have a higher probability of connecting to other nodes within the same neighborhood than to those outside of it. In biological networks, this means that nodes encoding molecules involved in the same biological process are often highly interconnected, thus forming modules that represent functional units. These modules are essential for understanding biological processes, as they often correspond to groups of molecules that work together to perform specific tasks within a cell. Over the years, numerous algorithms have been developed to identify network modules. Among these, kernel clustering, modularity optimization, random-walk-based methods, and local methods have emerged as the most effective. Each of these algorithms has its strengths and weaknesses, and their performance depends on the network characteristics. For instance, kernel clustering is

known for its ability to handle complex, non-linear relationships between nodes, while modularity optimization focuses on maximizing the modularity score, which measures the quality of the division of a network into modules (Newman & Girvan, 2004). Instead, random-walk-based methods like Walktrap or Infomap, simulate the movement of a random walker through the network to detect modules (Pons & Latapy, 2006; Rosvall et al., 2009). Finally, local methods like SPICi focus on the network topology to identify densely connected regions (Jiang & Singh, 2010). These algorithms were highlighted as top performers in the Disease Module Identification DREAM Challenge. However, the challenge also finds that there is no single best-performing method across all types of networks, indicating that the algorithm choosing depends on the specific network and research question (Choobdar et al., 2019).

Applications of Network Analysis in Biology

The analysis of biological networks has numerous applications, ranging from studying the molecular mechanism involved in human diseases to identifying DR candidates (Z. Yu et al., 2024). One of the primary applications of biological networks is the identification and prioritization of disease-causing candidate genes (Janyasupab et al., 2021; Jha et al., 2020; Solano Noriega et al., 2021; Stolfi et al., 2023; Tang et al., 2021; Y. Zhang et al., 2020). Researchers can identify key genes or proteins potentially involved in disease pathogenesis through network analysis, such as centrality measures, community detection, or network propagation algorithms (Cowen et al., 2017). For example, Nazarieh & Helms (2019) mixed the topological features from TF-miRNA co-regulatory network and the biological factors for ranking the gene candidates for breast invasive carcinoma and liver hepatocellular carcinoma. Instead, Yepes et al. (2020) applied a network propagation method to evaluate the similarity between candidate genes and known disease genes in PPI networks. With this approach, they identified five candidate genes for germline cutaneous melanoma susceptibility from whole-exome sequencing analysis of 98 cutaneous melanoma patients. Furthermore, network analysis is used to identify disease modules, subnetwork enriched in genes associated with specific diseases. Researchers detect modules that show a significant associations with disease phenotypes, thereby enhancing the understanding of molecular mechanisms involved in these pathologies (Badam et al., 2021; Choobdar et al., 2019; R. S. Wang & Loscalzo, 2018). The identification of disease modules provide insights into the molecular mechanisms underlying diseases and may also reveal potential therapeutic targets (A. L. Barabási et al., 2010). Target identification represents one of the first network analysis applications for drug discovery. Module detection methods were used to identify potential therapeutic targets for complex diseases, cancer (Hossain et al., 2021), and Alzheimer's disease (Parolo et al., 2023; Peng et al., 2020). Other researchers applied network analysis methods to identify drug targets in complex diseases. For example, J. Wang et al. (2021) proposed a pipeline that combines network analysis and greedy articulation point removal to identify therapeutic targets for vitiligo, successfully identifying 44 potential targets. Network analysis is also used to predict new therapeutic

effects of existing drugs (Ashburn & Thor, 2004; Z. Wu et al., 2018). The random walk with restart (RWR) algorithm has been extensively utilized in drug repositioning (Martínez et al., 2015; Su et al., 2023; Y. Wang et al., 2019). In 2023, Su et al. proposed the Drug Target Set Enrichment Analysis method (DTSEA), which employs the RWR in the gene functional interaction network. The methods also apply enrichment analysis based on the disease-gene associations and drug–target interactions for drug repositioning. Su and colleagues identified 50 potential drug candidates for COVID-19, 18 of which were evaluated in clinical trials (Su et al., 2023). Furthermore, several network-based methods, including network-based inference and path-based methods, were employed in the drug repositioning of various diseases, including cancer (Vitali et al., 2016), type 2 diabetes (Teng et al., 2022), Alzheimer's disease (Fang et al., 2021; Parolo et al., 2023), and metabolic syndrome (Misselbeck et al., 2019). Misselbeck et al., in a study performed at Fondazione COSBI, proposed a systems biology approach to identify DR candidates. The authors combined topological and functional similarities in a proximity score to describe the interaction between drugs and pathways and suggested a potential use of the BTK inhibitor ibrutinib as a pharmacological treatment for metabolic syndrome.

Linear Programming And Optimization Method

The previous section explored how various network-based models can analyze and describe biological data. Although these methods have effectively revealed insights into biological mechanisms, they do not provide a complete understanding of the underlying molecular mechanisms of disease. To address this limitation and achieve a more comprehensive understanding of biological mechanisms, optimization-based modeling techniques can be employed to extract a more detailed mechanistic representation of signaling processes and the dysregulation of signaling pathways. These models integrate biological data, such as expressions and prior knowledge of biological processes represented as networks, to identify the optimal solution that describes feasible molecular mechanisms. Perturbations introduced to the system, such as drug treatment or stimuli, are used as input, and the resulting changes in expression observed in the cells are represented as output. Several formalization and optimization models exist for parameter estimation of signaling system problems; in this thesis, we employed Integer Linear Programming (ILP).

Linear Programming and Integer Linear Programming.

Linear programming (LP) studies algorithms and techniques for mathematical optimization problems. A problem is called linear if both the objective function and the constraints are linear functions. LP optimization techniques aim to maximize or minimize a linear objective function subject to a set of linear equality or inequality constraints. LP problem can be defined as follows:

$$\text{Objective function: } max \text{ (or min) } c^T x, \qquad c \in R^n$$

$$\text{Subject to: } Ax \leq (or \geq) b, \qquad x \in R^n, b \in R^m$$

Where:

- c is a vector of coefficients for the objective function;
- x is the vector of decision variables;
- A is a $m \times n$ matrix representing the coefficients of the constraints;
- b is a vector representing those constraints' upper (or lower) bounds.

The number of rows and columns of A are equal to the number of constraints and variables, respectively. In standard LP problems, the x variables are allowed to take continuous values, while adding integer restrictions to constraint transforms it into an Integer Linear Programming problem (ILP). Formally, the ILP problem constraints can be written as:

$$\text{Objective function: } max \text{ (or min) } c^T x, \qquad c \in R^n$$

$$\text{Subject to: } Ax \leq (or \geq) b, \qquad x \in Z^n, b \in R^m$$

All the linear problems that accept both integer and continuous variables are called Mixed Integer Linear Programming (MILP). In the context of this thesis, the introduction of integer constraints is necessary due to the structure of the prior knowledge of biological processes, which are represented as directed and signed graphs. As a result of the integer restrictions, ILP problems are typically more challenging to solve than LP problems, as the solution space is discrete rather than continuous, leading to a significant increase in problem complexity through combinatorial explosion.

Application of Linear programming in system biology and drug discovery

LP has become increasingly important in systems biology in recent years, offering effective solutions to a wide range of complex optimization challenges within biological systems. In particular, ILP and MILP models have been shown to be effective in addressing optimisation problems, notably within the context of drug discovery and development.

LP has been demonstrated to be an effective approach in metabolic network analysis, particularly in the context of improving Flux Balance Analysis (FBA). The application of LP is crucial for optimizing metabolic pathways through the identification of gene knockouts or overexpressions that aim to maximize the production of target metabolites. LP plays a key role in identifying core gene sets that are essential for specific metabolic functions by effectively modeling gene-protein-reaction relationships. This, in turn, provides valuable guidance for experimental designs aimed at strain optimization. In a recent study, Tamura et al. (2023) introduced gDel_minRN, a MILP-based algorithm designed to efficiently identify gene deletion strategies for growth-coupled production in metabolic

networks. This approach reduces the number of reactions and facilitates the stoichiometrically feasible production of target metabolites.

Additionally, LP methods have been employed to refine genome-scale metabolic models by integrating condition-specific experimental data, such as transcriptomics or proteomics, enabling the development of context-specific models (Jensen et al., 2018; Richelle et al., 2019; Röhl & Bockmayr, 2017). Walakira et al. (2021) presents a framework for efficient GEMs extraction, with a particular focus on the selection of model extraction method (MEM) choice between different MILP and ILP-based methods. Ozen et al. (2022) proposes a network learning approach using ILP, which is useful for reducing the gap between the curated networks and experimental data, by trying to cope with the incompleteness of literature/resources on molecular pathways. ILP has also been used to optimize metabolic networks for drug discovery by identifying biological targets that can be manipulated to produce the desired effect with minimal side effects (Qiu et al., 2018).

In network reconstruction and pathway inference, ILP has been applied to signaling pathway reconstruction, where contributes to the identification of the most plausible pathway structures that are consistent with the observed proteomic and phosphorylation data. Ji et al. (2017) predict the efficacy of drugs by integrating ILP with biological data, such as phosphoproteomics, to infer pathways specific to certain cancer cells, helping identify potential treatment compounds. ILP plays a role in gene regulatory network inference, helping to uncover transcription factors and regulatory interactions that explain gene expression profiles (Liu et al., 2019; Melas et al., 2015). This application has been particularly useful when analyzing time-series gene expression data or gene perturbation experiments (Gjerga et al., 2021). Similarly, ILP has been effectively used to reconstruct signaling pathways from protein-protein interactions. LocPL improves the automatic reconstruction of signaling pathways from protein-protein interactions by incorporating protein localization information (Youssef et al., 2019)

Machine Learning and Deep Learning Models

Overview

Machine learning (ML) is a family of powerful approaches that have revolutionized various fields, including DR and drug development. These techniques get advantages from large datasets and innovative algorithms to identify patterns, make predictions, and assist in discovering and developing new drugs or finding new uses for existing ones.

Between ML paradigms, we can find supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning develops models that predict future values of data categories or continuous variables by training on known input and output data

relationships. In contrast, unsupervised learning is used for exploratory purposes, developing models that enable data clustering without user-specified categories. Supervised learning trains a model to predict future outputs from new inputs, whereas unsupervised learning discovers hidden patterns and intrinsic structures in input data to cluster it meaningfully. Reinforcement learning models differ from the supervised and unsupervised in that they address sequential decision problems. In these problems, the action to be taken at a given moment depends on the current state of the system, which in turn affects the subsequent state of the system. The quality of an action is quantified by a numerical value representing the "reward", inspired by the concept of reinforcement and designed to encourage correct behaviors. Finally, semi-supervised learning is a machine learning paradigm that combines a small amount of human-labeled data (usually found in supervised learning) with a large amount of unlabeled data (typical of unsupervised learning). The Positive Unlabeled (PU) problem is a form of semi-supervised learning in which a binary classifier is trained to learn only from positive and unlabeled occurrences. The PU learning scenario assumes the availability of two sample sets for training: the positive set P and a mixed set U, which contain both positive and negative samples without being labeled. These problems differ from other semi-supervised learning, where a labeled set contains examples of both classes.

Deep learning (DL) is a family of ML algorithms that employ neural networks to learn from data. Neural networks consist of a hierarchy of layers that are composed of neurons. A neuron can be represented as a weighted mean of the input signals whose result is then passed to a nonlinear activation function. The most common architectures are Feedforward Networks (FNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Autoencoder (AE), but more complex architectures can be obtained by arbitrarily combining them. In this dissertation, feedforward network architectures will be applied. The FNN is one of the simplest artificial neural networks characterized by a single direction of information flow. In these architectures, input features are processed through an input layer, passing through several nonlinear transformations in hidden layers (if present), and finish in an output layer that generates predictions. Unlike other types of neural networks, such as RNNs, feedforward neural networks do not contain cycles or loops. As previously outlined, the FNNs are composed of three types of layers (input, hidden, and output layer) interconnected by weights. The input layer is responsible for receiving and transmitting the input data to the subsequent layer. The number of neurons in the input layer depends on the input data dimensions. Each neuron in a hidden layer takes the weighted sum of the outputs from the previous layer, applies an activation function, and passes the result to the next layer. The output layer is the final layer and is responsible for producing the output given a set of inputs. The number of neurons in the last layer is determined by the number of possible outputs the network is designed to produce. A neural network is defined as a fully connected network when every neuron in a single layer is connected to every neuron in the subsequent layer. The connection between neurons can be quantified by weights. The learning process involves

updating these weights in order to reduce the error in the output. The feedforward neural network training process comprises the feedforward and the backpropagation phases. In the feedforward phase, the input data is fed into the network, and it propagates forward through the network until the output layer provides the prediction. The weighted sum of the inputs is calculated and passed through an activation function at each hidden layer. Activation functions have a key role since they introduce non-linearity into the network, allowing the model to learn more complex patterns. Common activation functions include the sigmoid, tanh, and ReLU (Rectified Linear Unit). In the backpropagation phase, the difference between the prediction and the ground truth (i.e., error) is calculated and propagated back into the network. The weight is adjusted using a gradient descent optimization algorithm to minimize the error. The combination of these two phases is repeated iteratively, passing the data set through the network several times and updating the weights to reduce the error in the prediction. This process continues until the network performs satisfactorily on the training data, and is known as *gradient descent*. While using FNNs is widely acknowledged as an efficient and powerful approach, it is important to recognize their limitations. The number of hidden layers and neurons in each layer significantly impacts the network's performance. Therefore, it is crucial to properly set the hyperparameters and design the training process in order to prevent overfitting.

Applications of ML and DL Models in Drug Discovery

The application of ML and DL is a valuable tool for the drug discovery process. All phases of drug development nowadays utilize ML and DL algorithms to identify new drug targets (J. Jeon et al., 2014; Y. Bin Wang et al., 2020; Wen et al., 2017), provide robust evidence for target-disease associations (Ferrero et al., 2017; M. Jeon et al., 2021), optimize small molecule design (Riniker et al., 2014), and gain insights into disease mechanisms (Alipanahi et al., 2015; Godinez et al., 2017). The application of ML and DL has the potential to advance our understanding of disease mechanisms, enhance our comprehension of disease and non-disease phenotypes, facilitate the identification of novel biomarkers for clinical prognosis and drug efficacy (Mamoshina et al., 2018; Rajpal et al., 2021; Tognetti et al., 2022), and enhance digital pathology imaging (Litjens et al., 2016). Furthermore, it can facilitate the extraction of high-content information from images at all levels of resolution (Aggarwal et al., 2021; Olsen et al., 2018).

2. An omics-driven computational workflow for drug target identification in rare diseases: application to cystinosis.

Introduction

In recent years, network-based computational methods have emerged as a valuable approach for integrating the heterogeneous biological knowledge extracted from various sources. Network analysis is a highly flexible approach for modeling the relationships among molecular entities such as genes, proteins, metabolites, and other broader biomedical concepts such as diseases, phenotypes, and pathways. A main advantage of network analysis is the study of the molecules of interest in their functional contest, beyond the single-molecule dysregulations observed using traditional approaches for the analysis of omics data. (A. L. Barabási et al., 2010; Choobdar et al., 2019; M. M. Li et al., 2022; Menche et al., 2015) For example, the identification of disease modules, i.e., portions of the network enriched in genes associated with a trait of interest, has proven to be particularly useful to study the disease-associated genes in their context and identify neighboring genes that may be more suitable as drug targets than the directly associated ones (Sadegh et al., 2021).

We and others have applied network analysis and disease module identification to identify DR candidates, i.e., new indications of existing drugs (Cheng et al., 2018; McGowan et al., 2023; Misselbeck et al., 2019; Morselli Gysi et al., 2021; Parolo et al., 2023; Sadegh et al., 2021). In addition to its applications to common, multifactorial diseases, network analysis has also been tested in studying rare diseases. DR represents a crucial strategy for the treatment of rare diseases. Rare diseases affect a small percentage of the population, making traditional drug development an economically challenging business due to the limited market. Repurposing existing drugs represents a cost-effective and time-efficient strategy for addressing the unmet medical needs of patients suffering from rare diseases (Brewer, 2009; Coté et al., 2010; Sardana et al., 2011). By repurposing drugs, pharmaceutical companies can reduce the investment required for de novo drug discovery, thereby making it financially feasible to develop treatments for small patient populations. Furthermore, since repurposed drugs have already passed the safety testing, the time to market is considerably reduced (Nosengo, 2016; Roessler et al., 2021). These approaches are particularly beneficial for rare disease patients, who often have limited treatment options and urgent medical needs.

Given the low prevalence of each rare disease, data availability is very limited, making it challenging to study genetic causes and molecular mechanisms and thus identify effective therapeutic solutions (Zhu et al., 2021). To aid in the diagnosis of rare diseases, a recent study used a DL framework based on a knowledge graph that integrated

information from multiple biological databases to identify candidate causative genes (Alsentzer et al., 2022). Moreover, it has been shown that network biology concepts usually applied to complex diseases, such as functional module identification, can be effectively extended to the rare disease field (Buphamalai et al., 2021). Notably, a few network-based repurposing applications for rare diseases have been proposed, supporting the approach's feasibility despite the paucity of data (McGowan et al., 2023; Sosa et al., 2020). In this study, we focused on Cystinosis, and we devised a network-based approach to study the molecular mechanisms underlying the disease and identify new therapeutic opportunities.

Cystinosis

Cystinosis is a rare, inherited metabolic disorder that leads to the accumulation of the amino acid cystine within the lysosomes of cells. The disorder is genetically inherited in an autosomal recessive pattern due to mutations in the CTNS gene, which encodes the lysosomal cystine transporter cystinosin. Cystinosin is critical for transporting cystine out of the lysosome; thus, mutations in this gene result in lysosomal cystine accumulation (Elmonem et al., 2016).

The molecular mechanisms that cause the clinical symptoms are not fully understood, even if new studies have started to shed light on possible causative mechanisms linking CTNS defects to the clinical phenotype. The epithelial cells lining the kidney's proximal tubule generate cystine via lysosomal degradation of disulfide-rich proteins that reabsorb from the glomerular filtrate through endocytosis. In the absence of a functional cystinosin protein, cystine accumulates and crystallizes in the lysosomal lumen of the proximal tubule cells. We recently reported that CTNS dysfunction leads to defective autophagic clearance of damaged mitochondria, generating oxidative stress that leads to cell proliferation and loss of epithelial reabsorptive activity. One of the key regulators of this process is mTORC1, which was shown to be constitutively active in proximal tubule cells with mutated CTNS, leading the cells to an anabolic proliferative state (Berquez et al., 2023; Festa et al., 2018; Luciani & Devuyst, 2024).

The clinical presentation of cystinosis is heterogeneous and can be classified into three primary forms based on the age of onset and the severity of symptoms: nephropathic, intermediate, and non-nephropathic. Nephropathic cystinosis, the most severe and prevalent form, manifests within the first year of life and is characterized by renal tubular dysfunction. This form leads to progressive renal failure, necessitating interventions such as electrolyte supplementation and renal transplantation. Intermediate cystinosis presents later, during childhood or adolescence, with a similar but less severe clinical progression. Non-nephropathic cystinosis typically presents in adulthood, primarily affecting the eyes, where cystine crystal deposition leads to photophobia and vision loss (Gahl et al., 2002). The overall incidence rates observed were between 1:115,000 and 1:260,000 live births. The highest incidence is observed in individuals of the Pakistani ethnic group in the West Midlands, UK (1:3,600 live births), and in populations with

detected founder mutations in the province of Brittany, France (1:26,000 live births)(Elmonem et al., 2016).

The diagnosis of cystinosis involves a combination of clinical evaluation, laboratory testing, and genetic analysis. Clinicians look for hallmark symptoms such as growth retardation, renal tubular dysfunction, and ocular abnormalities. Laboratory tests confirming elevated cystine levels in leukocytes are essential for diagnosis, and genetic testing of the CTNS gene can provide definitive confirmation.

Current cystinosis treatments aim to manage symptoms and slow disease progression. Symptomatic treatments address specific complications, such as hypothyroidism, diabetes, and growth failure. Cysteamine therapy, which reduces cystine levels within cells, is the only approved therapy for cystinosis. It has been shown to delay renal deterioration and other systemic manifestations. The prognosis for individuals with cystinosis has improved significantly with early diagnosis and the advent of cysteamine therapy, but despite treatments, there are currently no drugs that can prevent kidney failure. Consequently, renal transplantation remains an essential therapeutic intervention for patients with advanced forms of cystinosis-related kidney disease (Emma et al., 2014) and there is an urgent need to develop disease-modifying therapies (Medic et al., 2017).

Our Contribution: the phenotype-aware framework for DR

Here, we developed a phenotype-aware computational framework aimed at identifying therapeutic targets and repurposing opportunities for cystinosis, a rare genetic disorder. As illustrated in Figure 1, the framework is structured into three distinct blocks. The first block leverages a deep learning (DL) framework to predict candidate genes involved in proximal tubule function, a critical site of pathology in cystinosis. These predicted genes are then used to refine and restrict the search space in the second and third blocks, which focus on network-based strategies for identifying candidate drug targets and potential repurposing opportunities. The aim of this framework is twofold: methodological and biological. From a biological standpoint, the objective was to discover a signature or drug target specific to cystinosis that could guide the repurposing of existing drugs and lead to effective therapeutic interventions. However, equally important is the methodological aim of the framework, which is grounded in several key principle. First, the framework is designed to be tissue-specific, with a particular focus on the biological processes of the proximal tubule in the kidney, thereby ensuring that the identified targets are relevant to the pathophysiology of cystinosis. In addition, the framework integrates multiple types of data, including mutiomics data, ontology and drug repository information, to provide a more comprehensive and nuanced understanding of the disease. This integration of diverse data types allows to generate more complete biological insight. Moreover, the framework emphasizes a data-driven approach, relying heavily on machine learning techniques to extract meaningful insights from large datasets. This minimizes the need for manual intervention and reduces potential biases. The methodological goal also takes

into account the challenges posed by rare diseases, such as cystinosis, where limited data availability is a significant problem. To visualize the results interactively, we provide a web application at https://apps.cosbi.eu/awt234/.

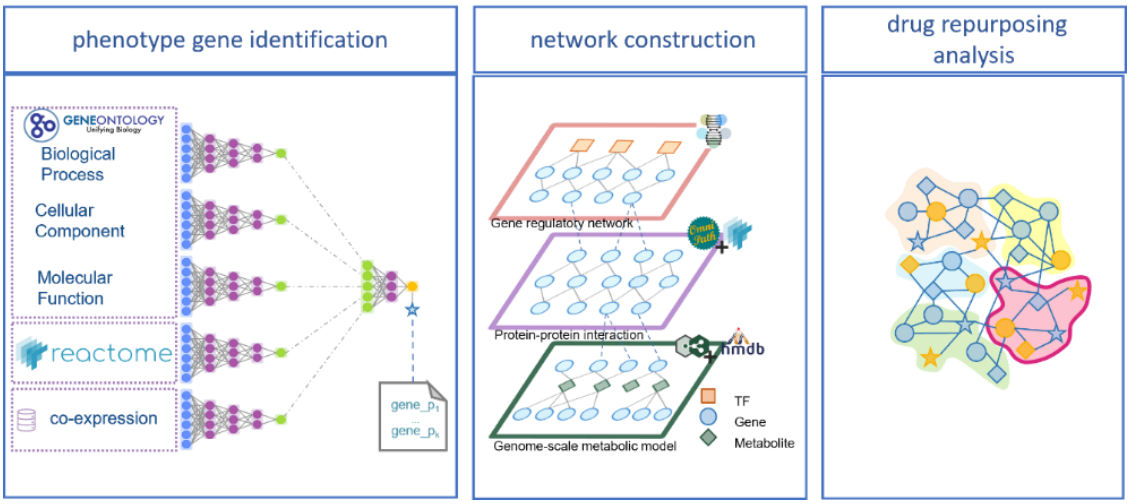


Figure 1 - Overview of the phenotype-aware framework. The pipeline is divided into three blocks. The first block is the ANN architecture, responsible for predicted candidate genes involved in the proximal tubule function. The second block illustrates the assembly of the tissue-specific multilayer network. The third block represents the network analysis for the drug repurposing identification.

Material and Methods

Resources and Background

The following section provides an overview of the resources employed in the development of the pipeline.

Ontologies

Human Phenotype Ontology

The Human Phenotype Ontology (HPO) (Gargano et al., 2024) is a project initiated in 2007 that provides an ontology of medically relevant phenotypes and disease-phenotype annotations. The ontology consists of a series of concepts (terms) and relationships between them. The HPO currently contains over 13,000 terms arranged in a directed acyclic graph (DAG). All the connections within the HPO consist of "is-a" relationships, meaning they are direct associations between classes and their subclasses. A term represents a more specific or limited instance of its parent term(s). Each term within the HPO describes a clinical abnormality.

Mammalian Phenotype Ontology

As part of the Mouse Genome Database (MGD) Project of Mouse Genome Informatics (MGI) portal, Smith & Eppig (2009) developed The Mammalian Phenotype Ontology (MPO). As HPO, MPO is a standardized structured vocabulary organized in a DAG where

each child term refers to a more specific condition. MGI also provides the annotations of the association between mouse phenotypes and genotypes, supported by one or more references. The genotype comprises one or more pairs of mutant alleles combined with associated genetic background strain information.

The Gene Ontology

The Gene Ontology (GO) was established in 1998 by a consortium of researchers studying the genomes of three model organisms. The consortium aimed to develop a complete computational model of biological systems (The Gene Ontology Consortium, 2023). The Gene Ontology Resource is the de facto standard for gene annotations. The GO Ontology includes three orthogonal ontologies structured as DAG. The first ontology describes the molecular activities of individual gene products (Molecular Function, MF). The second one defines the localization of gene products (Cellular Component, CC). The last ontology represents larger processes performed by multiple molecular activities (Biological Process, BP). GO annotations are created by associating genes and gene products to a GO term. When the GO terms related to two genes show semantic similarity, it is possible to deduce functional similarity between them. Therefore, numerous methodologies relying on semantic similarity were developed to estimate gene functional similarity. The DAG structure of GO, which facilitates quantitative semantic comparisons, has led to the development of several methods for measuring semantic similarity between GO terms; among the most widely used, we select methods involving graph-based measures. J. Z. Wang et al. (2007) introduced a graph-based approach that estimates similarity using the GO graph topology. The Wang metric decodes the semantics of GO terms into numerical values, considering the specificity of GO terms based on their positions within the GO graph structure and diverse semantic contributions from different relations. Genes can be associated with multiple GO terms. The scores derived from the multiple GO terms associated with each gene must be combined to assess the semantic similarity between genes. We used the Best-Match Average (BMA) (Schlicker et al., 2006) method to aggregate these scores. The BMA score utilizes the Best-Match Average strategy to determine similarity by computing the average of the maximum similarity values found in each row and column of the data. Given two genes, g_1 and g_2, annotated respectively in m and n GO terms, and defined the similarity between two GO terms as s_{ij}, the BMA aggregation method is defined as follows:

$$Sim_{BMA}(g_1, g_2) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} s_{ij} + \sum_{j=1}^{n} \max_{1 \le i \le m} s_{ij}}{m + n}$$

Annotation resources

ORPHANET

ORPHANET is a database dedicated to rare diseases and orphan drugs (Pavan et al., 2017). It offers information and resources for diagnosis, treatment, and research. It also provides detailed data on genes linked to rare genetic disorders, including their

nomenclature, type, chromosomal location, and cross-references with other databases. It categorizes genes based on their role in disease, like causative, modifiers, susceptibility factors, or phenotype contributors.

OMIM

OMIM (Amberger et al., 2019), Online Mendelian Inheritance in Man, is a comprehensive compendium of human genes and genetic phenotypes, initiated in the early 1960s and freely available. It contains information on all known mendelian disorders and over 16,000 genes and focuses on the relationship between phenotype and genotype.

GWAS Catalog

The GWAS Catalog, established in 2008 by the National Human Genome Research Institute (NHGRI), addresses a significant challenge in genetics: the accessibility of GWAS. The Catalog provides a complete and curated database of SNP-trait associations, facilitating global access for researchers and clinicians. The curators identify and evaluate the eligible GWA studies published through a literature search. They then extract the reported traits, significant SNP-trait associations, and sample metadata. Data are released weekly, and eligible studies are included within approximately one to two months after publication. In 2020, they will also accept submissions of unpublished GWAS data. The Catalog is a collaborative effort between EMBL-EBI and NHGRI, led by experienced geneticists and molecular biologists.

MitoCarta

MitoCarta (Rath et al., 2021) is a database that provides a comprehensive catalog of the proteins localized to the mitochondria in different organisms. It was developed for the first time in 2009 through a collaborative effort of researchers from the Broad Institute of MIT and Harvard University. The latest version, MitoCarta3.0, released in 2020, includes 1136 human and 1140 mouse genes encoding mitochondrially localized proteins, with submitochondrial compartment and pathway annotations. The curators performed mass spectrometry on mitochondria from fourteen tissues, checked protein locations with GFP tagging, and combined these findings with six other datasets on mitochondrial localization using Bayesian analysis. They manually revised the previous MitoCarta2.0, adding 78 genes, removing 100 genes, and providing annotations of sub-mitochondrial localization.

Network and pathways resources

Biological pathways represent a series of biochemical reactions within a cell to perform specific biological functions. These pathways are fundamental for understanding how genes, proteins, and other cellular components interact within the cell. Each pathway is associated with the specific functions it performs within the cell. Numerous databases collect and document the relationships between genes and these pathways, providing valuable insight into how genes function and interact in cellular processes. Biological pathways are also useful resources to measure similarities between genes.

Reactome

The Reactome project was founded in 2003, and it offers a freely accessible, open-source relational database containing signaling and metabolic molecules organized into biological pathways. Reactome primarily focuses on the reactions occurring within biological systems. These reactions involve a variety of entities, including nucleic acids, proteins, protein complexes, vaccines, therapeutic agents, and small molecules. These entities are categorized into pathways, which represent areas of biological interest, including intermediary metabolism, signaling, transcriptional control, apoptosis, and disease (Milacic et al., 2024).

KEGG

The KEGG database project was initiated in 1995 under the Japanese Human Genome Project, with the aim of producing a comprehensive resource that would enhance comprehension of biological systems through genomic sequence data. KEGG defines biological systems as molecular networks, which represent an invaluable tool for analyzing large biological datasets. The KEGG pathway is constructed through a methodical process of capturing and organizing experimental insights from published literature (Jin et al., 2023).

WikiPathways

WikiPathways is a platform dedicated to the curation of biological pathways. It is presented as a model for a pathway database that will enhance and complement ongoing efforts such as KEGG and Reactome. What distinguishes WikiPathways from these other databases is its open and public approach, allowing for broader community participation. This approach also shifts most of the peer-reviewing, editing, and maintaining to the community. WikiPathways contains 1913 human-edited pathways for 27 species, of which 866 describe human biology (Agrawal et al., 2024).

Omnipath

The Omnipath database, developed in the Saez and Korcsmaros labs, comprises molecular biology prior knowledge and encompasses five databases, each integrating data from multiple original resources (Türei et al., 2016). The OmniPath database includes nine network datasets constructed from 61 resources for protein-protein interactions.

DoRothEA

The DoRothEA database (Garcia-Alonso et al., 2019) contains signed TF - target gene interactions. This database encompasses 1,395 TFs that regulate 20,244 genes through 486,676 unique interactions. An empirical confidence level is also assigned to each TF based on the amount of supportive evidence for that TF/interaction. The confidence levels range from A (high quality) to E (low quality).

Metabolic Atlas

Metabolic Atlas is an open-source genome-scale metabolic models (GEMs) collection aimed to facilitate easy browsing and analysis of these models (J. L. Robinson et al., 2020). The project, headed by Professor Jens Nielsen at Chalmers University of Technology, systematically gathers and curates detailed information about metabolic processes from seven different species. Here, the analysis focused on human metabolism.

Human Metabolome Database

The Human Metabolome Database (HMDB) is an open-access and comprehensive repository of small molecule metabolites (Wishart et al., 2022). Established in 2007 by Dr. David Wishart and his team at the University of Alberta, the HMDB provides detailed data on the association between 220,945 metabolites and 8,610 transporters and enzymes, including their respective structures, functions and associated diseases. The database collects data from various sources, including scientific literature, metabolomic studies, and other databases, ensuring a comprehensive coverage of metabolites. To maintain accuracy and relevance, the information undergoes meticulous curation and annotation by experts in the field.

Identification of Genes Associated with Kidney Proximal Tubule Function Through Deep Learning

Identification of Literature-Derived Renal Tubule Phenotype Genes

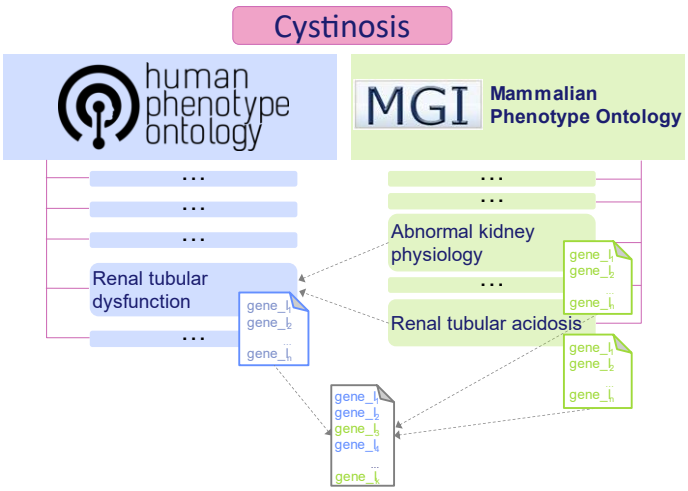


Figure 2 - Scheme of the phenotype genes extraction process from the literature.

Pre-processing of Human Phenotype Ontology Data

HPO in OBO format was retrieved from the ontology repository [1]. This file contains the full set of HPO terms and their parent-child relationship. From the HPO ontology, we selected the Renal tubular dysfunction (HP:0000124) phenotype and kept all their descendant terms for 12 phenotypes. To identify the genes associated with each HPO term, we used two HPO annotation resources: (A) a table of phenotype-gene associations[2] and (B)a table of phenotype-disease associations[3]. The *phentoype_to_genes.txt* file (A) provides the link between HPO terms and genes. HPO curators assign HPO terms associated with any disease caused by mutations in a gene to that gene. Instead, the phenotype.hpoa file (B) contains the phenotype-disease associations obtained by HPO combining ORPHANET and OMIM data. To identify the causative genes of the diseases derived from ORPHANET, we leveraged the table with the information on the gene-rare disease associations from the Orphadata portal[4]. We used the OMIM dataset assembled by Ehrhart et al. (2021), which includes 4166 rare monogenic diseases linked to 3163 causative genes[5]. The gene-disease associations from these two databases were then merged with the disease-phenotype associations in *phenotype.hpoa* (B), in order to have a list of associations between genes and phenotype. This list is then appended to the information in the *phentoype_to_genes.txt* file (A) to obtain all the associations between genes and phenotypes from human data.

Pre-processing of Mammalian Phenotype Ontology Data

To supplement the training dataset toward a more robust neural network training, we included data on associations between genes and phenotype deriving from the MPO. MPO and the corresponding annotation data were retrieved from the MGI database[6]. To identify phenotype-gene associations of interest, we downloaded the *HMD_HumanPhenotype.rpt, MGI_PhenoGenoMP.rpt, VOC_MammalianPhenotype.rpt,* and *MPheno_OBO.ontology* files. *HMD_HumanPhenotype.rpt* contains the MGI ID (MGI_Marker_Accession_ID) and the links between the mouse gene, the homologous human gene, and the high-level mouse phenotype. *MGI_PhenoGenoMP.rpt*, on the other hand, contains the MGI ID and the information relative to the specific mouse phenotype. *VOC_MammalianPhenotype.rpt* contains the specific mouse phenotype description. The information from the three files was merged to obtain a list of human genes linked to specific mouse phenotypes. The file *MPheno_OBO.ontology* contains the full set of MPO terms and their parent-child relationship. We use this ontology to filter the list of

[1] http://purl.obolibrary.org/obo/hp.obo downloaded on 25/06/2021
[2] http://purl.obolibrary.org/obo/hp/hpoa/phenotype_to_genes.txt downloaded on 25/06/2021
[3] http://purl.obolibrary.org/obo/hp/hpoa/phenotype.hpoa downloaded on 25/06/2021
[4] http://www.orphadata.org/cgi-bin/index.php downloaded on 28/06/2021
[5] https://figshare.com/articles/dataset/Gene-RD-Provenance_V2_1_txt/13373750 downloaded on 26/07/2021
[6] http://www.informatics.jax.org/ downloaded on 18/10/2021

associations, keeping only links between human genes and child phenotypes of Abnormal Renal Physiology (ARP, MP:0005502).

Mapping MPO Terms to the Corresponding HPO Terms

To map the MPO terms to the corresponding HPO terms, we applied two different approaches: (A) using public available HPO and MGI alignment files and (B) name-based matching of phenotype terms between ontologies.

For the first approach, we downloaded the file *mp2hp.txt*[7] from the PhenomeBLAST Computational Analysis of Mouse Phenotypes (CAMP) project (Oellrich et al., 2012). We also downloaded the file *alignment_type_2_HP_MP_b.rdf*[8] from the results of the Ontology Alignment Evaluation Initiative 2016 (OAEI2016).

For the second approach, we matched pairs of mammalian-human phenotype ontologies with the exact same name.

Merging HPO and MGI

Finally, the associations gene-phenotype lists derived from HPO and MGI were merged. The HPO ontology was augmented with its parent term information, and a filtration process was employed to generate a final list of genes associated with renal tubule dysfunction, hereafter defined as phenotype genes (PG). Figure 2 represents a schematic representation of extracting phenotype genes from HPO and MGI.

All the analysis on HPO and MGI data was carried out via R (vers 4.1.2 R Foundation for Statical Computing, Vienna, Austria), making use of the following packages:

- OntologyIndex: vers 2.10, (Greene et al., 2017)
- Igraph: vers 1.5.1, (Csárdi et al., 2023)
- Tidyr: vers 1.3.0,(Wickham et al., 2024)
- Dplyr: vers 1.1.3, (Wickham, François, et al., 2023)
- Stringr: vers 1.5.0, (Wickham, 2023)
- Rvest: vers 1.0.3, (Wickham, 2024)
- Xml2: vers 1.3.5, (Wickham, Hester, et al., 2023)
- BiomaRt: vers 2.54.0, (Durinck et al., 2009)

Identification of Genes Expressed in Proximal Tubule Cells

Since the proximal tubules have a crucial role in cystinosis, we searched public single-cell data containing the highest number of proximal tubule cells for healthy patients. After a literature review, we identified Stewart et al. (2019) as the best resource for our purposes. This study utilized high-throughput single-cell RNA sequencing (scRNA-seq) via

[7] https://code.google.com/archive/p/phenomeblast/wikis/CAMP.wiki downloaded on 26/10/2021
[8] https://github.com/bio-ontology-research-group/OAEI2016/tree/master/results downloaded on 26/10/2021

the 10x Genomics platform. The authors examined single-cell suspensions from fourteen mature human and six fetal kidneys. We downloaded the file *Mature_Full_v3.h5a*, which includes gene expression counts and normalized data, from the Kidney Cell Atlas[9]. Four CSV files were extracted: *X.csv*, which contains the normalized counts; *count.csv*, which contains the non-normalized counts; *var.csv*, which contains the gene metadata; and *obs.csv*, which contains the cell metadata. We parsed the dataset using the *Seurat* R package (vers 4.3, (Hao et al., 2021)), and we kept only data of cells annotated as proximal tubule cells derived from mature human kidney samples (33,694 cells in total). To minimize the number of lowly expressed genes in the result, we subset the matrix to include only those genes with more than two counts in more than three cells (population median).

This process identified 16,918 genes expressed in the healthy proximal tubule cells that were used to calculate ten distinct gene-gene similarity metrics.

Construction of Gene Similarity Matrices

Co-Expression and Proportionality

Similar genes have similar expression profiles, so the gene co-expression level can be used to describe the similarity between the gene and any other genes. Therefore, with scRNA-seq expression data, we calculated three different similarity measures between gene expressions: Pearson (PE), Rho (RH), and Phi_s (PH). The first is the Pearson correlation index, a gold standard of correlation measures, while the other two are alternative measures of proportionality (Skinnider et al., 2019). RNA-Seq gene expression data have two key properties: technical factors can impact the total number of counts, and the difference between values is only meaningful proportionally. Even if there are several normalization methodologies in the literature to cope with this issue, their application can lead to different results regarding the number and identity of differentially expressed genes. To this, we decided also to apply two proportionality measures, Rho (ρp) (Lovell et al., 2015) and Phi_s, (ϕ_s) (Erb & Notredame, 2016), to have less normalization-dependent metrics and to explore equally valid alternatives. The ρ_p measure spans the interval [-1, 1], exhibiting an analogy to correlation, whereas the ϕ_s measure extends over the range [0, ∞), paralleling the concept of dissimilarity.

Rho and Phi_s proportionality quantify the association between two feature vectors after log-ratio transformation. Consider the gene expression count matrix X with N samples (rows) and F features (columns). The metrics first involve the application of the centered log-ratio transformation (clr), which scales each sample vector (row) by its geometric mean (*g(x)*), calculated as:

[9] https://www.kidneycellatlas.org/ downloaded on 17/01/2022

$$\mathrm{clr}(x) = [\ln \frac{x_i}{g(x)}; \ldots; \ln \frac{x_F}{g(x)}]$$

Obtaining the new matrix A_{NxF}.

The proportionality index between each combination of A feature vectors (column), A_i and A_j ($i, j \in A$), is then computed as:

$$\phi_s(A_i, A_j) = \frac{var(A_i - A_j)}{var(A_i + A_j)}$$

$$\rho_p(A_i, A_j) = 1 - \frac{var(A_i - A_j)}{var(A_i) + var(A_j)}$$

All the metrics were calculated via R (vers 4.1.2) with the packages *dismay* (vers 0.0.1, (Skinnider et al., 2019)). This package wraps ten different distance metrics. To calculate the Pearson correlation, it imports the *cor* function from the *WGCNA* package (vers 1.51, (Langfelder & Horvath, 2008)). For ϕ_s and ρ_p instead, it imports phis and perb functions from the *propr* package (vers 3.1.8, (Quinn et al., 2017)). The Pearson correlation was calculated on normalized data, while ϕ_s and ρ_p on non-normalized data.

GO Semantic Similarity

The R package *GoSemSim* (vers 2.20.0, (G. Yu, 2020; G. Yu et al., 2010)) was used to compute the semantic similarities between genes. Three matrices were calculated, one for each GO ontology: biological process (BP), molecular function (MF), and cellular component (CC). We prepare the GO structure and the gene-to-GO terms mapping in the required format using the *godata* function. We use the *mgeneSim* function to calculate the Wang metric and to apply the BMA method.

Pathways Co-Membership

We collected the gene-pathway annotation from the Reactome database via the *Reactome.db R* library (vers 1.77.0, (Ligtenberg, 2019)). We collected the gene-pathway annotation from the KEGG database via *org.Hs.eg.db* R library (vers 3.14.0, (Carlson, 2019)).

To calculate the pathways co-membership, we used the *mgeneSim* function from the BioCor package (vers 1.18.0, (Revilla Sancho, 2024)), which implements the Sørensen-Dice index as a measure of pathway similarity. Dice similarity is between 0 and 1 and is twice the number of genes shared by the pathways divided by the sum of genes in each pathway. Given two pathways (a, b), the Dice index is defined as follows:

$$Dice(a, b) = \frac{2\,|a \cap b|}{|a| + |b|}$$

Since genes can be present in multiple pathways, the *mgeneSim* function calculates the dice similarities between each pathway and combines these to obtain the overall

similarities between genes. The package provides four different aggregation methods. We apply the MAX and BMA methods.

The BMA method is the same as described for GO similarities.

The MAX method calculates the maximum similarity over all combinations of pathways pair between the two genes. Given two genes, g_1 and g_2, annotated respectively in m and n pathways, and defined two pathways in which the two genes are present as p_{1i} and p_{2j}, the MAX aggregation method is defined as follows:

$$sim_{MAX}(g_1, g_2) = \max_{1 \le i \le m, 1 \le j \le n} sim(p_{1i}, p_{2j})$$

Prediction of New Renal Tubule Phenotype Genes

To predict new renal tubule genes, we designed and implemented a feed-forward neural network architecture, shown in Figure 3. The aim is to identify the more similar genes, according to varied aspects, to the already annotated PGs. We devised a two-step artificial neural network (ANN) optimized to solve a positive-unlabeled classification problem. In the model, the positive labels were the genes known to be associated with renal tubule function, whereas the input features were the similarity matrices described above.

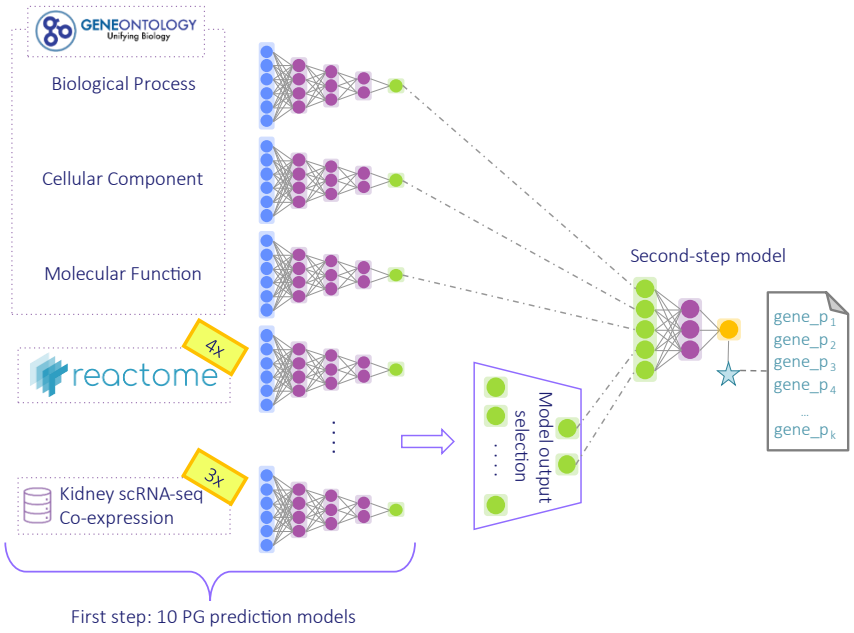


Figure 3 - ANN architecture. Ten different ANNs were trained to estimate pseudo-probabilities, indicating the probability that a gene is PG. Model output selection was used to determine the most efficient combination of sources. The second network took the outputs of the five selected ANNs as input and returned the overall pseudo-probability for each gene to be PG.

In the first step, ten distinct ANNs with the same architecture were set up and trained separately to estimate pseudo-probabilities indicating the likelihood that a gene is associated with renal tubule function based on the corresponding similarity matrix. After evaluating all the possible combinations of variables to determine the most efficient setup, the output values from the five chosen sources are employed to train an additional ANN. The second network takes the outputs of the five individual ANNs as input, merges them to assign appropriate importance to each source, and returns the overall pseudo-probability for each gene to be phenotype-related. The ANN model was developed using the *PyTorch* library (vers. 1.10.2, (Paszke et al., 2019)) in a Python environment (vers 3.11.2).

First Part – Prediction From Single Similarities

The first part consists of 10 feed-forward neural networks, each trained on a distinct similarity matrix. These include Pearson, Rho, Phi_s, GO semantic similarity BP, GO semantic similarity CC, GO semantic similarity MF, Pathways co-membership Reactome MAX, Pathways co-membership Reactome BMA, Pathways co-membership Kegg MAX, Pathways co-membership Kegg BMA.

Cross Validation

The training process of each ANN employes a 10-fold stratified cross-validation scheme to avoid overfitting, find the best hyperparameters, and ensure reproducibility (The MAQC Consortium, 2010). Each dataset underwent random partitioning into ten stratified folds. During each training session, one fold was designated as the test set, another as the validation set, and the remaining eight folds served as the training set. This procedure was iterated ten times, ensuring that each fold was utilized exactly once as a test set and once as a validation set. Figure 4 shows the cross validations scheme. The data splitting was done by personalizing the *StratifiedKFold* function from scikit-learn. To improve the model, we optimized hyperparameters such as the learning rate, number of hidden layers, number of hidden nodes, and weights based on validation set scores. We generated the prediction on the unseen test set to evaluate the model's performance.

Dimensionality Reduction

In the training session, a dimensionality reduction algorithm was applied to the similarity matrix as a preliminary step, specifically, the Principal Component Analysis (PCA). The PCA fitting was performed exclusively on the training set to prevent any information leakage between the training and test sets. The transformation was then applied to the validation and test sets. We used the PCA function from the *scikit-learn* package (vers 1.0.2, (Pedregosa et al., 2011)). An evaluation was conducted to determine the optimal number of PCs to be held. This entailed selecting the configuration that explained 0.496 of Pearson's correlation matrix variance, which was found to be the first 60 principal components. This threshold was employed for all networks; thus, each ANN takes the initial 60 PCs of similarity matrices as input.

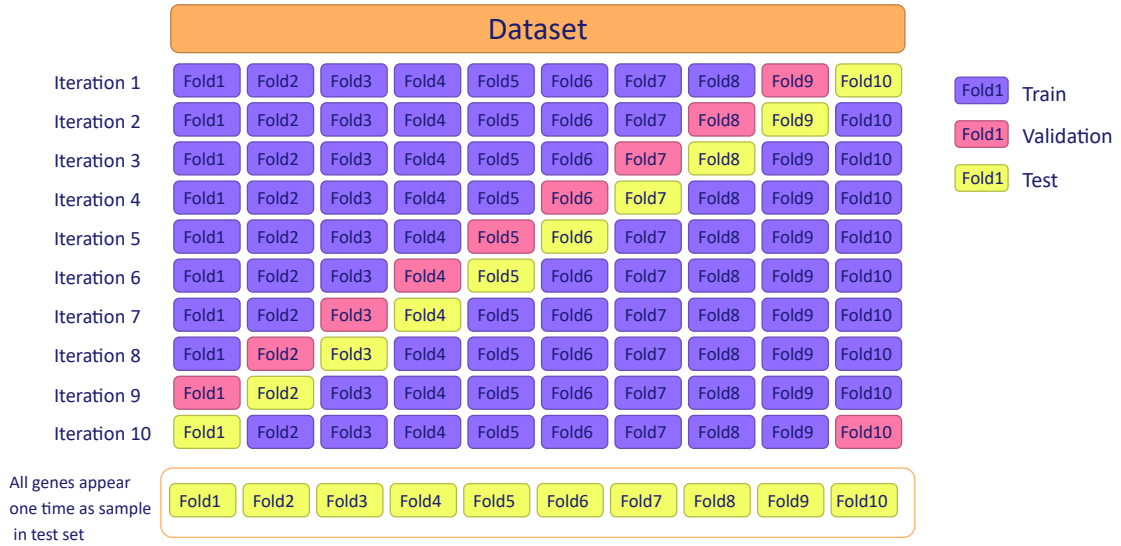


Figure 4 – The stratified 10-fold cross-validation scheme.

Layers

This network comprises four fully connected linear layers, where the first three are followed by a Rectified Linear Unit (ReLU) activation function.

The ReLU activation function is one of the most common activation functions used in neural networks, introduced by Fukushima (1969). It expresses the positive part of its argument and its formally defined as:

$$f(x) = x^{+} = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & if\ x > 0 \\ 0 & otherwise \end{cases}$$

where x represents the neuron input.

The layers are structured as follows:

- Linear layer with 60 input and 30 output neurons, followed by ReLU activation.
- Linear layer with 30 input and 20 output neurons, followed by ReLU activation.
- Linear layer with 20 input and 15 output neurons, followed by ReLU activation.
- Linear layer with 15 input neurons and one output neuron.

A sigmoid function was applied to obtain a probability-like value as output, which allows all outputs to be scaled between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{1 + e^{x}} = 1 - \sigma(-x)$$

This function was applied to the final linear layer but only during training. During the validation and testing phases, the *BCEWithLogitsLoss* function was utilized, integrating a sigmoid layer and the *BCELoss* directly, obviating the need to apply the sigmoid function.

The weights of the network are initialized using the Xavier initialization method. This method ensures that the weight variance in each layer is consistent throughout the

network, preventing the gradients from disappearing or exploding during the backpropagation training process. By initializing the weights according to the Xavier initialization method, the network is better prepared to learn effectively, resulting in faster convergence and improved overall performance.

Batch Size– Unbalanced Data – Weighted Sampler

A DataLoader is a utility provided by the PyTorch library that facilitates efficient data loading for training, validation, and testing processes in DL models. Their application is essential for helping the orderly loading of data into the neural network and is required due to memory limitations that preclude the simultaneous loading of the entire dataset. The *batch_size* parameter controls the amount of data loaded into memory and fed into the neural network, and the sampler parameter is utilized to manage how to sample the batch from the dataset. The batch size was set as 64 units. Given the highly imbalanced nature of the dataset, a weighted *RandomSampler* from PyTorch was applied within the training DataLoader. This configuration ensures that batches in the model training process contain an equal representation of labeled and unlabeled examples. The sampler weights are computed as the reciprocals of the class frequencies observed within the training set. For instance, if class frequencies are denoted as freq_1=34 and freq_0=52, the corresponding weights are calculated as weight_1=(34+52)/34 and weight_0=(34+52)/52, respectively. In contrast, the DataLoader function for the validation and test sets does not incorporate weighted sampling. This decision was made to preserve the integrity of the validation and testing processes, allowing them to faithfully represent the natural distribution of data encountered in real-world scenarios. The loss function assigns a higher weight to positive examples to address class imbalance within the validation and test sets. This strategy aims to mitigate the impact of class imbalance and promote the robustness of model performance across diverse datasets.

Loss Function and Adam Optimizer

In the case of binary classification, the cross-entropy loss function is preferred because it allows the models to assign probabilities to the labels independently of each other. Two different loss functions from the PyTorch package were used for the training and the validation sets. The *BCELoss* function was used for the training sets, while the *BCEWithLogitsLoss* function was used for the validation set. Although both functions are specific to binary classification cases, they were configured so that the *BCELoss* function assigns the same error weight when the model misclassifies positive and negative examples, while the *BCEWithLogitsLoss* function does not. This choice is due to the weights added to the DataLoader sampler during the training phase, which balance the batches. In the validation DataLoader, on the other hand, each sample has the same weight, resulting in an unbalanced batch with many unlabeled examples and only a few positive ones. *BCEWithLogitsLoss* produces worse scores when the model misclassifies a positive sample than when it misclassifies an unlabeled sample. The selected approach was designed to facilitate the validation and testing of the model with samples that closely approximate the original data set. In the proposed architecture, it is of critical

importance that the algorithm is able to correctly identify positive samples, as this enables it to identify observations that are highly similar to them subsequently. Consequently, a higher penalty was assigned to positive example misclassification, which is considered more severe than a misclassification error on a negative example. During the training phase, model weights were updated using the Adam optimizer, a common algorithm for stochastic gradient descent (SGD)-based optimization commonly used in DL models. Introduced in the paper "Adam: A Method for Stochastic Optimization," the Adam optimizer combines the best features of two other optimization algorithms, AdaGrad and RMSProp, to converge to the minimum of the cost function more efficiently (Kingma & Ba, 2014). The first and second moments of the gradients are employed to regulate the learning rate throughout the training process. The mean and variance of the gradients represent, respectively, the first and second moments. The Adam algorithm updates the parameters at each iteration by maintaining an exponentially decaying average of past and squared gradients. This makes Adam faster and more efficient at converging to the minimum of the cost function than conventional descent techniques. Adam's most important hyperparameters are the learning and exponential decay rates. To achieve optimal results, these hyperparameters can be adjusted for specific problems. The learning rate (lr) adjusts the step size at each iteration during the gradient descent. A higher learning rate can speed up convergence but can also cause the optimizer to outperform the optimal solution. The decay parameter (decay) controls how the learning rate decreases with time. In this case, it was set equal to 1e-5.

Epoch – Early Stopper

The maximum number of epochs was set to 1000. An early stopping function was implemented to prevent overfitting from interrupting the training process if the recall metrics did not improve for more than seven epochs. Typically, early stopping functions are based on loss function values. However, in this case, the objective was to optimize for recall, as the goal was to train a network that would be accurate in recognizing positive examples.

Model Evaluation

In order to identify the optimal combination of hyperparameters, including the number of layers, layer size, activation functions, and learning rate, a series of experiments were conducted. The models were evaluated to achieve a high recall and a low number of false positives to optimize the performance on the validation set. Once the most appropriate parameters were determined, all ten networks were trained, the optimal source combination was explored, and the results were saved.

A preliminary approach to combine the probabilities to obtain a single measure was to calculate, for each gene, a simple arithmetic average of the five probabilities obtained in the first training phase. Although this approach yielded reliable results, it was tested to use the last layer as input to a second neural network. This approach ensures that no single source carries the same weight when calculating the overall probability. Rather

than imposing a fixed weight on each source, the algorithm learns which sources have greater influence in calculating the final probability, thereby improving the reliability of the outcome.

In order to determine which of the selected sources should be chosen from the ten calculated sources, we evaluated all possible combinations that included all three GO metrics, one co-expression similarity metric (among Rho, Phi_s, and Pearson), and one of the four pathway comembership metrics (either KEGG or Reactome). The decision to maintain at least one for each of the three categories of similarity metrics was driven by the fact that their combination provided valuable insights into the functional relationships and biological roles of genes. Similar genes have similar expression profiles, so the scRNA-seq co-expression metrics provided insights into the co-regulation of genes. Furthermore, it helped to identify genes that are dynamically co-regulated in specific cell types or states, which may not be evident in pathway databases. Genes that are part of the same pathways are likely to work together in specific biological functions. Incorporating pathway co-membership revealed genes involved in common signaling cascades, metabolic processes, or other cellular mechanisms, even if they are not necessarily co-expressed. Lastly, GO semantic similarity captured the functional relatedness of genes since it measures the likeness in meaning between two genes based on their annotations to GO terms. Integrating GO semantic similarity helped identify functionally similar genes, even if they are not directly co-expressed or part of the same pathways. Combining these different similarity measures gives us a more comprehensive understanding of the functional relationships between genes.

In total, there are 12 possible metrics combinations. For each of these combinations, we calculated the probability output of each source and then combined these values using an arithmetic mean to determine the final probability distribution. The means were used to estimate which genes would be labeled as PG and which as non-PG using as thresholds all values between 0.5 and 1 included, with a span of every 0.05 (total of 11 thresholds).

Based on these results, we constructed the confusion matrices to calculate the recall, precision, and PU score according to the following equations:

$$Precision = \frac{True\ positive}{True\ positive\ +\ False\ positive}$$

$$Recall = \frac{True\ positive}{True\ positive\ +\ False\ negative}$$

$$PU_{\text{Score}} = \frac{p \times r}{Pr(y = 1)} = \frac{r^2}{Pr(\hat{y} = 1)}$$

Where:

p represents the precision,

r represents the recall,

$Pr(y = 1)$ represent the proportion of known positive labels in the prediction set and,

$Pr(\hat{y} = 1)$ represent the proportion of positive predictions made by the model.

Second Part – Merging Results to Obtain the Overall Prediction.

The second network takes the last layer of the previously trained networks as input and returns the probability of being a PG as output. For each gene, we took the values obtained on the test sets, from the last linear layer (there is no sigmoid in the last layer). The second network was trained using only the outputs of the five selected sources.

The architecture and workflow are very similar to the first networks. It is a feed-forward neural network with three linear layers alternated by ReLU functions. The input consists of the five values for each sample (i.e., the output of each ANN applied to each of the five selected sources), and the output is a single value that passes through a sigmoid function to obtain a pseudo probability. A 10-fold cross-validation was applied, with balanced batches in the training and unbalanced batches in the validation and testing. The same loss functions are used as in the first network: *BCELoss* for the training set and *BCEWithLogitsLoss* for the validation and test set. The early stopping function is again applied, terminating the training after seven epochs of no improvement on the recall metric. Adam was chosen as the optimization algorithm, and the weights were initialized using the *xavier_normal* function. The model was trained for a maximum of 1000 epochs with an LR of 1e-5. To obtain more reliable estimates, we trained the network 100 times using different random seeds, impacting the initialization of the network and, thus, the learning process. We calculated 100 recall indices and determined the mean and confidence interval (CI). For each gene, we obtained 100 different estimates of the probability of it being a PG. We labeled those with a mean of 100 probabilities greater than 0.5 as new PG.

To compare the performance of the ANN-based model with alternative, more shallow models, a Random Forest (RF) model was trained, and the Receiver Operating Characteristic (ROC) curves were plotted. It was found that the ANN model achieved a better score than RF (Figure 5), both in training and in the validation set. The ANN model achieved an area under the ROC of 0.71 in the training and validation set, while the RF model only reached 0.66 in the training and 0.60 in the validation set.

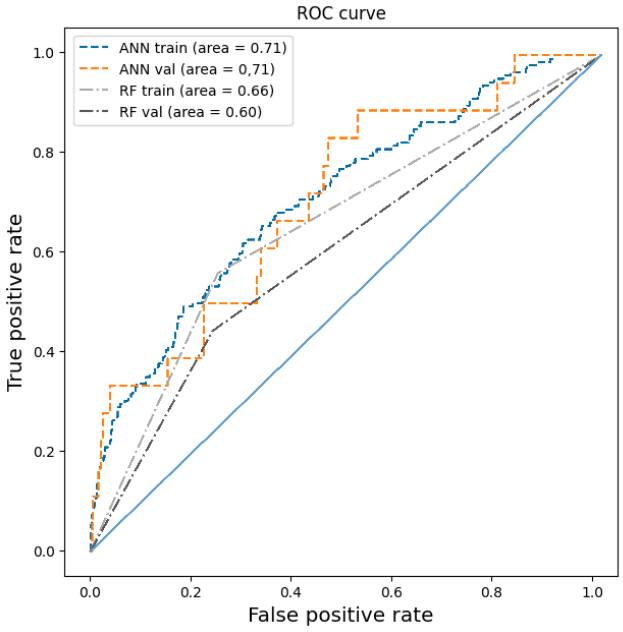


Figure 5 - ROC curves of ANN model and RF model. The orange and blue lines represent the ANN RC curves on the train and validation sets. The light gray and dark gray lines represent the RF ROC curves on the train and validation sets, respectively.

Sensitivity Analysis of the Selected Source

To determine the impact of each source on the probability of being PG, we systematically retrained the second part of the architecture, removing one of the five sources in each iteration. We ran 100 training iterations for each model. We computed the exact probabilities and their corresponding CI for all genes. We calculated the recall value for each 100 repetitions and determined the mean and CI. We applied a one-tailed t-test to determine the statistical significance of the difference in recall performance between the complete model and the model without one source.

Characterization of Phenotype Genes and External Validation

To verify the biological significance of our findings, we decided to explore the predicted PGs' biological functions by utilizing two pathway databases that have not been previously employed for gene-gene similarity analysis, namely WikiPathways and KEGG. Additionally, to gain deeper insights into the role of the identified genes in kidney disease, we compared the PGs with genes connected to chronic kidney disease (CKD) via genome-wide association studies. Furthermore, we analyzed the agreement between the characterization of the PGs and mitochondrial genes by utilizing the resources available on Mitocarta.

KEGG and WikiPathways gene pathway annotations were obtained from the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) using the MSigDBr R package (vers.

7.5.1, (Dolgalev, 2022)). Enrichment analysis was performed with the *enricher* function of the clusterProfiler package (vers 4.9.1, (T. Wu et al., 2021; G. Yu et al., 2012)).

The GWAS variants associated with CKD were downloaded from the GWAS catalog[10]. We performed a one-tailed Fisher's exact test using the *fisher.test* function in R to assess whether the enrichment in the number of genes present in CKD-GWAS was statistically significant.

The MitoCarta3.0 gene set was downloaded from the Broad Institute portal[11]. To assess whether the enrichment in the number of genes present in MitoCarta was statistically significant, we performed a one-tailed Fisher's exact test using the *fisher.test* function in R.

Identification of Disease-Related Molecules

The omics data were obtained from (Berquez et al., 2023). The RNA sequencing data from the mouse model is available on the Gene Expression Omnibus (GEO) of the NCBI, ID GSE226397[12]. The proteomic data is available at the ProteomeXchange Consortium as PXD040450[13] , and the metabolomics data at ZENODO[14]. The Bioconductor limma package (Ritchie et al., 2015) was used for each omic to identify differentially expressed genes or metabolites between CTNS knockout ($CTNS^{KO}$) mouse versus control mouse. The Benjamini-Hochberg (BH) method was used to correct p-values for multiple testing. For metabolomic data, human gene lists were extracted from the mouse metabolite lists using the MetaboAnalyst[15] portal. In the process, some mouse compounds were translated to the same human gene. In some cases, multiple Log Fold Changes (LogFC) have opposite signs. Therefore, we decided to eliminate all such cases, even though they were significant because they reflected diametrically opposite trends. Genes and metabolites with an adjusted p-value < 0.05 and a human homolog were identified as disease molecules (DMs). The top 300 proteins (sorted by nominal p-value) with -0.5 < LogFC < 1.5 and a human homolog were identified as DMs.

Multilayer Network

A *multilayer network* is a type of network that consists of multiple layers, where each layer represents a different type of relationship or interaction between nodes. The nodes

[10] https://www.ebi.ac.uk/gwas/efotraits/EFO_0003884 downloaded on 07/06/2023
[11] https://personal.broadinstitute.org/scalvo/MitoCarta3.0/Human.MitoCarta3.0.xls downloaded on 07/06/2023
[12] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE226397
[13] https://www.ebi.ac.uk/pride/archive/projects/PXD040450
[14] https://zenodo.org/records/7655201
[15] https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml

in a multilayer network are interconnected across the different layers, providing a more comprehensive and realistic representation of complex systems (De Domenico, 2023).

The multilayer network was constructed by combining three different types of interactions from five different sources. We included directed protein-protein interactions, transcription factor-gene interactions (GRN), and protein-metabolite interactions. Figure 6 provides a schematic illustration of the network construction process with the sources' references.

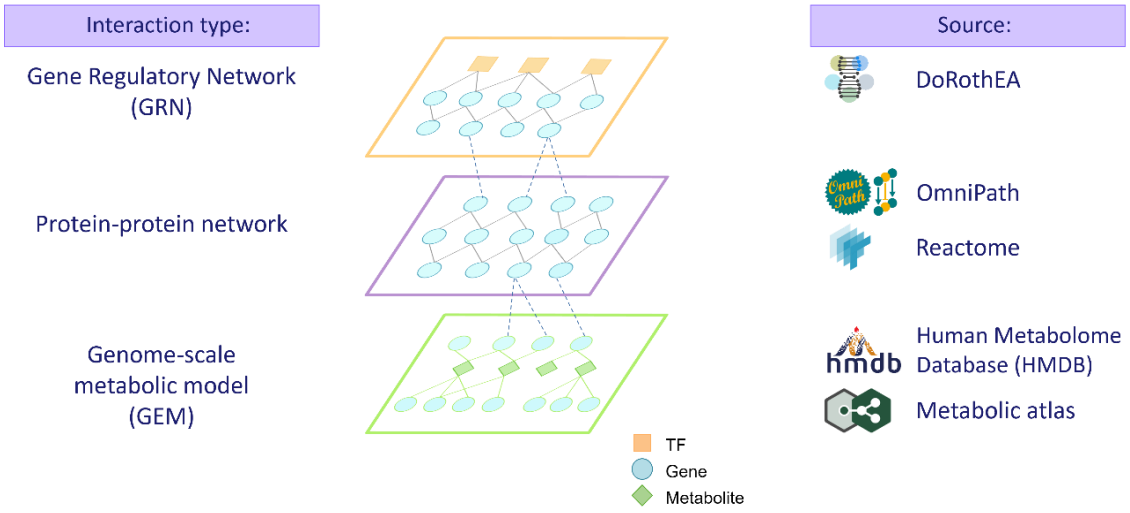


Figure 6 - Multilayer network scheme. The orange layer represents the TF-gene interactions. The purple layer represents the protein-protein interactions, and the green layer represents the protein-metabolite interactions.

Protein-Protein Interactions

We utilized the Omnipath database and Reactome pathways to collect the protein-protein interactions. The first was retrieved using the OmnipathR R package (version 3.5.24, (Valdeolivas et al., 2016)) on December 14, 2022. The *import_omnipath_interactions* function was used to download the interactions, resulting in 40,014 interactions. After converting gene symbols to Entrez IDs and removing interactions that included proteins without Entrez IDs, we obtained a directed network with 7,723 nodes and 37,677 edges. The Reactome pathways were downloaded from the Pathway Commons portal[16], containing 365,961 interactions. After applying a filter, we retained only those interactions related to controls-state-change-of, controls-transport-of, controls-phosphorylation-of, controls-expression-of, and catalysis-precedes. The resulting data frame was transformed into a directed network, collecting 8,045 nodes and 182,278 edges.

[16] https://www.pathwaycommons.org/archives/PC2/v12/ downloaded on 05/09/2023

Gene Regulatory Networks

In addition to protein-protein interactions, we also considered GRN. To acquire the DoRothEA data, we used the decoupleR package (version 2.3.1, (Badia-i-Mompel et al., 2022)) and retrieved all gene and TF interactions using the *get_dorothea* function. The interactions were filtered only to include those with the highest confidence levels (A, B, and C). After filtering, the data frame was converted to a directed network of 9,298 nodes and 32,273 edges.

Protein-Metabolite Interactions

In addition to the two previously mentioned forms of interaction, we have also included genome-scale metabolic networks as a third category. Protein-metabolite interactions were derived from Metabolic Atlas (J. L. Robinson et al., 2020) and the Human Metabolome Database (HMDB) (Wishart et al., 2022).

We downloaded the XML file with the description of the metabolic reactions from the Metabolic Atlas repository[17] (vers 1.12.0). We parsed the file with the BeautifulSoup python library (vers 4.11, (Richardson, 2007)) and extracted the list of reactants, products, and associated proteins for each reaction. We then created a list of reactant-to-protein and protein-to-product interactions. For reversible reactions, we also included protein-to-reactant and product-to-protein interactions. To simplify the network, the information about the subcellular compartment where the reaction occurs was removed, and common metabolites such as hydrogen, water, and oxygen were excluded from the interactions. The final network included 6,566 nodes and 46,341 interactions.

We downloaded the XML file with the information for all metabolites from the HMDB portal[18]. We selected the protein interactions with metabolites detected in the metabolome analysis on the cystinosis animal model (all metabolites independently of their differential expression between WT and KO animals). These selected interactions were merged with those from the other sources to obtain a complete network.

Merging

Finally, all the previous networks were merged into a directed network with 1,103,515 edges and 40,262 nodes. After obtaining the complete network, we applied a filter to retain only the proteins expressed in the PT, as described in the *"Identification of Genes Expressed in Proximal Tubule Cells"* section. We removed some edges from the HMBD database, keeping only edges involving metabolites detected in the animal model experiments, regardless of their significance. All other HMDB interactions were excluded.

[17] https://github.com/SysBioChalmers/Human-GEM/tree/main/model downloaded on 29/06/2022
[18] https://hmdb.ca/downloads downloaded on 28/09/2022

After carefully analyzing the complete network before filtering, it was determined that it was too large to be useful for our purposes. Some metabolites, which are involved in numerous reactions, had thousands of edges, which made the network difficult to analyze. This result led us to decide to remove these edges from the network. The resulting multilayer network consists of 14,936 nodes and 193,992 edges.

Identification of Network Modules and Enrichment Analysis

To divide the network into smaller subnetworks with a maximum of 100 nodes, we used the *cluster_walktrap* function from the igraph R package (vers 1.5.1, (Csárdi et al., 2023)) that implements the Walktrap Community Finding algorithm implementation (Pons & Latapy, 2006). To avoid having large modules, following the approach of Choobdar et al. (2019), if we identified modules larger than 100 nodes, we applied the algorithm recursively up to six times. We identified 271 modules, 12 of which contained more than 100 nodes.

The enrichment of DMs and PGs in all modules was assessed using the piano package (version 2.13, (Väremo et al., 2013)). The gene sets were loaded using the *loadGSC* function, and the enrichment was calculated using the *runGSAhyper* function, which implements Fisher's exact test. The FDR method was used to adjust the p-values. All genes/metabolites in the network were initially included in the universe. We then changed our strategy to evaluate the enrichment of the molecules by using only those genes/metabolites detected in single omics. Modules were considered enriched if they had an FDR-adjusted p-value < 0.05 in both PGs and at least one omic.

Pathway and Gene-Set Analysis of the Modules

The enrichment of the GO Cellular Components terms was evaluated for each of the 22 modules to study the subcellular compositions of the protein constituents of the modules. All genes in a given module were extracted, and the enrichment was assessed using all genes in the complete network as a reference universe. Enrichment was calculated using the *enrichGO* function of the clusterProfiler package (vers 4.9.1 (T. Wu et al., 2021; G. Yu et al., 2012)). P-values were adjusted using the BH method, and a threshold of 0.05 was applied to select the significant terms. Once the significantly enriched CCs were identified, we selected only the most significant term for each module (with the smallest adjusted p-value). A simplified representation of the results was created by connecting the GO terms to their parent GO terms describing cell macrostructures reported by Paganin et al. (2023) These terms include Membranes, Cytoplasm, Nucleoplasm, Extracellular region, Endoplasmic reticulum, Golgi apparatus, Ribosome, Actin cytoskeleton, Mitochondrion, Vesicle, Nucleolus, Nuclear body, Lysosome, Endosome, Chromosome, Microtubule cytoskeleton, and Lipid droplet.

The go.obo file was obtained from the Gene Ontology portal[19] and was transformed into a DAG with the igraph package. Only those relations pertinent to the CC ontology were retained. The *distances* function from igraph was employed to calculate the distance between each enriched term and any generic terms it was connected to. The rationale behind this approach was to identify the generic term with the fewest degrees of separation from the specific term. If a specific term was a child of multiple generic terms, the generic term closest to the specific term was retained. Additionally, we evaluated the enrichment of the KEGG pathway and Hallmark gene sets. For the enrichment analysis, we employed the *enricher* function of the clusterProfiler package (vers 4.9.1 (T. Wu et al., 2021; G. Yu et al., 2012)) while the MSigDB (Liberzon et al., 2011) gene sets were obtained using the *msigdbr* function of the MSigDBr R package (vers. 7.5.1, (Dolgalev, 2022)). In both cases, we used all genes in the complete network as the universe. The p-values were adjusted via the BH method. A module is considered enriched if the adjusted p-value is less than 0.05.

Drug Repurposing Analysis

To get drug information, we utilized DrugBank, which is a comprehensive portal that includes information on drugs, both those that have already been approved and those that are currently in development. On June 29, 2023, we downloaded the complete database from the DrugBank portal (version 5.1.10, (Wishart et al., 2018)). A Python script was employed to extract the information from the XML file and transform it into a text table. The extracted information comprises the following: drug ID, drug name, status, target ID, organism, action, and gene name. The drugs were then filtered and selected based on the following criteria. Firstly, only those drugs with human targets and without DNA or RNA as targets were retained. Furthermore, only drugs with a known activating or inhibitory mechanism of action (or synonymous terms) were included. In the case of a drug associated with more than one mechanism of action, we only retained the record if all of the mechanisms were in mutual agreement. For example, if activation and inhibition were listed in the mechanism of action field, we decided to delete the record since it was impossible to determine which mechanism to label the drug. A data frame is created that contains all drugs whose targets are present in the various modules, along with the module in which they are placed. For each module, a data frame that contains only DMs is extracted. Disease genes can be dysregulated in both transcriptomics and proteomics. If a gene is dysregulated in only one of these, the data from the dysregulated source is retained. If a gene results dysregulated in both proteomics and RNA-Seq, we have opted to maintain the FC identified in proteomics, as it is more downstream. The first neighbor list is extracted for each drug in each module. The first neighbor of a drug is labeled as either up or downregulated. The number of these first neighbors of the drug that are

[19] https://geneontology.org/docs/download-ontology/ downloaded on 29/06/2023

upregulated and the number that are downregulated are calculated. The number of up and downregulated genes in the module is also calculated. Subsequently, for each quartet of drugs, targets, modules, and mechanisms of action (MoAs), the following four pieces of information will be provided:

1. The number of DMs that are up-regulated and directly connected to the drug;
2. The number of DMs that are down-regulated and directly connected to the drug;
3. The number of DMs that are up-regulated in the entire module;
4. The number of DMs that are down-regulated in the entire module.

Each drug's mechanism of action is identified and classified as either an inhibitor or activator. The drug-module combination is defined as consistent if the number of DMs in the module with an opposite sense to the drug is greater than those with a concordant sense. This labeling is achieved by verifying that the number of up-regulated DMs is greater than the number of down-regulated DMs if the drug has an inhibitory-like MoA. Similarly, if the drug has an activator effect, the number of down-regulated DMs in the module is compared to the number of up-regulated DMs. This consistency check is also performed with the number of DMs of the drug's first neighbors.

Web Application

Given the considerable heterogeneity, complexity, and quantity of the outcomes produced, developing a software application (app) to visualize the results interactively was necessary. The app, developed with the shinyapp R package (vers 1.7.4, (Chang et al., 2024)), included an initial tab in which it is possible to view the level of enrichment in the various omics modules. Clicking on a module of interest from this tab would direct the user to the second tab, which provided an overview of the results of the individual module. This tab also included an interactive plot of the module's network. In addition, descriptive statistics of the module were available, including the number of nodes, differentiated by type, number of PGs, and number of DMs. Moreover, the same tab enabled the observation of the results of Hallmark and KEGG enrichments present in the module. Finally, a table was provided with all the nodules present in the module, along with all the features of interest. These included LogFC of omics on $CTNS^{KO}$ mouse model and data related to DrugBank. Finally, a third tab enabled the user to visualize and explore the connection between different modules through an interactive plot. By selecting a module of interest, the user could also visualize its enrichment in both Hallmark and KEGG.

Figure 7 – Web app for interactive exploration of pipeline results. Module tab.

Results

Identification of Genes Associated with Kidney Proximal Tubule Function Through Deep Learning

Identification of Literature-Derived Renal Tubule Phenotype Genes

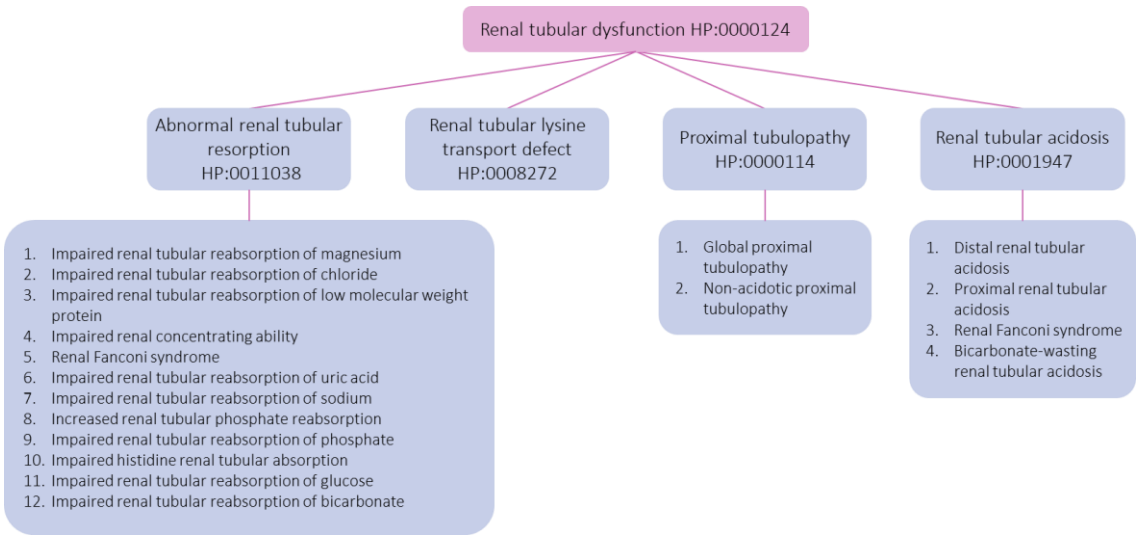


Figure 8 - Child phenotype of Renal tubular dysfunction

The proximal tubule has a pivotal role in cystinosis. Consequently, this study focused on analyzing the *Renal Tubular Dysfunction* phenotype, identified by HPO with the identifier HP:0000124, and its associated child phenotypes. In total, this phenotype is related to 19 child phenotypes, as illustrated in Figure 8. Information from the HPO was combined with MGI to define a list of gene-phenotype associations, as described in the Methods section. To obtain a list of phenotype-gene associations, we annotated genes causative of rare diseases or mutated in a disease mouse model as related to the phenotypes of that disease/model. We then selected the gene-phenotype associations related to renal tubular dysfunction. In total, by combining HPO and MGI data, we identified a set of 189 genes associated with renal tubular dysfunction. Table A7 presents a comprehensive list of PGs and detailed information regarding the source from which the link was derived. Table 1 represents the counts of genes derived at each resource.

Table 1 - Number of genes found in each resource.

Source	N of distinct genes
HPO	103
HPO, MGI	8
MGI	78

Identification of Genes Expressed in Proximal Tubule Cells

The data published by Stewart et al. (2019) consisted of 33,694 cells and 40,268 genes. After applying the filters described in the Methods section, the final array was identified as comprising 16,918 genes expressed in 33,694 cells. Both the tSNE (Belkina et al., 2019) and UMAP (McInnes et al., 2018) algorithms were employed to represent these data. Figure 9 represents the first and second UMAP components, while Figure 10 represents the first and second tSNE components.

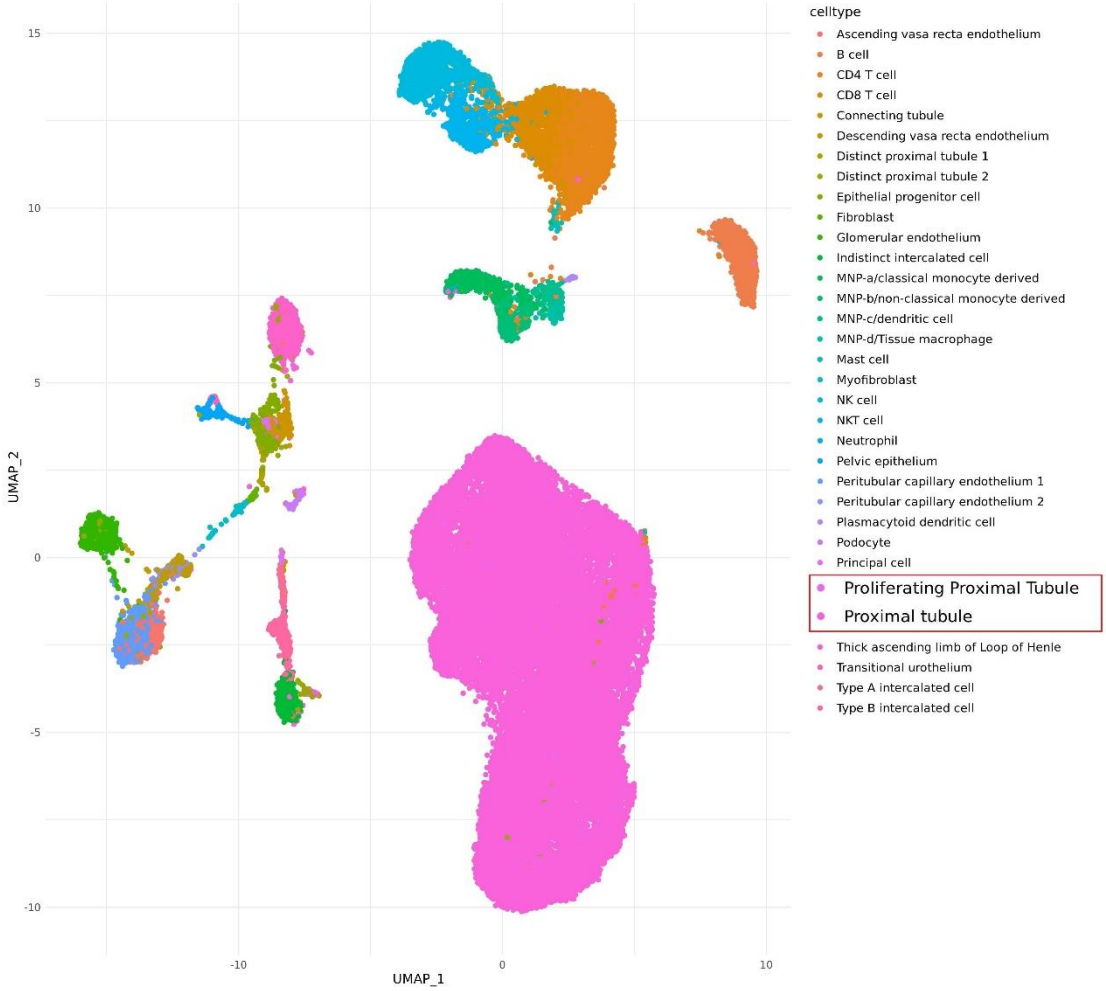


Figure 9 - UMAP representation of single-cell data. The pink dot represents proximal tubule (PT) cells.

In both plots, the samples are colored according to the cells' classification. Figure 9 represents all cell types, whereas Figure 10 distinguishes only PT cells (in pink) from others. In both figures, the proportion of PT cells is predominant, and the cluster is well separated from the other clusters.

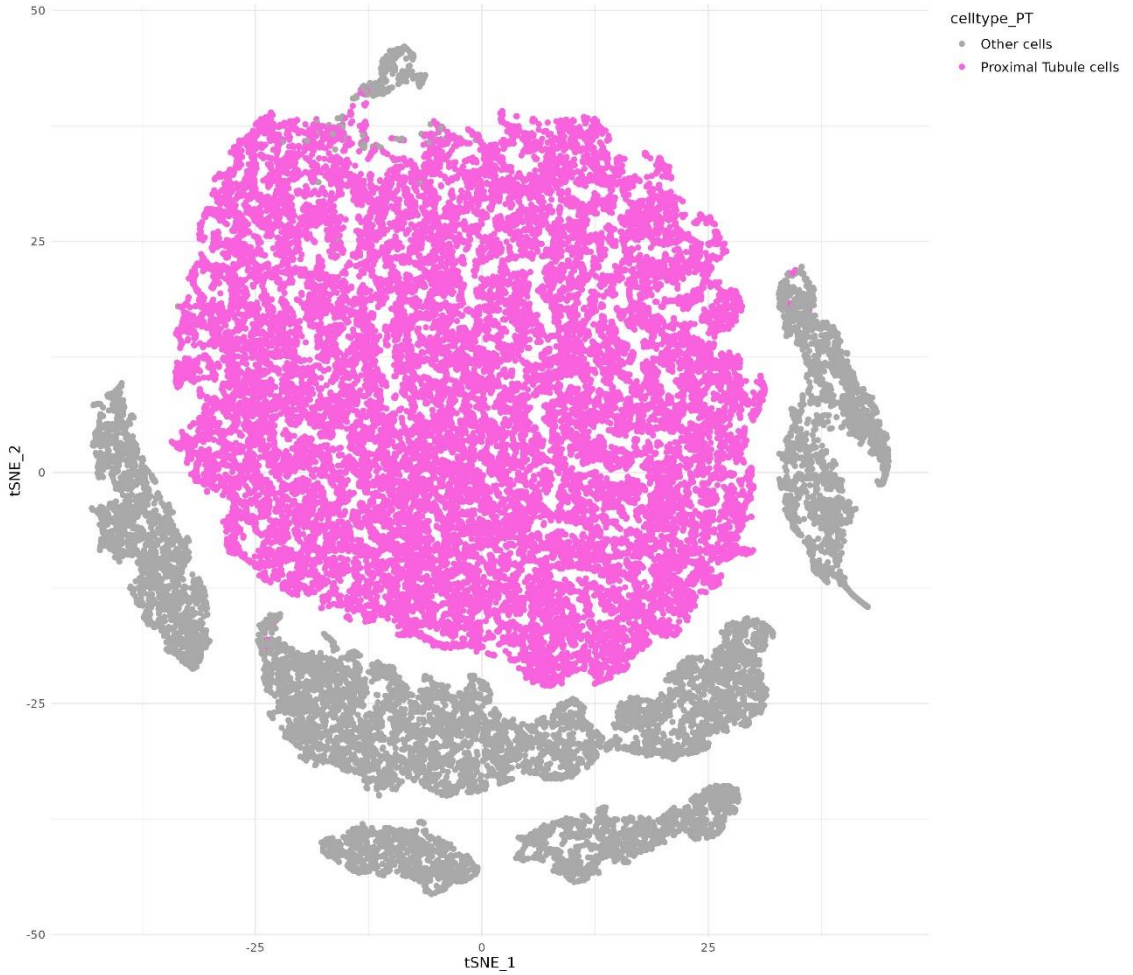


Figure 10 - tSNE representation of single-cell data. The pink dot represents proximal tubule (PT) cells.

Gene Similarities Matrices

The expression observed in these cells was used to calculate the Pearson index, the Rho and Phi_s proportionality coefficients. A list of 24,368 unique genes was created to calculate semantic similarity indices. These genes were derived from combining those expressed within the proximal tubule and PGs found in the literature. It should be noted that not all genes have annotations in all ontologies. Consequently, the similarity matrices are smaller than the complete list of genes. Specifically:

- The matrix with the similarity derived from the BP ontology has 15,846 rows.
- The matrix with the similarity derived from the CC ontology has 16,950 rows.
- The matrix with the similarity derived from the MF ontology has 16,926 rows.
- The four matrices derived from the KEGG and Reactome pathways have 5,868 and 11,148 rows, respectively.

Prediction of New Renal Tubule Phenotype Genes

First Part – Prediction From Single Similarities

The set of ontology-derived genes was used as the positive set in a positive-unlabeled ML framework based on neural networks. This framework integrates information from multiple biological modalities to predict novel phenotype-related genes based on their similarity with the known ones. The initial phase of the ANN architecture entails training all networks individually. Each network accepts one of ten similarity matrices as input and then returns a probability of each gene being a PG. Table 2 illustrates the precision and recall metrics, as well as the PU scores, calculated on the test sets for each source taken individually. Additionally, the table presents the four columns representing the confusion matrix, which include true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). With a recall of 0.757, the BMA_kegg source produced the best results. On the other hand, GO_MF is the worst-performing source, with a PU score of 1.284 and a recall of 0.509.

Table 2 - Results of the single source ANN. (TN: true negative, FP: false positive, FN: false negative, TP: true positive, r: recall, p: precision)

Source	TN	FP	FN	TP	r	p	PU score
PE	12169	3094	47	65	0.580	0.020	1.639
PH	12048	3215	48	64	0.571	0.019	1.531
RH	11461	3802	42	70	0.625	0.018	1.551
GO_BP	10522	1501	38	73	0.658	0.046	3.334
GO_CC	10803	1772	43	69	0.616	0.037	2.616
GO_MF	10303	2562	54	56	0.509	0.021	1.284
MAX_react	6473	1791	29	72	0.713	0.039	2.282
BMA_react	6914	1350	29	72	0.713	0.051	2.989
MAX_kegg	3501	870	24	42	0.636	0.046	1.970
BMA_kegg	2902	1469	16	50	0.757	0.033	1.676

Model Evaluation

The second part of the architecture involves the combination of the estimates obtained from a selection of five sources out of the total of ten obtained in the first part. Table 3 shows the first ten rows of the table containing the results of all thresholds for all source combinations, sorted by highest recall value (total of 132 rows). The table illustrates the four columns representing the confusion matrix, as well as the recall, the precision, and the PU scores.

The optimal combination was found to include the following sources: GO Biological Process (BP), GO Molecular Function (MF), and GO Cellular Component (CC) semantic similarity, Reactome pathway similarity aggregated with MAX, and gene co-expression level in kidney proximal tubule cells based on the Rho proportionality coefficient (rh).

Combining these variables yields an estimated number of 1121 new PGs, with a recall of 0.759 and a PU-score of 7.343, resulting from setting a threshold of 0.5.

Table 3 - Results of the top ten source combinations (higher recall values). (TN: true negative, FP: false positive, FN: false negative, TP: true positive, r: recall, p: precision, th: threshold).

Source combination	TN	FP	FN	TP	r	p	PU score	th
RH, BP, CC, MF, MAX_react	14142	1121	27	85	0.759	0.070	7.343	0.5
PE, BP, CC, MF, MAX_react	14170	1093	29	83	0.741	0.071	7.180	0.5
PH, BP, CC, MF, MAX_react	14223	1040	30	82	0.732	0.073	7.345	0.5
PE, BP, CC, MF, BMA_react	14285	978	31	81	0.723	0.076	7.594	0.5
RH, BP, CC, MF, BMA_react	14255	1008	32	80	0.714	0.074	7.210	0.5
PH, BP, CC, MF, BMA_react	14320	943	33	79	0.705	0.077	7.485	0.5
PE, BP, CC, MF, MAX_kegg	14269	994	33	79	0.705	0.074	7.129	0.5
PE, BP, CC, MF, BMA_kegg	14223	1040	33	79	0.705	0.071	6.836	0.5
PH, BP, CC, MF, MAX_kegg	14299	964	34	78	0.696	0.075	7.156	0.5
RH, BP, CC, MF, MAX_kegg	14237	1026	34	78	0.696	0.071	6.755	0.5

Second Part – Merging Results to Obtain the Overall Prediction.

The final values of the last layers of the five selected neural networks were utilized as inputs for the second phase of the model. An additional neural network was employed to calculate the final probability for each gene to be phenotype-related. Differing from the previous phase, 100 networks were trained with different weights initialization. The average and CI of the 100 pseudoprobabilities were then calculated for each gene, enhancing the result's robustness. The mean value of 0.5 was used as the threshold for identifying new potential PGs, resulting in 1170 genes potentially involved in renal tubule function. These were combined with the ontology-derived genes, leading to a total of 1269 PGs. Table 4 shows the confusion matrix of the results. The average recall obtained in the training phase is 0.762, with a 95% CI (0.734, 0.790). In the validation phase, the average recall is 0.735 with a 95% CI (0.708, 0.762). Finally, in the test phase, the average recall is 0.750 with a 95% CI (0.723, 0.778). Figure 11 shows the violin plots of the final probabilities of each gene. The two distributions on the left represent the PGs found in the literature, while the two on the right represent the unlabeled examples. The two pink violins represent the negative predicted genes, while those colored blue are the positive predicted ones. It is evident that the true NON_PGs have a much denser and uniform distribution, centered toward the mean, with relatively short tails.

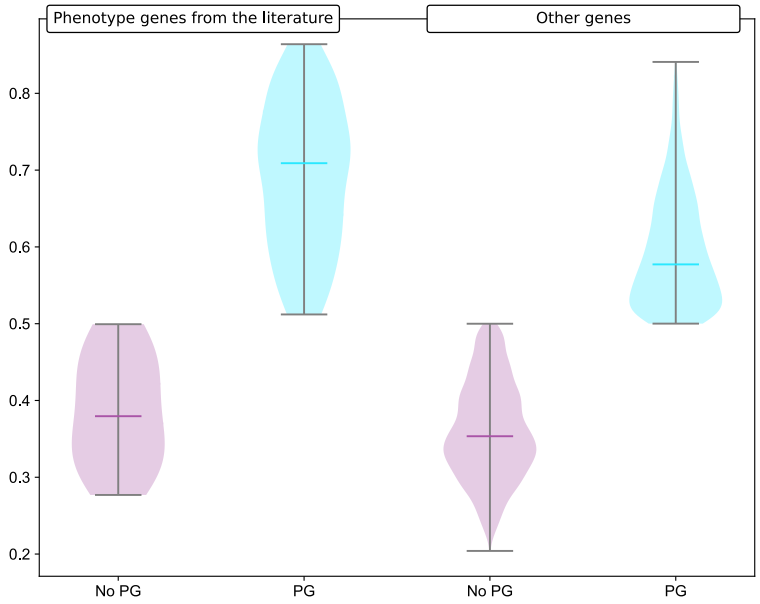


Figure 11 - Probability distributions for the phenotype genes obtained from the ML framework. The model probability is shown separately from the literature-derived PGs (on the left) and all the other genes (on the right).

Conversely, the distribution of new PGs (i.e., false positives, the relevant focus of our analysis) is markedly skewed, exhibiting a pronounced and prolonged tail extending towards the upper values. Notably, this same tail is not apparent in the context of true negatives, as evidenced by the lack of a similar tail on the side towards zero in the distribution of true NON_PGs.

Table 4 - Confusion matrix of prediction.

Observed / Predicted	Positive	Negative
Positive	83	*1170*
Negative	16	6450

Sensitivity Analysis of the Selected Source

We integrated information from five distinct sources to obtain the final probabilities, with the algorithm assigning distinct weights to each. In order to evaluate the influence and the impact of each source on the overall probability, we repeated the second-step training process five times, with one source being excluded each time. Figure 12 illustrates the recall performance of the five one-source-out models compared to that of the full model. The dots represent the mean recall, while the bars represent the corresponding 95% CI limits. The training, validation, and test sets results are shown for each model, colored green, light blue, and yellow, respectively. The dashed vertical lines

represent the limits of the 95% CI of the full model obtained on the test set. The table presents the values of recall mean with their respective 95% CIs obtained on the test sets. The p-value represents one-tailed Fisher's exact test results used to assess the significance of the difference between the one-source-out model and the full model. It can be observed that in almost all scenarios, the removal of a source leads to a deterioration in performance. In particular, the source that appears to exert the highest impact is Reactome, which, when removed, results in a notable decline in performance, with the mean recall of 0.75 reduced to 0.694 (p-value: 0.006, Figure 12). Even the removal of GO_CC and GO_BP decreases mean recall with slightly significant p-values, respectively 0.076 and 0.059. In contrast, the results remain roughly similar when the rho proportionality index is eliminated. Finally, the GO_MF removal results in a very small increase in performance, but this difference is not statistically significant, having a p-value of 0.652.

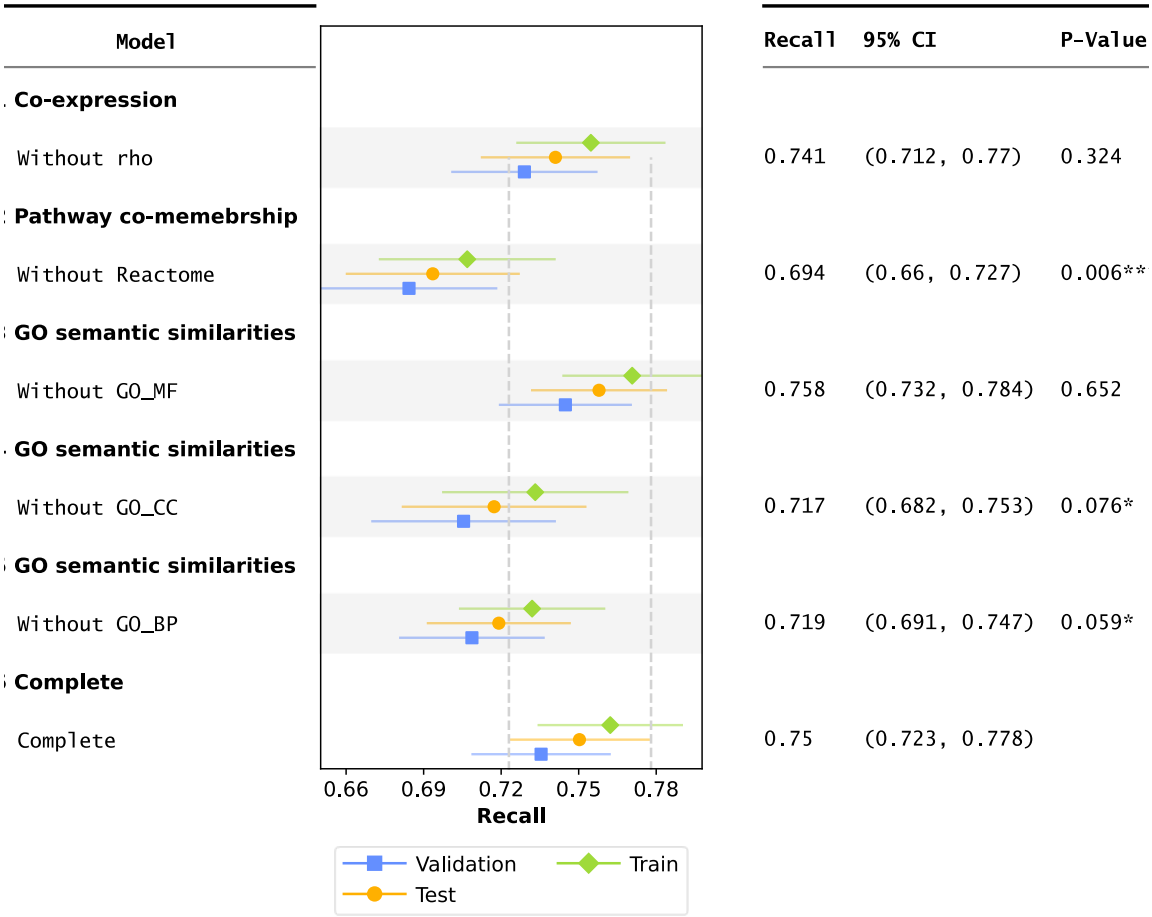


Figure 12 - Chart showing the mean (dot, square, and diamond) and the CI of the model recall from the complete ANN architecture and the models trained without the indicated data source. The vertical dashed lines indicate the CI of the complete model with all data sources.

To assess the impact of each source, we also analyzed the number of genes shared by the final model prediction and the single models. Figure 13 illustrates the upset plot with the intersection between the predicted PGs generated by the full model and those generated by the individual models. Our analysis revealed that the Reactome-only model identified 80 genes that were not identified by any other source. Furthermore, the largest intersections between the genes derived from the complete and single models consistently included Reactome. These results confirm the importance observed in previous results. In the second place, with a difference of only one gene from the first, we found the intersection between all sources, indicating a substantial number of genes demonstrating concordant results in all sources. Our findings reveal a notable overlap between the Reactome, the GO for Biological Process, and Molecular Function. This result can be attributed to the fact that the three measures in question encompass information comparable to that of the two remaining sources (rho and GO_CC). The first three measures, in fact, can be seen as related to biological processes and reactions, interactions between entities. On the other hand, the two excluded ones represent diametrically opposed concepts, such as expression levels and cellular components, and therefore, more single characteristics than group characteristics. Furthermore, the full model identified 31 PGs that were not identified from any of the other sources, highlighting the importance of integrating multiple sources of information.

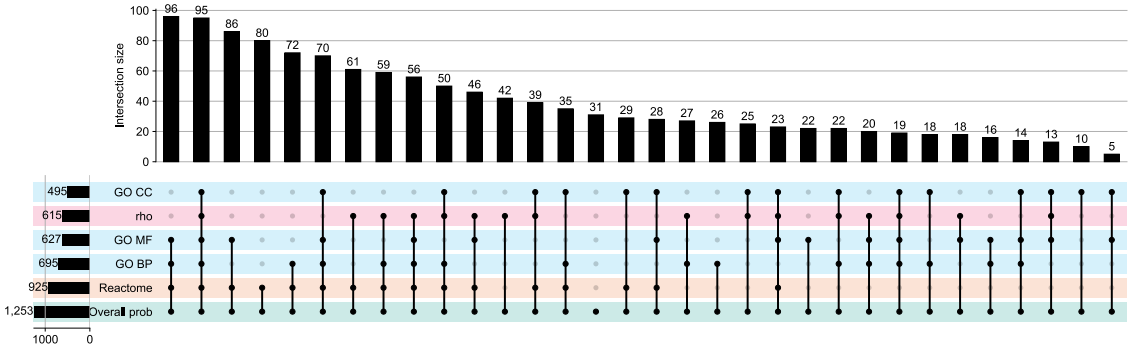


Figure 13 - Upset plot. Intersections among genes predicted as PGs in the final neural network (step 1 + step 2) and genes predicted as PGs in each of the first-step NNs.

Characterization of Phenotype Genes and External Validation

As a first positive outcome of external validation, it can be observed that the pathway analysis carried out on the Wikipathways database revealed significant enrichment in proximal tubule transport pathways. There was also a significant enrichment in pathways related to mitochondrial functions, such as the *ETC oxphos system in mitochondria* and *Mitochondrial long chain fatty acid beta-oxidation* (Figure 14). Furthermore, enrichment in the *oxidative phosphorylation* pathway was observed in both datasets, consistent with findings reported in the literature (Jackson et al., 1962; Sansanwal et al., 2010). (Figure 14, Figure 15).

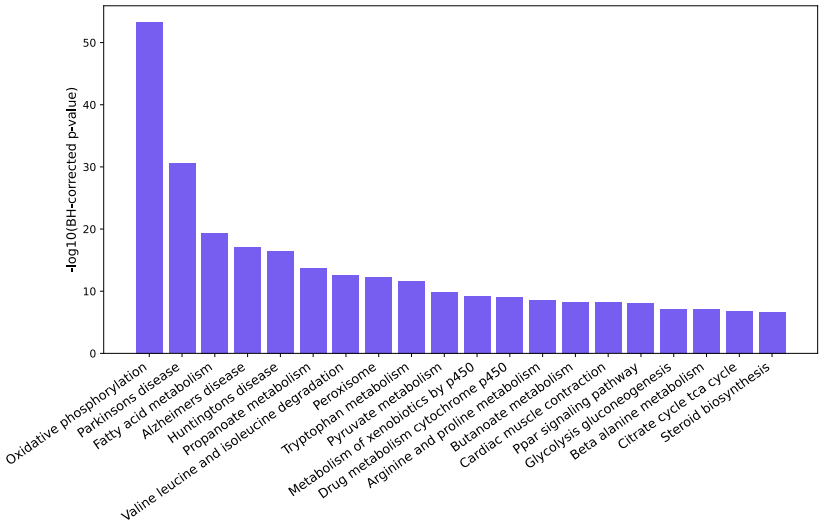


Figure 14 - WikiPathways enrichment analysis results. The BH-adjusted p-value of the top-twenty significant pathways is shown.

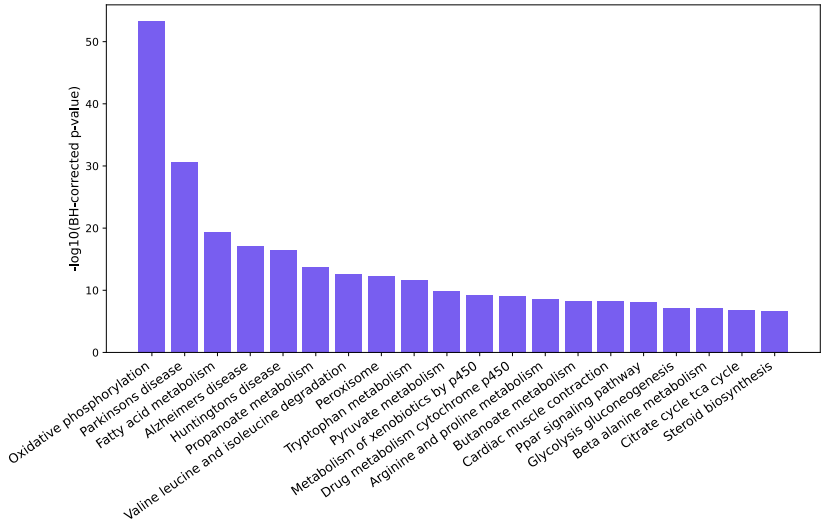
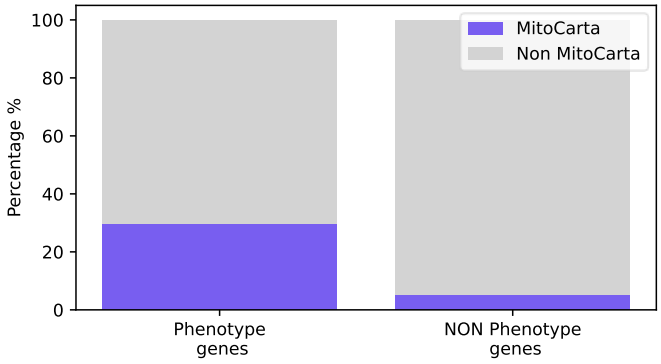


Figure 15 - KEGG enrichment analysis results. The BH-adjusted p-value of the top-twenty significant pathways is shown.

To investigate the relevance of the role of mitochondria, we have evaluated the overlap between PGs and genes found in MitoCarta, a database of proteins with mitochondrial localization. Of the 1269 PGs, 377 are also included in the MitoCarta database. We conducted a one-sided Fisher's exact test to assess whether the enrichment of the number of genes present in MitoCarta is statistically significant. The test yielded a p-value of 2.2e-16, indicating that the enrichment of mitochondrial genes among PGs is statistically significant. Figure 16 represents a plot with bars representing the percentage of mitochondrial genes present in the group of PGs and non-PGs. The table below shows the absolute frequencies.

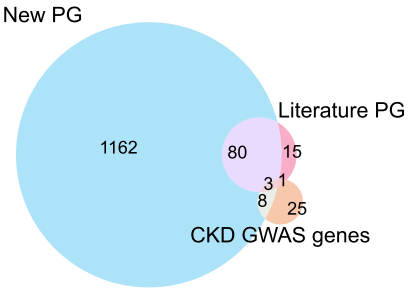


MitoCarta	377	346
Non MitoCarta	892	6140

Figure 16 - Frequency of proteins with mitochondrial localization in PGs and non-PGs according to MitoCarta annotation.

The overlap between PGs and genes associated with CKD by GWAS was evaluated to confirm the adherence of PGs in the context of kidney disease. Figure 17 shows the Venn plot between PGs in the literature, new PGs, and CKD-GWAS.

The table below represents the confusion matrix of PGs and CKD-GWAS. A one-tailed Fischer exact test was conducted, which yielded a p-value of 0.012, thereby confirming the significant enrichment.



	PG	Non PG
CKD_GWAS	12	25
Non CKD_GWAS	1257	6425

Figure 17 - Overlap between CKD-GWAS genes and PGs obtained from the literature or predicted by the model.

Disease-Related Molecules

After expanding the list of ontology-derived PGs with new PGs predicted by the model, the functional context of the proteins encoded by these genes was investigated, and their interaction with the molecules modulated in a mouse model of cystinosis was recently developed (Berquez et al., 2023; Festa et al., 2018). The transcriptomic, proteomic, and metabolomic profiles of the renal proximal tubule segments of the $CTNS^{KO}$ mouse model of cystinosis were utilized to identify the set of DMs. These include 824 upregulated and 502 downregulated expressed genes (transcriptomics), 122 upregulated and 178 downregulated proteins (proteomics), 98 upregulated and 45 downregulated metabolites (metabolomics). Figure 18A represents the volcano plot, which illustrates the LogFC and the log base 10 of the p-value (log10pval) for proteins. Instead, Figure 18B displays volcano plots depicting the abovementioned parameters for genes, while Figure 18C shows plots of the same parameters for metabolites.

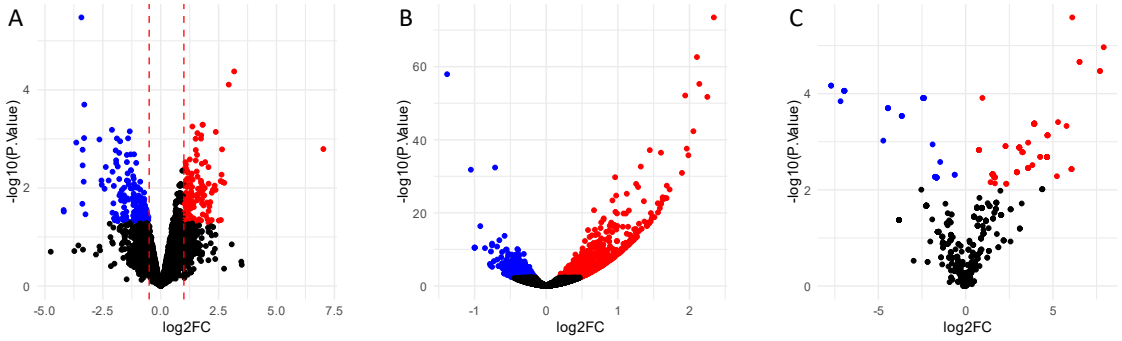


Figure 18 - Volcano plot of differential (A) protein abundance analysis, (B) gene expression analysis, and (C) metabolite abundance analysis. Red and blue points indicate upregulated and downregulated molecules. The vertical bars in (B) indicate the fold change thresholds (-0.5 and 1) used to select the significant proteins.

Multilayer Network

We then built a multi-layer molecular network including TF-gene interactions from Dorothea (Garcia-Alonso et al., 2019), PPI from OmniPath (Türei et al., 2016) and Reactome (Milacic et al., 2024), and metabolite-protein interactions from Metabolic Atlas (J. L. Robinson et al., 2020) and HMDB (Wishart et al., 2022). Genes without evidence of expression in proximal tubule cells were excluded from the network. The final network has 14,936 nodes and 193,992 directed edges. The nodes can be divided into three categories: genes, metabolites, and TFs. In the final network, 73.42% of the nodes are genes, 24.30% are metabolites, and 2.28% are TFs. Of the 1269 PGs, 1241 were mapped onto the network. Of these, 7.82% were derived from literature, while the remaining 92.18% were derived from ANN predictions. Regarding DMs, 1296 unique molecules were mapped on the network as belonging to at least one of the three

categories. In detail, there are 302 disease proteins, 62 disease metabolites, and 954 disease genes. Of the 193,399 edges, 12.60% of the interactions are attributable to Dorothea, 22.03% represent interactions from metabolic patterns, and 65.37% are simple protein-protein interactions. Table 5 provides the absolute frequencies of the nodes and edges categories in the final network.

Table 5 - Multilayer network composition

Category	Description	Count
Nodes source	Genes	10966
	Metabolites	3630
	Transcription Factors	340
	Total	14936
Phenotype genes	From literature	97
	Predicted from ANNs	1144
	Total	1241
Disease genes	Proteomics	302
	Metabolomics	62
	Rna-seq data	954
	Total	1296
Edges source	Dorothea	24442
	HMDB	5203
	Metabolic Atlas	33856
	Metabolic Atlas + HMDB	3679
	Omnipath	26225
	Omnipath + Dorothea	311
	Omnipath + Reactome	2892
	Omnipath + Reactome + Dorothea	103
	Reactome	97079
	Reactome + Dorothea	202
	Total	193992

Identification of network modules and enrichment analysis

The network was partitioned into communities of highly interacting nodes, with a maximum size of 100 nodes. This division, as described in the methods section, yielded 271 modules. The enrichment of DMs and PGs in each module was analyzed. 128 modules were identified as enriched in disease genes, 49 in disease proteins, and six in disease metabolites. Instead, 32 modules were found to be enriched in PGs. The 22 modules that were found to be enriched in both PGs and at least one class of DMs (PG-DM modules) were subsequently analyzed. Figure 19 shows the 22 enriched modules, with the blue bars representing the DMs frequency in each module, the purple bars

representing the PGs frequencies, and the orange squares representing the normalized adjusted p-value (min-max normalization is performed independently for each omics data and only for significant (<0.05) p-values.

Table 6 contains, for each module, the number of PGs and DMs for each of the three categories separately and the adjusted p-value of enrichment. Additionally, the last columns contain the total module numerosity, the number of metabolites, the number of genes, and the relative frequencies of DMs and PGs.

Table 6 - Enriched modules composition.

Modules	p.adj PG	N PG	p.adj Meta	N D Meta	p.adj Prot	N D Prot	p.adj Transc	N D Transc	N TOT	N Meta	N Genes	Freq rel DM	Freq rel PG	Freq abs DM
M_12	<0.001	19	0.024	3		0	0.108	1	65	28	37	0.062	0.292	4
M_122	<0.001	38		0	0.018	2	<0.001	5	90		90	0.078	0.422	7
M_130	<0.001	24	0.190	1	0.091	1	<0.001	6	71	29	42	0.113	0.338	8
M_131	<0.001	24	0.190	1	0.018	2	<0.001	6	83	7	76	0.108	0.289	9
M_132	<0.001	24	<0.001	6	0.018	2	<0.001	8	98	50	48	0.163	0.245	16
M_134	<0.001	8	0.190	1		0	<0.001	4	15	4	11	0.333	0.533	5
M_137	<0.001	29		0	0.018	2	0.003	3	88	28	60	0.057	0.33	5
M_14	0.001	8		0	0.018	2		0	18	5	13	0.111	0.444	2
M_143	0.039	12		0	0.091	1	<0.001	6	62	22	40	0.113	0.194	7
M_151	<0.001	23		0		0	<0.001	5	55	18	37	0.091	0.418	5
M_26	0.006	12		0	0.091	1	<0.001	4	49	25	24	0.102	0.245	5
M_265	<0.001	77		0	0.018	2	<0.001	18	253	4	249	0.079	0.304	20
M_33	<0.001	29		0	0.091	1	<0.001	9	49	6	43	0.204	0.592	10
M_34	<0.001	8		0	0.091	1	<0.001	4	17	5	12	0.294	0.471	5
M_47	0.008	13	0.087	2		0	0.016	2	58	27	31	0.069	0.224	4
M_49	0.031	12		0		0	<0.001	6	60	25	35	0.1	0.2	6
M_5	<0.001	28	0.007	4	0.018	2	<0.001	8	96	34	62	0.146	0.292	14
M_54	<0.001	10		0	0.091	1	0.016	2	26		26	0.115	0.385	3
M_72	0.001	13		0	0.001	4	0.016	2	48	1	47	0.125	0.271	6
M_89	<0.001	13		0		0	<0.001	9	31	4	27	0.29	0.419	9
M_9	<0.001	47		0	0.018	2	0.016	2	71	5	66	0.056	0.662	4
M_95	0.031	6		0	0.091	1	0.003	3	19	1	18	0.211	0.316	4

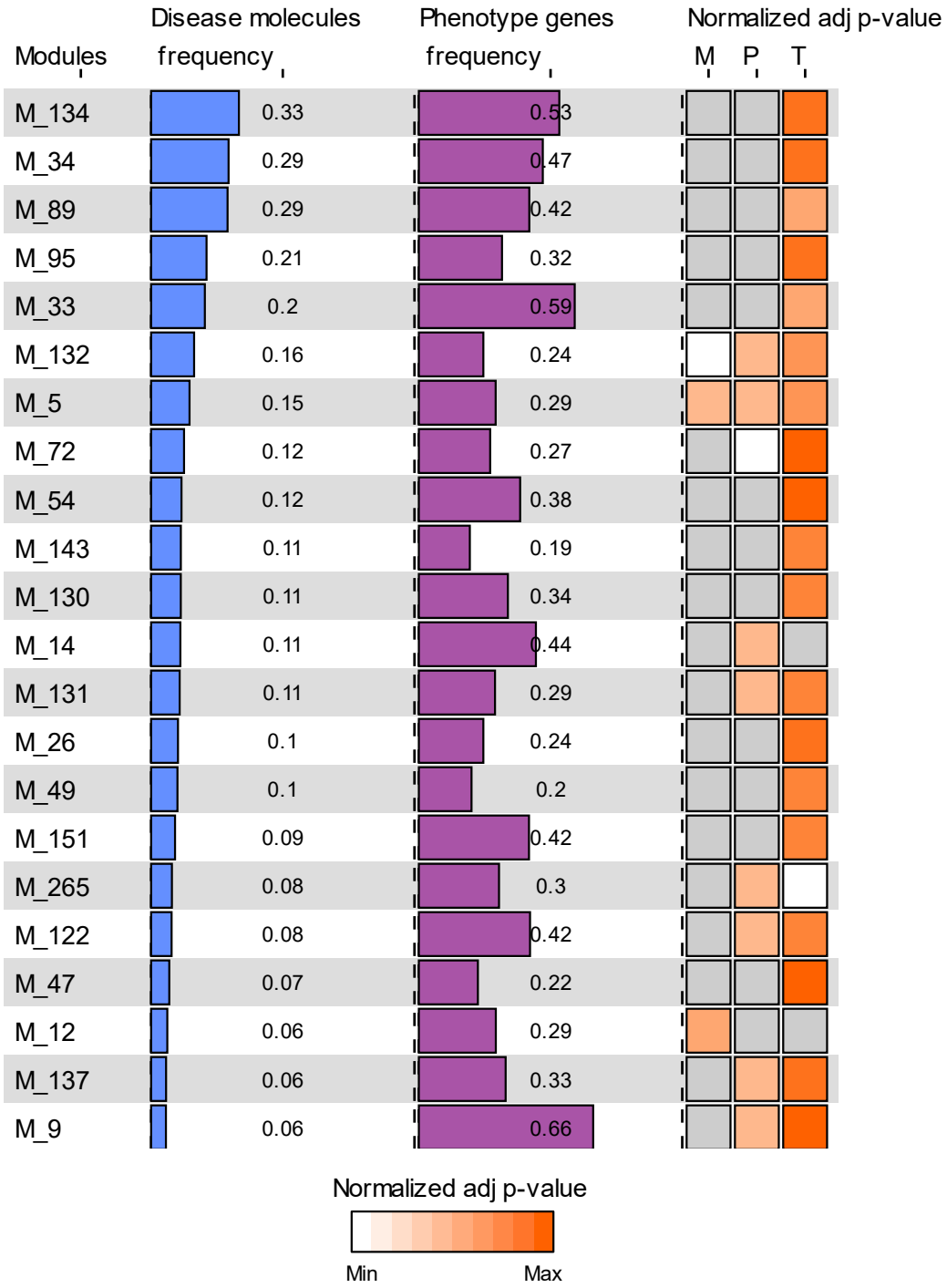


Figure 19 - PG-DM module characterization. For each PG-DM module are reported: the frequency of disease molecules (light blue), the frequency of phenotype-related genes (magenta), and the normalized -Log$_{10}$ BH-corrected p-value (orange) of the module enrichment in metabolites (M), proteins (P), and transcripts (T). A gray square indicates a non-significant enrichment (BH-corrected p-value > 0.05).

Pathway and gene-set analysis of the modules

Using pathway and gene-set enrichment analysis, we investigated the PG-DM modules' biological function and the identified proteins' cellular localization. Oxidative phosphorylation emerged as the most significant Hallmark pathway shared across modules (present in modules M_130, M_134, M_151, M_33, M_54, and M_9), and the mitochondrion was the most enriched cellular component. (Figure 20, Figure 21, Table A8 and Table A9).

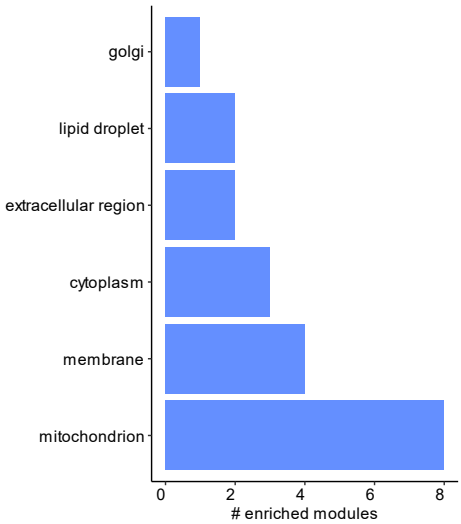


Figure 20 - Most represented cellular compartments of the proteins within the PG-DM module. For each PG-DM module, we selected the most enriched GO Cellular Component term and identified its parent term.

Module M_130 includes alpha-ketoglutarate (oxoglutaric acid), a metabolite known to couple CTNS deficiency with defective autophagy, apoptosis, and PT dysfunction (Jamalpoor et al., 2021). In line, alpha-ketoglutarate levels were significantly increased in primary cells derived from microdissected PT segments of $CTNS^{KO}$/cystinosis-affected mice. KEGG pathway analysis of this module and modules M_265 and M_33 identified significant enrichment in the pathway “Proximal tubule bicarbonate reclamation” (Figure 22), which is of particular relevance for cystinosis because the patients, among the symptoms, present with metabolic acidosis (Emma et al., 2014).

Among the other modules associated with oxidative phosphorylation, module M_33 includes several subunits of the vacuolar ATPase (V-ATPase), an enzyme that mediates the acidification of endosomes and lysosomes. Recent findings indicate that active mTORC1 inhibits this process, thereby controlling the lysosomal catabolic activity (Ratto et al., 2022).

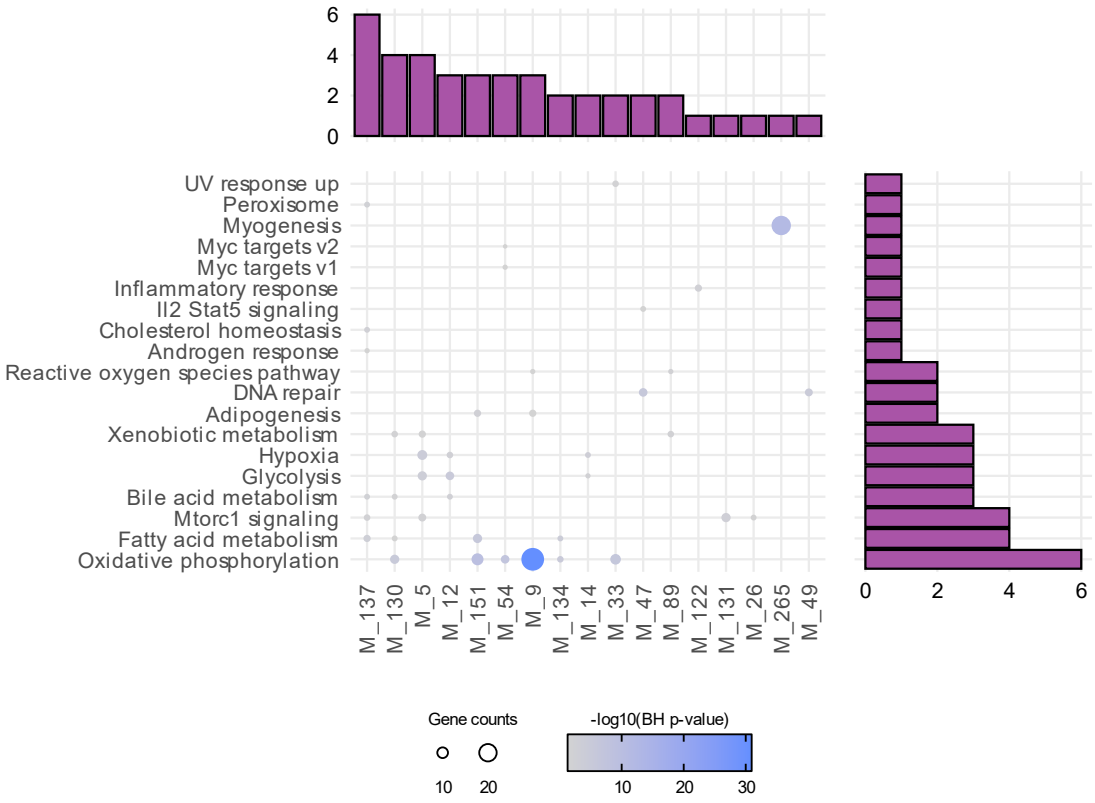


Figure 21 - The significantly enriched Hallmark pathways are shown for each PG-DM module. The color of the circles is proportional to the -log(10) BH-corrected p-value of the enrichment, and the size is proportional to the number of genes. The purple bars on the right and on the top indicate the number of significant modules per pathway and the number of significant pathways per module, respectively.

Interestingly, a preprint study reported a patient with ATP6V1B1 gene deficiency and proximal tubulopathy and confirmed by in vitro experiments that ATP6V1B1-/- cells have impaired autophagy (Jamalpoor et al., 2024).

Other modules, even if not enriched in proteins from the oxidative phosphorylation pathway, included molecules from highly related pathways. For example, M_89 is enriched in xenobiotic metabolism and reactive oxygen species (ROS) pathway and includes glutathione, a major antioxidant against oxidative stress, already associated with cystinosis by previous studies (Wilmer et al., 2010, 2011). In agreement with the literature, in our metabolomic dataset, glutathione levels were downregulated in PT cells, albeit not significantly (Log2FC = -1.41, p-value = 0.25). In addition to module M_89, Glutathione metabolism was significantly enriched by KEGG analysis in modules M_131, M_132, and M_49 (Figure 22).

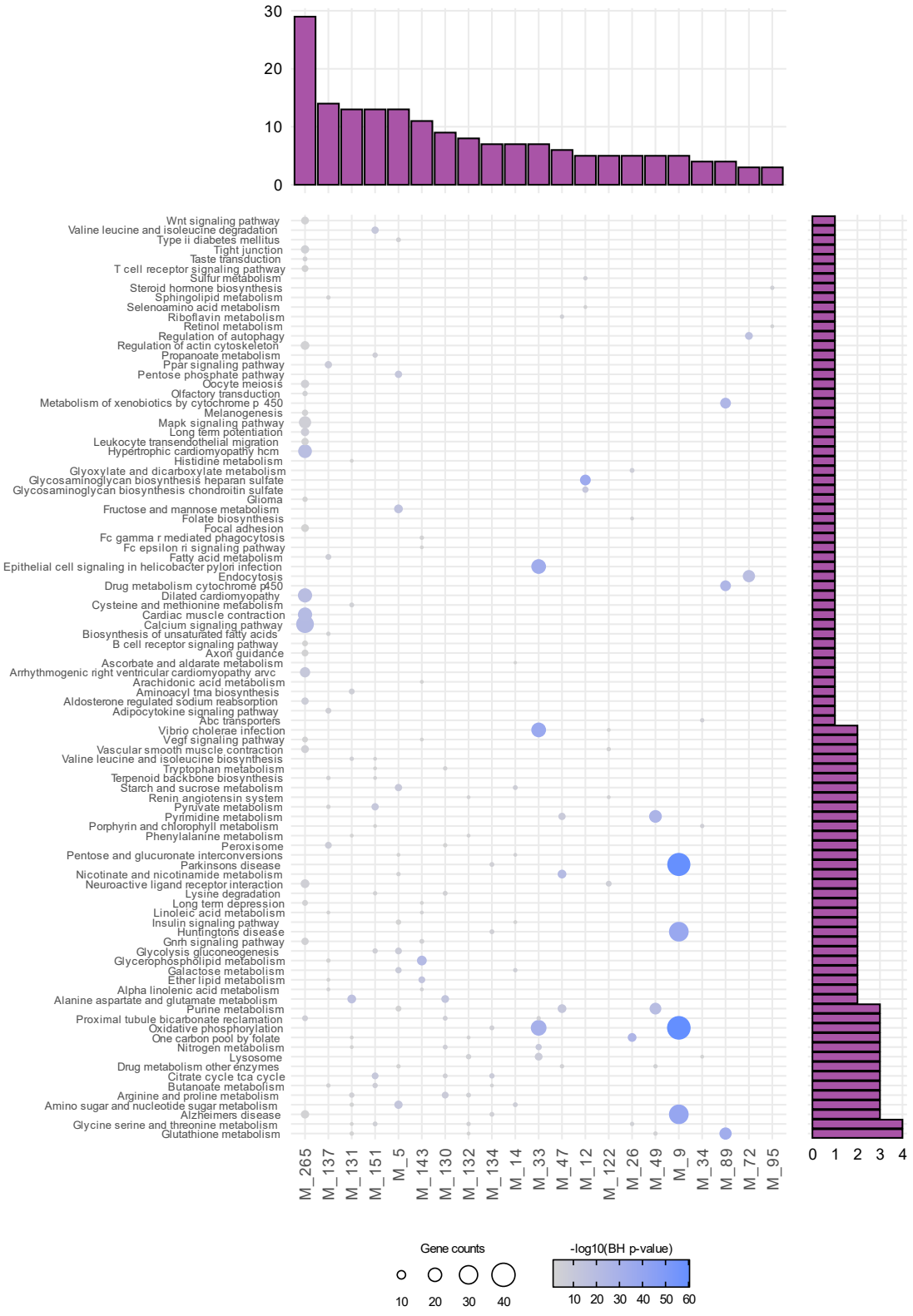


Figure 22 - The significantly enriched KEGG pathways are shown for each PG-DM module. The color of the circles is proportional to the -log(10) BH-corrected p-value of the enrichment, and the size is proportional to the number of genes. The purple bars on the right and on the top indicate the number of significant modules per pathway and the number of significant pathways per module, respectively.

Hallmark gene-set enrichment analysis also revealed a close connection between mTORC1 signaling and cystinosis-dysregulated metabolism. Four modules (M_131, M_137, M_26, and M_5) had mTORC1 signaling among the significantly enriched Hallmark pathways (Figure 21). For M_131 and M_26, mTORC1 is the most enriched pathway with 9/76 and 4/24 genes, respectively (Figure 21). In addition to the mTORC1 connection, both modules are related to amino acid metabolism and redox defense, with module M_131 including glutamate and module M_26 including several folate metabolism-related molecules (Figure 21).

Drug Repurposing Analysis

To identify novel therapeutic opportunities for cystinosis, we mapped the human targets of existing drugs with known pharmacological action onto the network and identified those mapping to any of the 22 PG-DM modules. To mitigate the risk of selecting drugs that may exacerbate the disease instead of treating it, we evaluated the consistency between the pharmacological action and the module dysregulation for each drug-target pair by considering the FC of the differentially expressed molecules. As DR candidates, we kept only inhibitors acting on targets within upregulated modules (with a higher number of upregulated molecules compared to the downregulated ones), whereas, for activators, we kept those with a target within downregulated modules (with a higher number of downregulated molecules compared to the upregulated ones).

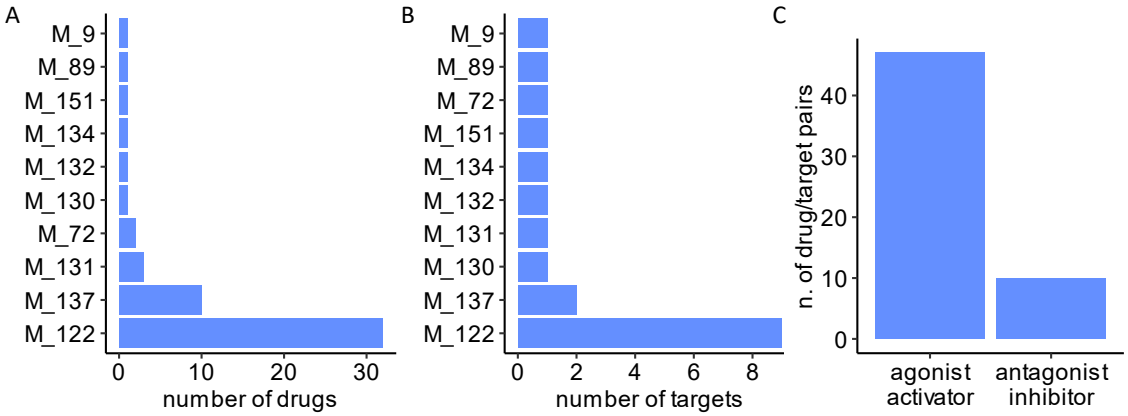


Figure 23 - Number of (A) drug repurposing candidates and (B) drug targets for each module, as well as (C) number of agonists and antagonists among the identified drug/target pairs.

We identified 51 drugs targeting 19 proteins in ten PG-DM modules (Figure 23Figure 23A-B, Table A10). In ten cases, the drugs have an inhibitory effect on their targets, while in the others, they induce the activation of the target (Figure 23C). As expected, the targets mainly relate to metabolism (four oxidoreductases and four metabolite-converting enzymes, Figure 24). In addition, the ion channel class includes three subunits of the gamma-aminobutyric acid (GABA) receptor (GABRB2, GABRA2, GABRB3), possibly participating in the renal GABAergic system (Takano et al., 2014; Wildman et al., 2023).

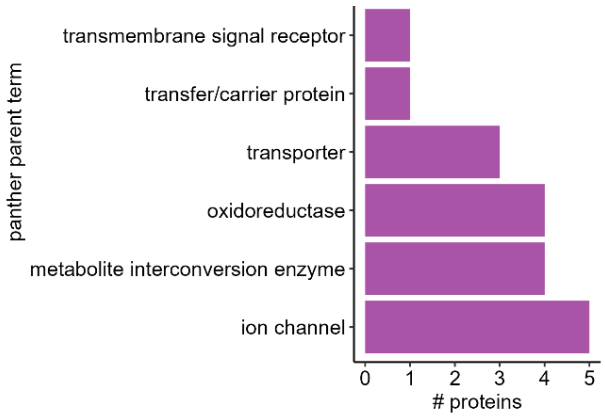


Figure 24 - Protein classification of the drug targets using the Panther database.

Among the drugs connected to the top PG-DM modules, acetylcysteine, also known as N-acetylcysteine (NAC), was identified as a repurposing candidate linked to M_89, the glutathione-related module (Figure 26). Noteworthy, this molecule has already been tested in cystinosis clinical trials, and it was shown to reduce oxidative stress and improve renal function[20] (Pache De Faria Guimaraes et al., 2014; Sahasrabudhe et al., 2023). Coenzyme Q10 (CoQ10 - ubidecarenone) was instead linked to modules M_9 and M_134, which are related to oxidative phosphorylation and ROS pathways (Figure 25). In agreement with these findings, CoQ10 is an essential component of the mitochondrial electron transport chain, and, similar to NAC, it acts as an antioxidant. CoQ10 supplementation has already been proposed to treat various diseases and to counteract aging since the endogenous synthesis showed an age-dependent reduction (Hernández-Camacho et al., 2018; Marappan & Marappan, 2020).

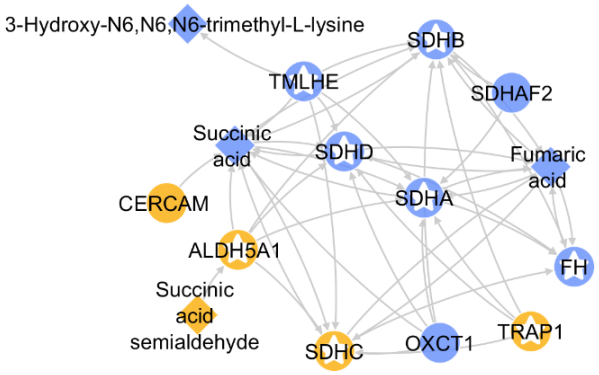


Figure 25 - Network visualization of the module M_134. The shape of the nodes indicates the class of molecule represented: diamonds are metabolites, and circles are genes/proteins. Yellow nodes represent disease molecules (differentially expressed abundant). A white star within the circle indicates a phenotype gene/protein. The SDHA gene is a target of ubidecarenone (CoQ10).

[20] https://classic.clinicaltrials.gov/ct2/show/NCT01614431

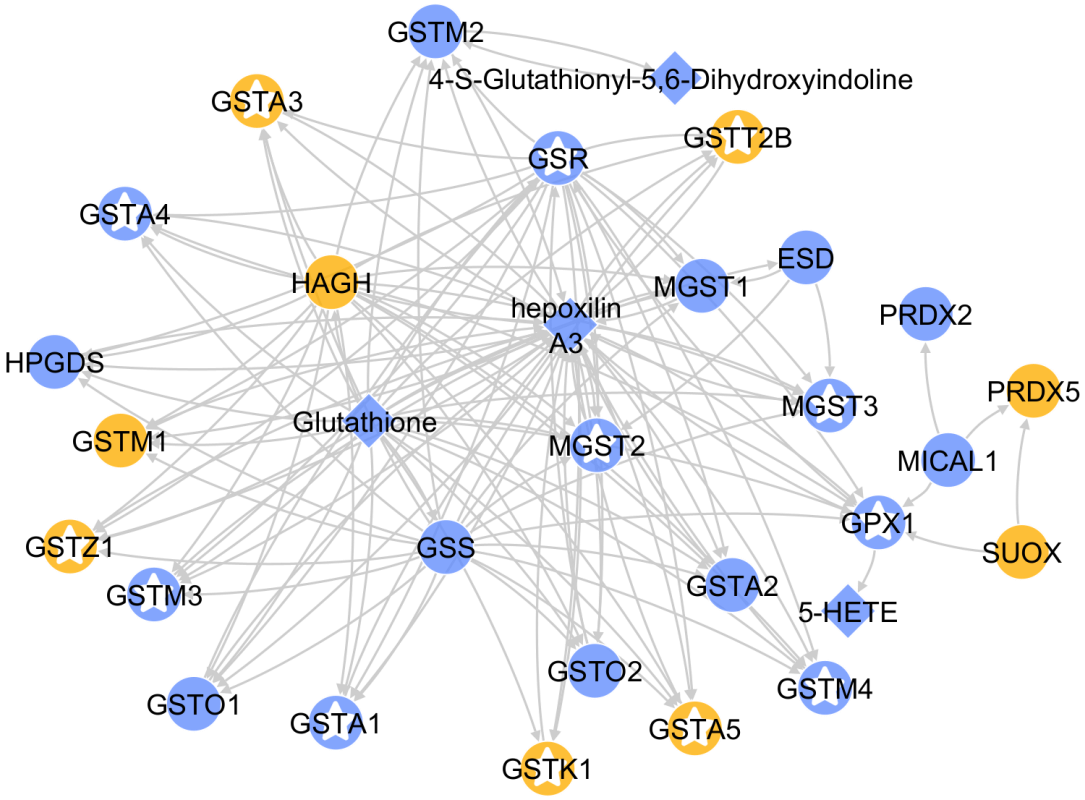


Figure 26 - Network visualization of the module M_89. The shape of the nodes indicates the class of molecule represented: diamonds are metabolites, and circles are genes/proteins. Yellow nodes represent disease molecules (differentially expressed/abundant). A white star within the circle indicates a phenotype gene/protein. The GSS gene is a target of N-acetylcysteine (NAC).

Discussion

Cystinosis is a rare lysosomal storage disorder with unmet medical needs. To contribute to understanding the molecular mechanisms involved in the disease and support the development of new therapies, we devised a computational framework that integrates different data types from heterogeneous sources and identifies DR candidates.

Our approach started by defining a list of relevant genes for the clinical phenotype of interest, in our case, the renal tubule function. Indeed, even if cystinosis is a systemic disease involving multiple organs and tissues, kidneys, and specifically the epithelial cells lining the kidney proximal tubules, are primarily affected, and their impairment leads to renal failure requiring kidney transplantation, which is itself associated with severe complications (Gahl et al., 2002). For this reason, we decided to focus our study on this cell type, with the final aim of pointing out candidate therapies acting on dysfunctional molecular mechanisms of these cells, possibly preventing kidney failure.

Berquez et al. previously showed that pathway enrichment analysis of transcriptomic differential gene expression results mainly identified cell cycle and proliferation as significantly upregulated pathways in $CTNS^{KO}$ mouse model PT cells. This is, however, a consequence of PT cell dedifferentiation and loss of their endocytic function. Starting the search for new therapeutic opportunities from these pathways would have led us to focus on downstream effects, not disease-specific. Interestingly, a study in yeast showed that mRNA profiling frequently detects metabolic responses, while genetic screens tend to identify response regulators, and connecting the two was proposed as a strategy to identify drug targets (Yeger-Lotem et al., 2009). The gene-phenotype associations reported in the ontologies we parsed to identify renal tubule genes can be conceptually paired with the results of yeast genetic screens. Indeed, these associations were included in the ontology annotations because mutations causing human diseases or animal models that present renal tubular dysfunction exist. Based on this, our approach aimed to connect the genes associated with renal tubule function to the results of the omics profiles of cystinosis PT cells.

We developed an ML model to expand the ontology-derived list of PT genes. ML-based approaches are increasingly applied in the drug discovery process (Vamathevan et al., 2019). Among the ML methods based on gene-gene similarity, GO similarity between drug transcriptional signatures has been employed for drug target prediction (Chen et al., 2024). Furthermore, gene-gene similarity has been utilized by a neural network model to predict pain targets (M. Jeon et al., 2021).

The ML predictions were mapped onto a multi-layered molecular network together with the disease molecules. The main advantage of using a network approach for data integration is the possibility of using data generated from different samples (Cai et al., 2022).

Our results revealed a key role of mitochondrial processes, specifically oxidative phosphorylation, in the disease. These findings are in line with previous studies highlighting the role of mitochondrial oxidative stress in the kidney manifestations of the disease (Festa et al., 2018). Indeed, the electron transport chain in oxidative phosphorylation is a main source of mitochondrial ROS, and this has a major impact on kidney epithelial PT cells, given their high mitochondrial density due to the energy demands needed to support the reabsorption of molecules from the glomerular filtrate (Khan et al., 2020).

In this study, cystinosis omics data were derived from an animal model. Even if this model well recapitulates the disease, specifically the PT dysfunction, the availability of patient-derived omics data would enhance our understanding of the disease pathogenesis, allowing us to deepen our analysis, potentially highlighting disease pathways not captured using animal models alone. Moreover, in future studies, instead of focusing only on existing drugs with a repurposing-based approach, the PG-DM modules could be further investigated to identify candidate drug targets and start an ad-hoc de novo drug design process, possibly relying on ML techniques (Atz et al., 2024; Kamya et al., 2023).

3. Quantifying Signaling Model Diversity Through Solver-Agnostic Solution Sampling with CORNETO

The following chapter presents the results of a research project conducted during a visiting period at SaezLab [21]in the Institute for Computational Biomedicine of Heidelberg University, Germany, from September 2023 to February 2024.

Introduction

In the preceding chapter, we investigate identifying potential drug targets and repurposing opportunities through applying ML methodologies, network topological properties, and functional annotation. The growing literature underscores the promise of ML in pinpointing repurposing opportunities within drug discovery (Cong & Endo, 2022; Priya et al., 2022). Nonetheless, the interpretability of results derived from ML and DL methods remains a challenge (Aittokallio, 2022; Ekins et al., 2019; Elbadawi et al., 2021). This problem necessitates integrating approaches that offer a more profound mechanistic understanding of disease biology and drug effects, thereby supporting the accuracy and reliability of drug target identification (Mitsos et al., 2009).

Mechanistic models furnish invaluable insights into biological mechanisms, com(Alber et al., 2019; Erdem & Birtwistle, 2023; Zampieri et al., 2019). In this chapter, we present the challenge of inferring mechanistic signaling networks from omics data due to the non-identifiability of related ILP problems, which results in multiple valid solutions.

Cell Signaling and Its Importance

Cells, as life's basic units, continuously encounter internal and external stimuli. These stimuli can range from environmental alterations, such as growth factors or toxins, to internal factors, like DNA damage. The cellular response to these stimuli is facilitated through an intricate communication mechanism known as cell signaling. This communication is pivotal for intercellular and extracellular interactions, with signaling pathways comprising a cascade of interactions and chemical reactions that transduce signals. Receptors on cell membranes bind specific molecules, initiating intracellular signaling cascades mediated by proteins or secondary messengers, which eventually end in the expression of specific genes (Nair et al., 2019). Each signaling pathway is defined by a unique ensemble of signaling proteins with varying roles. Abnormalities in these pathways can precipitate a range of diseases, from cancer to metabolic and immune disorders (Barata & Oliveira, 2019). Hence, comprehending the regulatory mechanisms

[21] https://saezlab.org/

governing altered signaling pathways is crucial for understanding disease pathology and devising effective therapeutic strategies (Kochen et al., 2022).

Modeling Signaling Networks

A signaling regulatory network can be represented as an interaction graph (Klamt & von Kamp, 2009), with nodes denoting signaling proteins and edges representing interactions between them. In cell signaling, the interactions are predominantly represented with directed edges, categorized as activatory or inhibitory. The signed representation is helpful to describe the causal relationships between proteins in the network. Activatory interactions tend to upregulate the activity of downstream proteins, while inhibitory interactions can down-regulate them.

A prevalent approach to modeling mechanistic signaling networks involves refining interactions from a Prior Knowledge Network (PKN) by integrating data-driven constraints. (Ideker et al., 2001; Saez‐Rodriguez et al., 2009; Terfve et al., 2012). This process leverages combinatorial optimization and mathematical solvers to derive context-specific subnetworks called network inference. The CARNIVAL method (Liu et al., 2019) is used to identify the upstream regulatory signaling processes that drive gene expression changes inferred from gene expression profiling experiments. It is a method that integrates PKN of protein-protein interactions obtained from various sources, with TFs targets, formulated as an integer linear programming (ILP) problem.

Challenges in Network Inference

However, identifying an optimal solution does not guarantee its exclusivity. The inherent non-identifiability of network inference problems implies the existence of multiple equally valid solutions (Guziolowski et al., 2013). In addition, beyond the possible optimal alternative solutions, there may be alternative solutions that do not reach the optimality threshold but may provide critical biological interpretations or hypotheses. In other words, for a given experimental data, PKN, and perturbations, exist an unknown amount of possible sub-networks (solutions) that are equally valid (optimal) in terms of agreement with experimental data, but only one of these solutions (without knowing which one a priori) is returned by Carnival, as also the commonly used context-specific reconstruction methods. Overlooking these alternative solutions can yield an incomplete understanding of the hypothesis space of consistent mechanistic signaling networks (Siegenthaler & Gunawan, 2014). Furthermore, identifying all potential alternative solutions is a challenging, if possible, task, particularly when large PKNs are involved due to the limitations of current technology (Guziolowski et al., 2013).

Despite increasing recognition of these limitations, there remains a need for studies that explore the computational difficulties and offer methods to explore the optimal space of alternative networks. One pertinent study, "DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks" (Rodríguez-Mier et al., 2021), proposes

four methodologies for enumerating context-specific alternative solutions within metabolic networks, all implemented in a MATLAB library. This study serves as a foundation for this work.

Our Contribution: Sensitivity Analysis Through CORNETO

Building upon the ideas from DEXOM, we have extended one of the methodologies to all mechanistic network inference problems, integrating this methodology into CORNETO—a novel package under development in the Saez lab. The purpose of CORNETO is to create a unified network inference framework implemented in Python, which will address many common network inference problems encountered in biology (Rodríguez Mier et al., 2024). One of the key contributions of this work is the implementation of an algorithm to explore the space of alternative solutions and to conduct sensitivity analysis on the optimal solution. Sensitivity analysis refers to a set of methods for evaluating the impact of changes in an input variable on the output of a model. The purpose of this type of analysis is to determine how uncertainty in the output of a mathematical model or system can be attributed to different levels in its inputs, thus identifying which variables have the most significant influence on the results, as well as quantifying the magnitude of their impact. The specific aim of the sensitivity analysis in this study was to evaluate and quantify the importance of the nodes and edges of the optimal solution, exploring the space of alternative solutions. By introducing constraints to the original problem, slight input perturbation is generated, thereby enabling the exploration of a subspace of alternative solutions proximal to the optimal solution. This process allows us to determine whether, upon the elimination of a node, the optimizer identifies distinct or analogous alternative solutions. If eliminating a node result in the generation of substantially different alternative solutions, it can be hypothesized that the node plays a critical role in the signal transduction pathway and is therefore pivotal to the network's function. Conversely, if the removal of a node results in solutions that are highly similar to the original, this suggests that the node may not be as essential in the solution and could potentially be replaced by alternative pathways that are only minimally longer, yet still lead to comparable biological interpretations. This sensitivity analysis is essential for assessing the robustness of the inferred network and provides deeper insights into the behavior of the model under varying conditions. The assessment of a gene's or reaction's significance within a solution contributes to provides valuable insights into the functional relevance and regulatory role of specific components within the network. This, in turn, offers valuable guidance in prioritizing targets for further experimental validation, thereby facilitating a more focused investigation into the underlying biological mechanisms driving the system.

Material and Methods

CORNETO and Carnival Method

CORNETO (Constraint-based Optimization for the Reconstruction of Networks from Omics) is an ongoing network inference framework developed in the Saez laboratory at Heidelberg University. Implemented in Python, CORNETO leverages onstrained optimization and mathematical programming to transform diverse network inference challenges common in biology into unified mathematical formulations expressed through flow networks. This approach provides a flexible and modular infrastructure for addressing diverse network reconstruction problems in a unified manner. In this project, it represents the fundamental framework on which the functions were developed. Specifically, it was employed by involving the CARNIVAL method (Causal Reasoning pipeline for Network identification using Integer VALue programming), which allows contextualizing causal signaling networks from expression footprints.

As previously described, cells respond to external stimuli, such as the detection of specific extracellular molecules or alterations in the cellular microenvironment, through signaling cascades that ultimately regulate gene expression, protein activation, and, consequently, the function or state of the cell in response to the perturbation. To represent the cellular response and the signal cascades, the Systems Biology Graphical Notation standard is used, according to which the signal is transmitted through pathways and networks of proteins. Specifically, the activity flow representation is employed, where each molecule in the network (node) is connected by directed edges that can be either activators (arrow) or inhibitors (hyphen).

CARNIVAL is a computational tool designed to infer the upstream signaling network responsible for the observed alterations in gene expression in RNA-seq experiments. It is based on the ILP algorithm of the Melas et al. method (Melas et al., 2015). This method integrates a signed, directed PKN of protein-protein interactions with the top estimated TFs activity changes. The standard CARNIVAL pipeline is applied in the context of a known perturbation (e.g., disease or drug treatment). It connects the resulting downstream changes in TFs activity with upstream signaling components through the PKN up to the perturbation targets provided by the user as being activated or inhibited. Figure 27 provides a graphical representation of the CARNIVAL method.

The prior knowledge network forms a graph G consisting of a set of nodes $\mathrm{V} = \{v_j, \mathrm{j} = 1, \ldots, \mathrm{n}_s\}$ representing proteins and a set of signed, directed edges $\mathrm{E} = \{e_i, \mathrm{i} = 1, \ldots, \mathrm{n}_r\}$ representing interactions between proteins. Each edge is characterized by an ordered pair of distinct source and target nodes $e_i = (\mathrm{s}_i, t_i)$, $\mathrm{s}_i, t_i \in \mathrm{V}$, and has an interaction sign $\sigma_i \in \{-1, 1\}$. When s_i has an inhibitory effect on t_i, then $\sigma_i = -1$, while when s_i has an activating effect on t_i, then $\sigma_i = 1$.

The user provides, in addition to the PKN:

1. A set of perturbed input nodes $V^I \subset V$ with corresponding activation states $I_j \in \{1, -1\}$.
2. A set of measured nodes $V^M \subset V$ with corresponding measurement scores. These measures are usually the set of predicted TFs activities, with $\alpha_j = abs(score)$ and $m_j = sign(score)$.

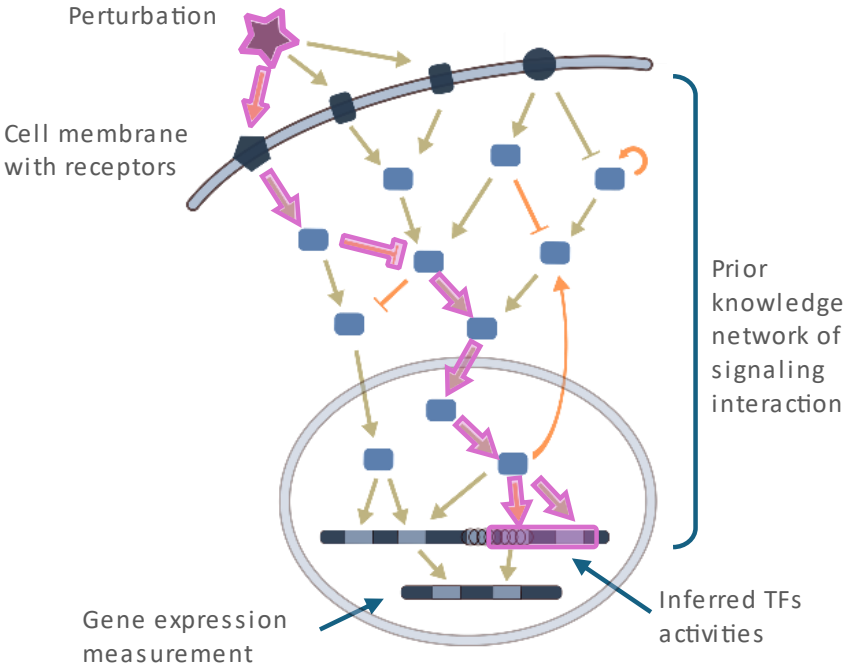


Figure 27 - CARNIVAL method scheme, adapted from (Rodríguez Mier & Türei, 2023).

The rules of signal transduction can be modeled as linear equality/inequality constraints. It is possible to find the formal ILP constraint in (Gjerga, 2020). Figure 28 shows the scheme of causal reasoning rules by which the signaling cascades are inferred through the PKN.

To evaluate a signaling network's ability to explain the predicted changes in TFs activity and the observed expression patterns given the initial perturbation, CARNIVAL adapted the objective function from Melas et al. (2015) as follows:

$$min\left(\sum_{V^M}|\alpha_j|\cdot|x_j - m_j|\right) + \left(\sum_V \beta\cdot(1-\gamma_j)\cdot x_j^+ + \sum_V \beta\cdot(1+\gamma_j)\cdot x_j^-\right)$$

The function is composed of two terms and balances two main criteria. Firstly, the network should provide a good fit for the data. For this purpose, the first term aims to incentivize the inclusion of as many of the user-provided measured nodes in the solution network as possible and ensure their activity state x_j matches the sign of the measurement score, m_j. If a measured node is not included in the network $(x_j = 0)$ a penalty is incurred, scaled by the magnitude of the user-provided measurement score

(α_j). The penalty is doubled if the activity state x_j of the measured node takes the opposite sign to that of the user-provided measurement (m_j).

Secondly, the network should be parsimonious and meet the first criterion while retaining a relatively small number of nodes. This criterion is based on the assumption that real biological networks evolve to be parsimonious (Leclerc, 2008), and therefore, the simplest network that explains the data is considered the best. The second term of the objective function incentivizes parsimony, increasing the value of the objective function when the number of nodes in the network rises. The γ_j weights allow the user to adjust the penalty depending on node activity. The balance between including more measured nodes and maintaining a smaller solution network can be tuned by adjusting the node penalty β. From here, we will refer to the first component of F as "error" and to the second component as "regularization term" (RT). The signaling network that meets a balance of these two criteria, minimizing the objective function and being within the constraints of the ILP problem, is found by a solver.

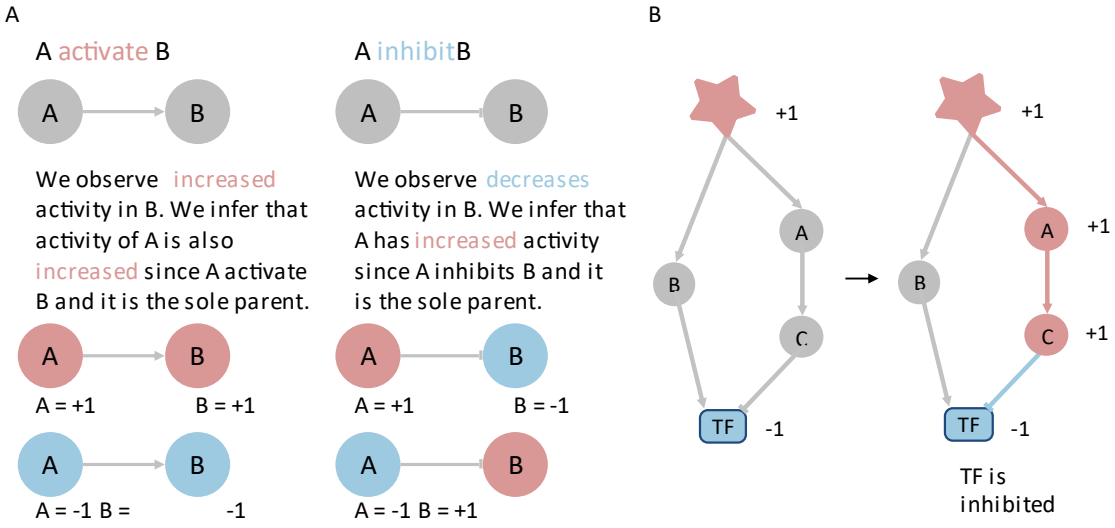


Figure 28 - The scheme of causal reasoning rules about changes in activity. (A) represents the causal inference of a protein that activates another protein, while (B) explains the causal reasoning of a protein that inhibits a second protein - figure from (Rodríguez Mier & Türei, 2023).

The optimal solution network inferred by CARNIVAL comprises a set of nodes $V' \subset V$, their corresponding activation states $x_j \in \{-1, 0, 1\} \forall v_j \in V$, and a set of directed edges $E' \subset E$ weighted by the interaction sign $\sigma_i \in \{-1, 0, 1\} \forall e_i \in E$.

The solver identifies an optimal solution, but as mentioned above, it is not always unique, and alternative solutions may exist. This study aims to emphasize the importance of exploring the space of alternative solutions. An *alternative solution* is defined as any solution other than the first one provided by the solver. Alternative solutions are classified as optimal and suboptimal. The former are all those solutions distinct from the first optimal solution that achieve the same level of optimality. At the same time, the

latter are valid solutions to the problem but with higher loss value, thus potentially exhibiting higher error or RT.

Implement *expl_alt_edge* and *expl_alt_node* Functions to Explore Alternative Solutions.

This study addressed the CARNIVAL problem by leveraging the principles outlined in DEXOM for context-specific network reconstruction. A set of constraints was introduced to identify alternative solutions, forcing the activation or deactivation of individual nodes or edges while monitoring the associated metrics for each viable solution.

In the initial version of our algorithm (Figure 29), iterative constraints were applied to all nodes or edges. The algorithm focused on either including or excluding individual nodes from the solution. If a node was present in the optimal solution—whether activated or inhibited—the algorithm sought alternative solutions that excluded that node. Conversely, if a node was absent from the original optimal solution, a constraint was added, forcing the optimizer to identify a solution that includes the node, regardless of its activation state. To expedite the search process, the algorithm was parallelized.

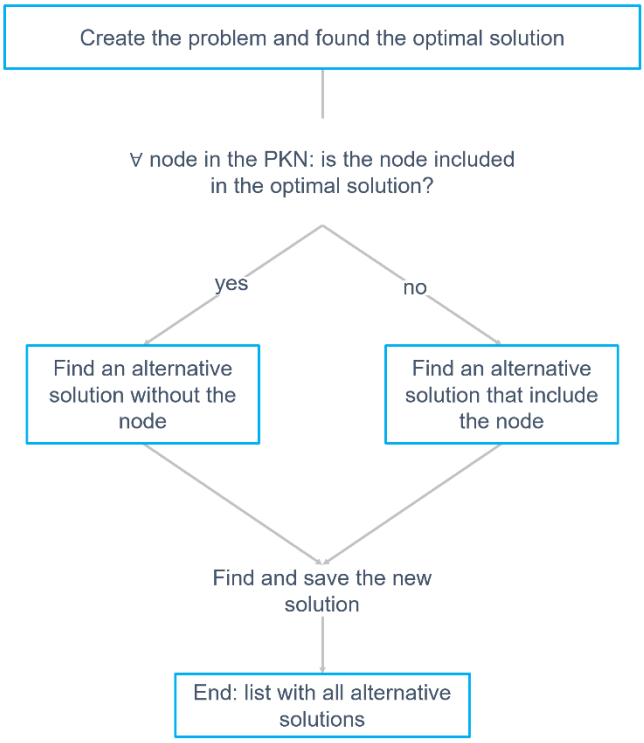


Figure 29 - Algorithm diagram, first version.

This initial version, however, proved to be computationally demanding, particularly when applied to large PNKs. Considering the computational challenges encountered with the initial version, a revised algorithm was developed that explores alternative solutions by adding constraints to one node at a time (Figure 30). In contrast to the previous

approach, which indiscriminately applied constraints to all nodes, the revised algorithm adds constraints only at nodes present in the initial optimal solution and those introduced in subsequent alternative solutions.

Algorithm Workflow

Optimal Solution Search: The algorithm initiates by identifying the optimal solution for the given problem.

Constraint List Creation: A list of constraints to be explored is generated, comprising all nodes from the optimal solution. Each node is listed twice, once with an opposite sign and once with an exclusion indicator.

Constraint list Update: If the newly identified solution incorporates nodes not initially selected in the optimal solution, the corresponding constraints for those nodes are appended to the constraint list for further testing. Notably, only the constraint with the opposite sign to that in the new solution is included, with the exclusion constraint being omitted as it led to the initial optimal solution.

Search Termination: Both algorithms, the node and edges versions, include a parameter that allows the search to conclude after a predetermined number of alternative solutions are identified.

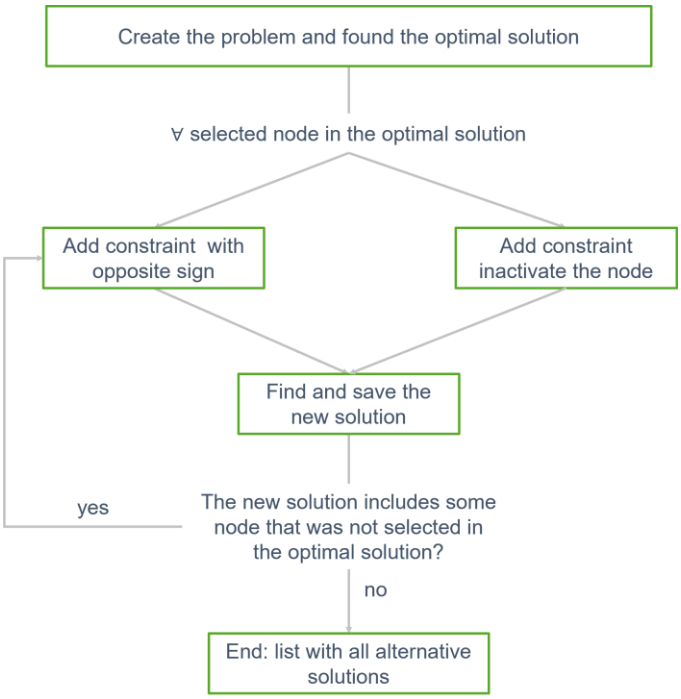


Figure 30 - Algorithm diagram, second version.

A toy example (Figure 31) was designed to provide a small experimental setting that would facilitate the development, testing, and validation of the functions. Additionally, the toy example would help present the functions. A PNK with 11 nodes and 15 edges

was constructed. Two nodes were designated as the perturbation input, and three were designated as the measurement nodes.

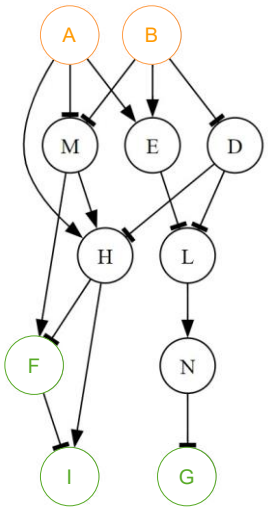


Figure 31 - Toy example. The plot represents the toy PKN with the perturbed nodes in orange and the measured nodes in green. Arrow-shaped edges represent activation, while hyphen-shaped edges represent inhibition of interactions between nodes.

Software and Tools

Both algorithm versions were implemented using Python programming language, leveraging various libraries to facilitate the development and execution of the algorithms. Below is a list of the specific packages employed in our study, along with their respective versions:

- Python: Version 3.11.9 (Van Rossum & Drake, 2009)
- Corneto: Version 0.9.1-alpha.5 (Rodríguez Mier et al., 2024)
- NumPy: Version 1.26.0 (Harris et al., 2020)
- Pandas: Version 2.1.1 (McKinney & others, 2010)
- Joblib: Version 1.3.2 (Joblib Development Team, 2020)
- SciPy: Version 1.11.3 (Virtanen et al., 2020)
- Scikit-learn: Version 1.3.2 (Pedregosa et al., 2011)
- Graphviz: Version 0.20.1 (Gansner & North, 1999)
- Gurobipy: 10.0.3 (Gurobi Optimization, 2023)
- Igraph: Version 0.10.8 (Csardi & Nepusz, 2006)
- Matplotlib: Version 3.8.0 (Hunter, 2007)
- NetworkX: Version 3.2.1 (Hagberg et al., 2008)
- Omnipath: Version 1.0.7 (Valdeolivas et al., 2016)
- Seaborn: Version 0.13.0 (Waskom, 2021)

These libraries were instrumental in implementing the algorithms, handling data structures, parallelizing computations, and performing various scientific computations

required for solving the CARNIVAL problem efficiently. These well-established libraries ensured our computational methods' reliability, scalability, and performance optimization.

Application of the Function to Real Data

Panacea Data

In this study, the primary application of real-world data was the PANACEA gene expression dataset. This dataset was retrieved through the NCBI GEO portal under the accession number GSE186341 (Douglass et al., 2022). The PANACEA database, developed by the Columbia CTD2 Center, offers a comprehensive collection of dose-response curves and molecular profiles, capturing diverse cellular contexts representative of neoplasms. This dataset encompasses samples from 32 distinct treatments across 11 cell lines, providing a valuable resource for studying cellular responses to various drug perturbations.

The present study focused on the PANC-1 cell line derived from a pancreatic cancer patient. The PANC-1 cell line was selected for its relevance to pancreatic cancer research and representation of the disease's molecular landscape. The cell lines were treated with Bafetinib, a kinase inhibitor primarily targeting ABL1 and LYN kinases (Kimura et al., 2005).

FLOP Pipeline

The FLOP (Functional Linkage of Omics Profiles) pipeline (Paton et al., 2024) was employed to analyze TFs activity within the Panacea dataset. FLOP is a comprehensive NetFlow-based workflow integrating normalization, differential expression (DE), and functional analysis methods for RNA-seq data.

The FLOP workflow includes the following steps:

- Downloading three gene set collections: PROGENy, DoRothEA, and MSigDB hallmarks.
- Executing six alternative DE analysis pipelines, such as tmm-limma, vsn-limma, voom-limma, log quantile-limma, edgeR, and DESeq2.
- Merging pipeline outputs for further analysis using the decoupleR module.

FLOP incorporates filtering strategies to eliminate lowly expressed genes and contrasts lacking sufficient signal. Four normalization methods -vsn, TMM, log quantile normalization, and voom- are employed, along with three DE analysis methods: limma, DESeq2, and edgeR.

In this study, the output of the filtered TMM-limma method was analyzed, and 34 TFs were taken as measurements, showing an adjusted p-value < 0.25.

Prior Knowledge Network

As a PKN, the SIGNOR (Lo Surdo et al., 2023) repository was selected for analysis and was accessed via Omnipath. SIGNOR provides a resource for annotating experimental evidence about causal interactions between proteins and other entities of biological relevance. These include stimuli, phenotypes, enzyme inhibitors, complexes, and protein families. Each entry in the database links to the experimental evidence that supports the interaction and is enriched with pertinent metadata. These data may include the effect of the interaction on the target entity's activity or the molecular mechanism underlying this effect. The SIGNOR interactions were obtained from the Omnipath database (Türei et al., 2021) using the corresponding Python client (version 1.0.7) with the *op.interactions.OmniPath.get* function. Initially, a filter was applied to restrict the network to only those connections with a curation effort above 6 to minimize the computational effort. Afterward, the complete network was analyzed. The network was pruned, and all edges and nodes above the perturbation nodes and below the measurement nodes were deleted.

Optimizer and Parameter

The Gurobi (Gurobi Optimization, 2023) optimizer was employed to identify the optimal solutions. The process of exploring alternative solutions is forced to conclude after the 200th attempt. The parameters were then set according to the following specifications:

- The value of the beta weight was set to 0.1.
- The MipGap value was set to 0.05.
- The time limit was set to 1000 seconds.

All alternative solutions identified are stored in a list.

Visualization

A scatter plot was generated to provide a visual representation of all the alternative solutions metrics. The plots display the absolute variation of the error and the percentage variation of the RT. The bar plots were generated on the marginal axes to illustrate the univariate distributions of error and RT. The plot was created using the *JointGrid* function of the seaborn package. An example of this representation is provided in Figure 32A.

The analysis concentrated on evaluating and analyzing the impact on the solution's optimality of knocking out a node (or edge). The study aims to quantify the impact of removing a node (or edge) on the solution's optimality by determining the number of additional nodes (or edges) that must be added to compensate for the knockout and the number of TFs that cannot be fitted.

A visualization was created to highlight the differences between the solutions' metrics when a given node (or edge) is absent. The set of alternative solutions was divided into

two groups based on the presence or absence of a given node (or edge). The median of each group is calculated, and the ratio is transformed using the *Log2* function. A scatter plot with a marginal distribution is used to visualize the results. Each dot represents a node. If the node dot is in the bottom-right quadrant, the solution without that node has an increase in solution size but a decrease in error. The same analysis is repeated for the edges. The two most interesting nodes and edges are highlighted. An example of this representation is provided in Figure 32B.

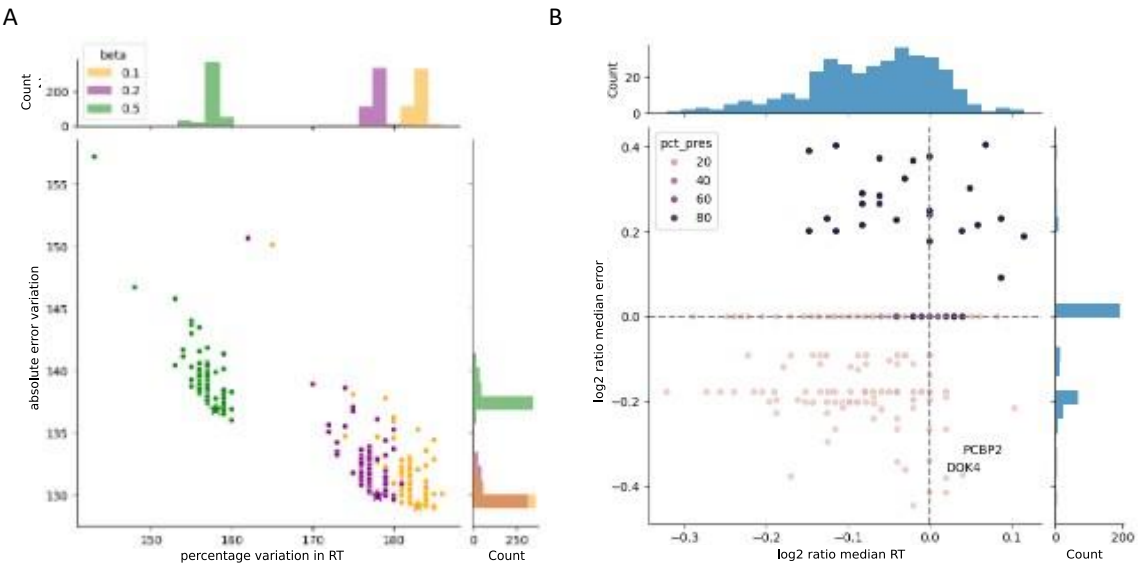


Figure 32 - Example of the two graph visualizations proposed. (A) Representation of alternative solution metrics. Each point represents an alternative solution, while the star represents the optimal solution. The different colors represent the results obtained in different experimental settings, changing the beta parameters. (B) Representation of sensitivity analysis results. Each point represents a node. The log2 ratio between the RT median of the solutions with and without the given node is plotted on the x-axis. The log2 ratio between the error median of the solutions with and without the given node is plotted on the y-axis. Each point is colored according to the percentage of its presence in the set of alternative solutions.

Results

Toy Example

We designed a toy example with a PNK with 11 nodes and 15 edges. We set two nodes as the perturbation input and three as the measurement. The optimal solution includes nine nodes, and the error is 0. The *expl_alt_node* function was tested, and nine alternative solutions were identified. Figure 33 illustrates the graph of the optimal and the alternative solutions.

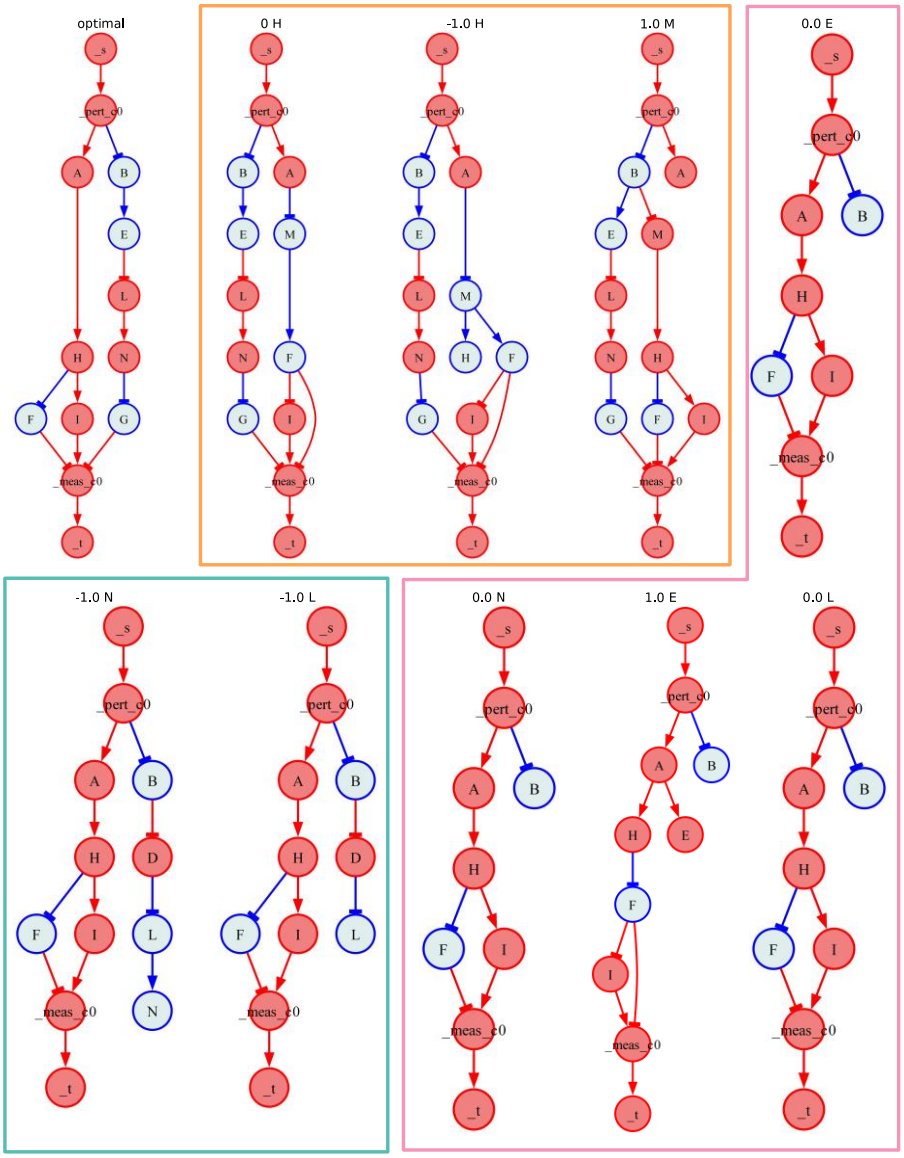


Figure 33 - Alternative solution found with expl_alt_node. The top-left plot represents the optimal solution; the others are all the alternative solutions found with expl_alt_node. The alternative solutions are grouped into 3 clusters.

These solutions were transformed into 1/0 matrices, where 1 represents the presence of the node, and 0 represents the absence. The solutions were clustered using the *linkage* function of the *Scipy* package with the complete method and Jaccard metric. The dendrogram was plotted (Figure 34A), and the threshold of 0.2 was identified as the optimal value to cut the tree and thus to identify three clusters.

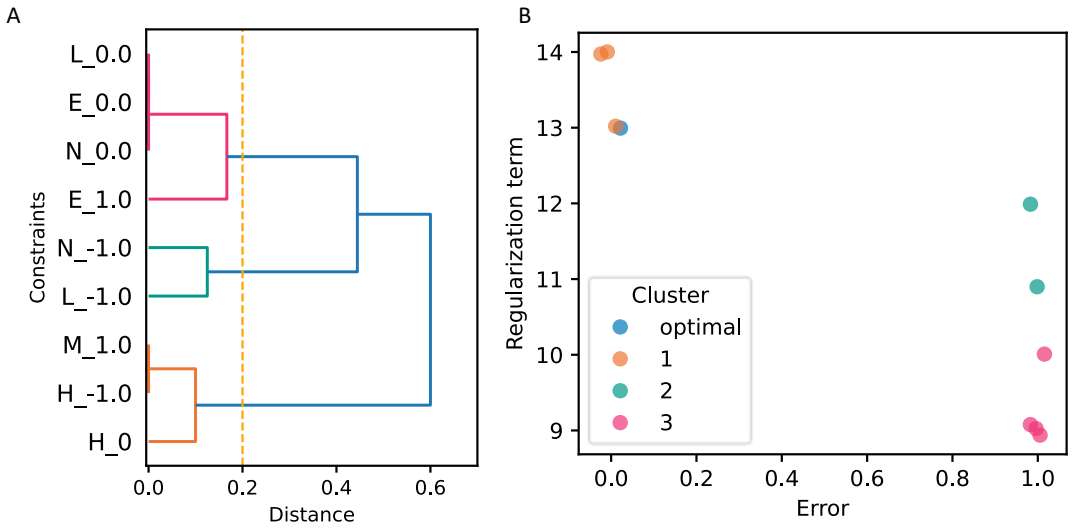


Figure 34 - Dendrogram (A) and scatter plot (B) representation of clustering alternative solutions obtained with expl_alt_node based on nodes presence/absence.

Cluster 1 (orange) Includes all solutions with the same error (0) but a higher number of nodes than the optimal solution. Cluster 2 (green) and Cluster 3 (pink) include solutions with a higher error but fewer nodes than Cluster 1. In cluster 2, we find solutions with a higher number of nodes than the optimal solution. In contrast, cluster 3 contains solutions with a similar number of nodes to the optimal solution but which fail to include all the measurements and perturbations. Figure 34B depicts the scatter plot of the solutions, with each solution color-coded according to its respective cluster. The blue data point represents the optimal solution, while the orange, green, and pink data points correspond to clusters 1, 2, and 3, respectively. Additionally, the edge function was tested on the toy example, identifying 15 alternative solutions, shown in Figure 35.

The same approach was applied to the analysis of the *expl_alt_node* results, whereby the alternative solutions were clustered. This finding was consistent with the previous results. The dendrogram Figure 36A was truncated at a level of 0.23, resulting in the identification of three clusters. Cluster 1 comprised all solutions with an error of zero, except for one solution with an error of two, which exhibited more nodes. Clusters 2 and 3 contain solutions with errors equal to one; however, they have fewer nodes than the other clusters. As in the case of the node, Figure 36B shows the scatter plot of the solutions, with each solution color-coded according to its respective cluster.

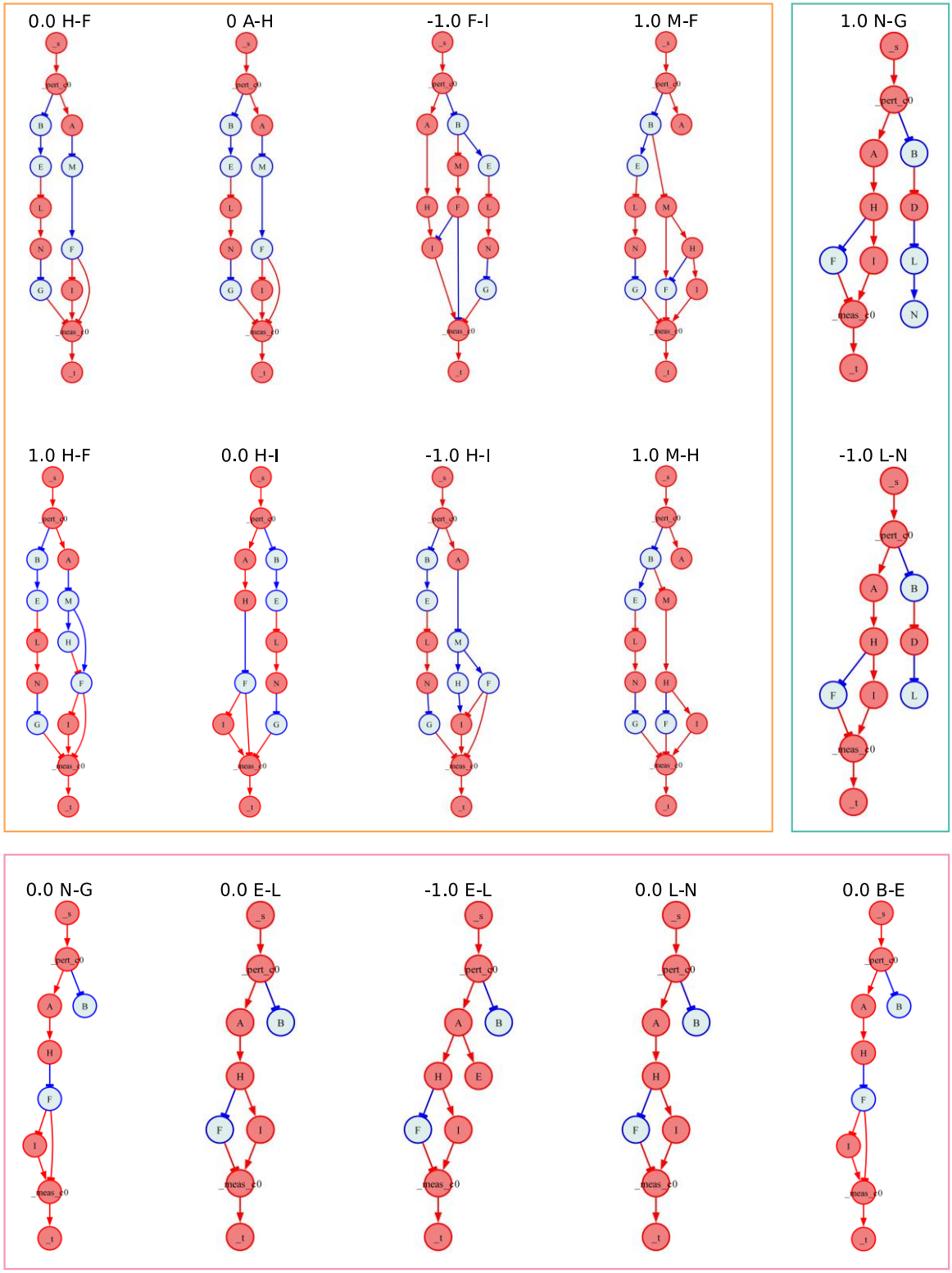


Figure 35 - The alternative solutions found with expl_alt_edge. The alternative solutions are grouped into 3 clusters.

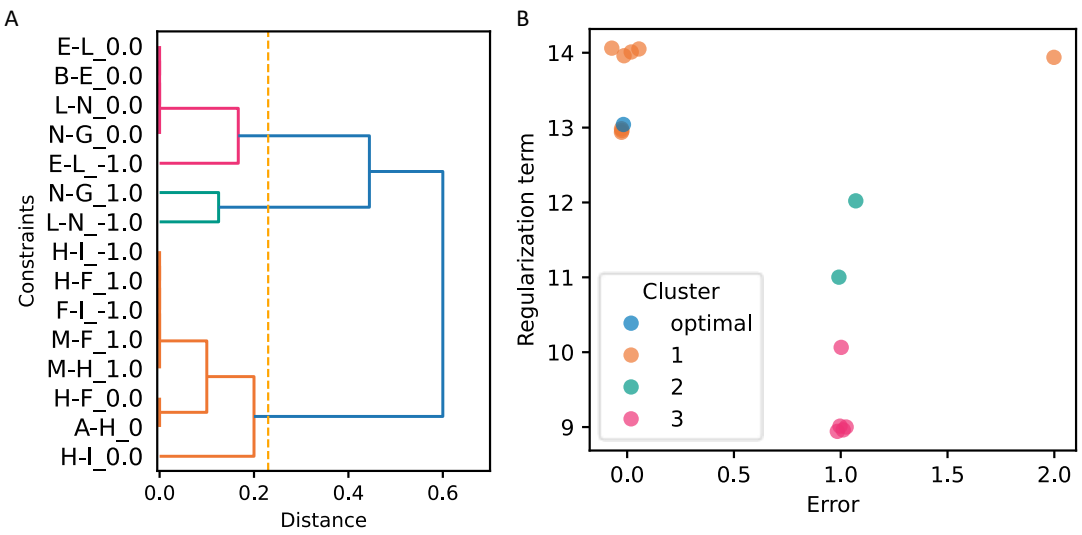


Figure 36 - Dendrogram (A) and scatter plot (B) representation of clustering alternative solutions obtained with expl_alt_edge based on nodes presence/absence.

Application to Panacea Data

The SIGNOR database was downloaded via Omnipath, which includes 60,857 edges and 5,522 nodes. Inhibitory-only interactions are assigned an edge weight of −1, while stimulus-only interactions are assigned an edge weight of 1. In the case of indications of both inhibitory and stimulus interactions, zero is assigned as the edge weight. In order to account for the effect of Bafetinib on two genes, namely ABL1 and LYN, which are known to be targets of this drug, two nodes were set as perturbations. As the two genes are kinase inhibitors, the signs for these nodes were set to -1. LogFC and relative adjusted p-values for TFs were calculated from the Panacea data analyses with the FLOP pipeline. TFs with an adjusted p-value > 0.25 were selected, and 34 TFs were set as measurements. Figure 37 shows the volcano plot, with the genes used as measurements highlighted. The TFs colored in blue are deactivated in bafetinib-treated cells compared to placebo-treated cells, while the red ones are activated.

The pruning process was applied to the PKN, eliminating all nodes and edges that fell above the perturbations and below the measurements. The final PKN consisted of 3853 edges and 1017 nodes. The optimal solution to the problem has an error of 15.25 and RT of 69.

Experimental Node Results

Firstly, the *expl_alt_node* function was applied, and the search was halted after 200 alternative solutions had been identified. Figure 38 shows the scatter plot where the absolute variation in the error of the alternative solution compared to the optimal solution is represented on the x-axis, while the percentage variation in RT is represented on the y-axis. The star represents the optimal solution. On the right margin, it can be

observed that two-thirds of the alternative solutions (153 out of 200) have an identical error to the optimal solution. In contrast, only a minor subset with solutions with a higher error than the optimal.

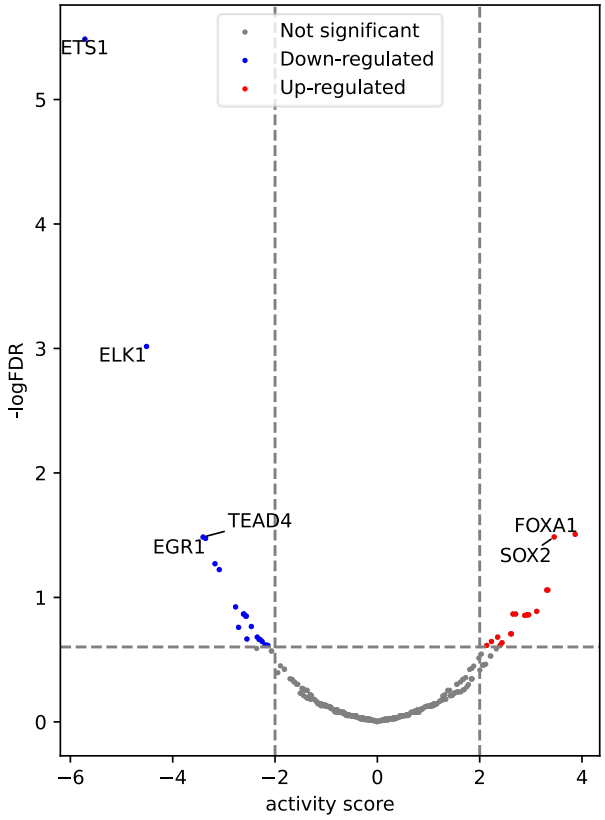


Figure 37 - Volcano plot of FLOP analysis. Red and blue points indicate upregulated and downregulated TFs selected as measurements.

These results suggest that the optimizer successfully identifies solutions that align with as many set measures as possible while only modifying the number of nodes. On the other hand, the bar plot of the upper marginal reveals the distribution of the percentage variation in the size of the solutions.

In contrast to the distribution observed for the error variance, a normal-shaped distribution with a mean of -0.28 and a median of -1.38 can be observed for the right margin. Among solutions with larger errors (49), it is possible to find solutions ranging from 6% more nodes to 16% fewer nodes, with a higher frequency of smaller solutions. The comparison between solutions with a smaller RT and those with a larger RT revealed 40 solutions with an RT below 0 versus four solutions with an RT above 0. Additionally, there were five solutions with the same RT as the optimal solution. When comparing the solutions with the same error as the optimal solution (151), it was found that approximately 45% had a smaller size than the optimal solution (67), while 17 solutions had the same size as the optimal solution, which can be considered equivalent optimal

solutions. Finally, the remaining 67 solutions had a larger size than the optimal solution. Notably, some solutions are identified that are more effective in terms of both error and size than the optimal solution. We also noticed that some of the identified solutions are better than the optimal solution in both terms of error and size because the parameter MipGap has been set to 0.05. This parameter allows the solver to sacrifice some accuracy in finding the optimal solution in favor of greater speed so that the solver finds and proposes an optimal solution with a small degree of uncertainty. This implies that there may be alternative, more optimal solutions, as we are observing. Since the primary purpose of this study is not to find the most optimal solution possible but rather to highlight and quantify the importance of studying alternative solutions, this lower precision is acceptable in this context. Moreover, the solutions identified and their respective experimental error are derived from prior knowledge, which is inherently biased and incomplete, allowing the limitations of the lower precision.

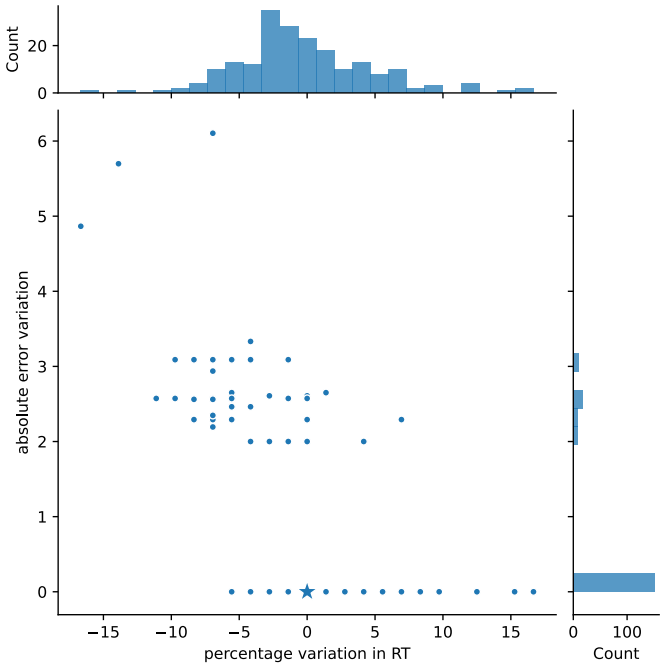


Figure 38 - Results of expl_alt_node on panacea data. Scatter plot of alternative solutions metrics.

Figure 39 illustrates the sensitivity analysis results conducted on nodes (A) and edges (B). Each point on the graphs represents the log ratio between the median of the distribution of solutions with a given node (or edge) and the median of the distribution of solutions without that node (or edge). The x-axis represents the log ratio between the medians of the RTs, while the y-axis shows the log ratio of the error medians. A point in the lower-right quadrant in Figure 39A represents a node (or an edge for Figure 39B) that, when absent from the solution, leads to an increase in median solution size but a decrease in median error. In this quadrant, one finds those nodes (or edges) that, although including them would result in sacrificing the parsimony criterion, are considered acceptable in the solution, as they lead to a significant decrease in error and a better fit to the data.

In Figure 39A, the nodes ERBB2 and EPHB2 are highlighted. Additionally, EPHB2 is present at the end of one of the two edges highlighted in Figure 39B, suggesting that these nodes may be crucial in explaining a portion of the biology of the problem that would otherwise remain unexplained if the solutions did not include them. It can be hypothesized that these nodes function as central hubs, representing well-characterized proteins with numerous interactions that contribute to explaining the activity of TFs within the context of the PKN. It is therefore possible to suppose that the removal of these nodes may result in a significant number of TFs remaining unexplained, which could indicate that these nodes are of critical importance. An alternative hypothesis is that these nodes have been the subject of extensive study, yet alternative compensatory signaling mechanisms may exist, but are not included in the PKN. An additional explanation is that the aforementioned nodes have been the subject of extensive study; however, other compensatory signaling mechanisms may exist but are not included in the PKN.

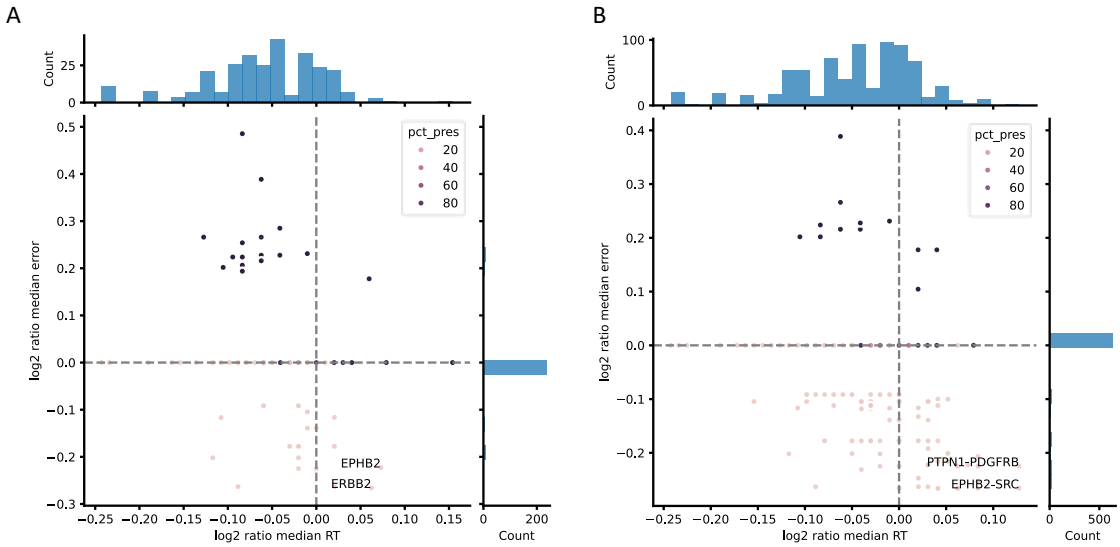


Figure 39 - Scatter plot of error and size variation when a node (A) or edge (B) is removed.

The difference between the optimal solution and one of the most different solutions was analyzed to highlight the importance of exploring and considering alternative solutions to find different biological insights. The Jaccard index between the nodes of the alternative solutions and the nodes of the optimal solution was calculated. The solution with the smallest index is found by adding constraints on node DLX5 = 1. The enrichment of the pathways from the WikiPathways gene set was evaluated, and Figure 40 shows the five most enriched pathways. It can be noticed that the two solutions share the enrichment of only one pathway, while the other four are different. The common pathway is B cell receptor signaling (WP23). The optimal solution is enriched in Brain-derived neurotrophic factor (BDNF) signaling (WP2380), IL11 signaling (WP2332), Thyroid stimulating hormone (TSH) signaling (WP2032), and IL5 signaling (WP127). The alternative solution is enriched in Pancreatic adenocarcinoma pathway (WP4263), VEGFA-VEGFR2 signaling (WP3888), Integrated cancer pathway (WP1971),

RAC1/PAK1/p38/MMP2 pathway (WP3303). Notably, optimal solution pathways are primarily associated with neurodevelopment, thyroid function, and immune responses, whereas alternative solution pathways are primarily associated with angiogenesis, cancer progression, and mechanisms associated with tumorigenesis. Optimal solution pathways tend to focus on specific molecular interactions and signaling cascades, such as neurotrophins (BDNF), cytokines (IL11 and IL5), and hormones (TSH). Whereas alternative solution pathways are more complex, involving mechanisms related to angiogenesis (VEGFA-VEGFR2 signaling), cell migration, and metastasis (RAC1/PAK1/p38/MMP2). Optimal solution pathways are primarily involved in regulating cell growth, differentiation, and survival, whereas alternative solution pathways are primarily involved in regulating cell proliferation and migration.

In summary, although both solutions describe signaling mechanisms, they differ significantly in their biological context, complexity, cellular specificity, and biological outcomes.

Figure 40 - Enrichment analysis results with WikiPathways gene set.

Experimental Edge Results

The same analyses were conducted on the alternative solutions identified with *expl_alt_edge*, setting the constraints on the edges. As illustrated in Figure 41, the scatter plot reveals a similar pattern observed in the alternative solutions of *expl_alt_node*. Most of the alternative solutions demonstrate the same error as the optimal solution. In fact, 132 alternative solutions have error variance from the optimal solution equal to zero. A total of 15 optimal alternative solutions were identified, resulting in the same error and size as the optimal solution. However, 112 has a larger dimension, while the last five solutions have smaller RT than the optimal solution. No solutions with smaller errors than the optimal solution were found.

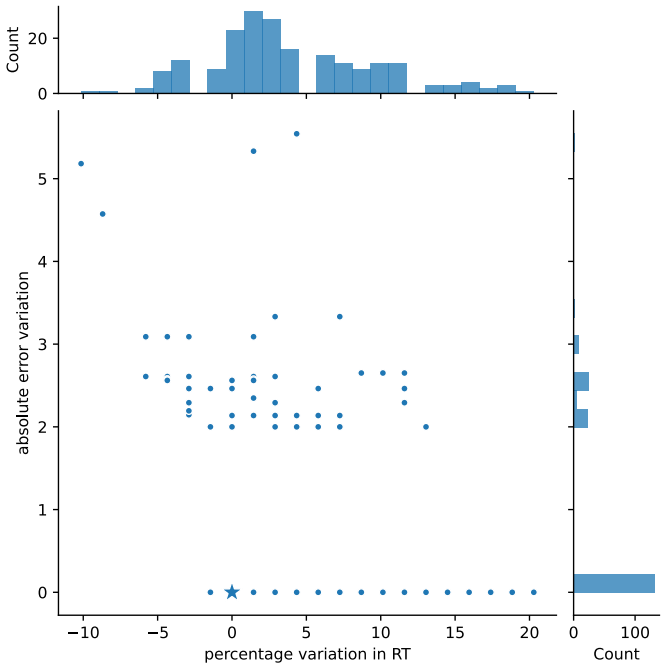


Figure 41 - Results of expl_alt_edge on panacea data. Scatter plot of alternative solutions metrics.

The upper bar plot indicates that most alternative solutions had a higher RT than the optimal solution. Specifically, 145 solutions reveal a greater than 0 variation in RT. Furthermore, the larger solution size also exhibits more variation. A maximum of 10% fewer nodes is observed in the smaller solutions, while the larger solutions reach up to 20% more nodes. In Figure 42, the plots highlight the nodes and edges that significantly influence the decrease in error. Observation indicates that the RXRB and RXRA nodes, when included in the solutions, result in a reduction of -0.43 of the log2 ratio median error. However, an increase in solution complexity accompanies this. Additionally, analysis reveals that the RXRB gene is situated at the extremes of the two most influential edges (RXRB-THRB and RXRB-THRA). Following the findings presented in Figure 41, since most solutions have an error equal to the optimal solution, the individual nodes or edges,

in most cases, also do not exhibit significant differences in error magnitude. Instead, the log2 ratio between medians of RTs has a more heterogeneous distribution.

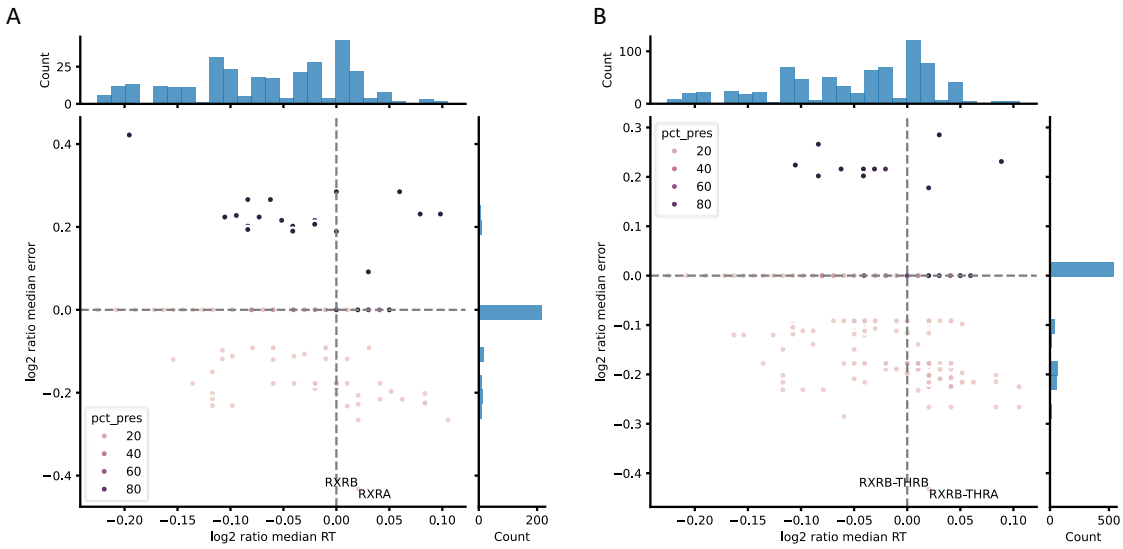


Figure 42 - Scatter plot of error and size variation when a node (A) or edge (B) is removed.

Discussion

In this study, we set up the building blocks to create a standardized workflow for exploring alternative solutions in contextualizing causal signaling networks from gene expression problems. Two functions for finding alternative solutions and a representation of their metrics were proposed. The functionality of the functions was tested on a toy example and then applied to a real context, using gene expression data from a pancreatic cancer cell perturbed with Bafetinib. Particular attention was then given to quantifying the importance of the individual nodes and edges included in the optimal solution. Furthermore, an example was presented to demonstrate the necessity of evaluating alternative solutions to provide a comprehensive biological interpretation.

The limitations of this study are twofold. First, the functions were tested on a single dataset of limited size, reducing the time required for the experiments. This is an ideal approach in a new function development set; however, it limited the generalization of the results. Second, while an approach to explore alternative solutions and different ways to visualize and interpret the results was successfully proposed, extracting their biological insight in a structured and rigorous manner remains an open question.

In the future, there is considerable potential to extend the application of these functions to more extensive networks, leveraging the computational power of a cluster. While this approach may introduce complexity, it would provide insights from a more comprehensive biological landscape, thus facilitating novel interpretations. Additionally,

it would be interesting to validate the methodology by applying it to new real-world datasets and different types of network inference problems, thereby exploring its efficacy across diverse contexts.

Another potential for further development could be the improvement and refinement of the enrichment analysis component. Creating different sets for enrichment analysis applications, such as clustering alternative solutions or investigating differences in nodes/edges between optimal and most different solutions, presents an intriguing prospect. Additionally, exploring the subset of nodes unique to each solution could offer valuable insights into the specific reactions they mediate.

4. Discussion, Conclusion, and Future Prospectives

In Chapter 1, we introduced the foundational concepts essential for comprehending the two projects discussed in the thesis. We describe the drug discovery process, focusing on the concepts, advantages, and strengths of DR. This approach allows the identification of new uses for existing drugs, which is crucial for rare diseases. We reviewed the principal applications and techniques in this field, presenting the current state of the art. We illustrated the fundamental principles of network analysis, primary data sources, and the employed techniques.The focus shifted to LP and optimization methods, to cocluded with ML and DL methodologies for drug repurposing ambitions.

In Chapter 2, we start presenting cystinosis, the rare disease. We detailed the causes, epidemiology, various forms, and existing cystinosis treatments, emphasizing the urgent need for alternative therapy beyond the currently available drug. Then we introduced a novel pipeline that integrates multi-omics, single-cell, and ontology data, combining network analysis and DL techniques, in order to understand the molecular mechanisms involved in cystinosis and support the development of new therapies. Our approach was focused on the renal tubule, as the kidney proximal tubules are primarily affected by cystinosis, leading to renal failure. Our previous findings indicated that cell cycle and proliferation pathways were upregulated in proximal tubule cells of the cystinosis mouse model. However, these were consequences rather than disease-specific mechanisms. We expanded an ontology-derived list of renal tubule genes using ML to identify more specific therapeutic targets. The predictions were mapped onto a molecular network with disease molecules. Our results highlighted the crucial role of mitochondrial oxidative phosphorylation processes, consistent with previous studies on the impact of mitochondrial oxidative stress in cystinosis kidney disease (Festa et al., 2018; Khan et al., 2020). We identified 51 drugs targeting 19 proteins potentially helpful in treating Cystinosis, like Coenzyme Q10 and N-acetylcysteine. The strengths of this study include the expansion of literature-derived phenotype genes and the development of a cell-specific network that integrates different data types. Currently, the workflow focuses only on repurposing existing compounds rather than identifying new drug targets. Future steps could involve wrapping the pipeline into a user-friendly set of functions to ensure the pipeline's reproducibility for other diseases. Additionally, it could be interesting to include patient-derived data instead of mouse model-derived data to enhance our understanding of the disease pathogenesis. Moreover, upgrading the pipeline with advanced techniques such as graph neural networks (GNN) or integrating network inference methodologies are potential improvements. GNNs are designed to process data structured as graphs, making them particularly effective for tasks involving complex relationships and interactions. By leveraging GNNs, it is possible to improve the model's ability to capture and analyze relation between disease molecule, phenotype genes and drug target, leading to more accurate drug target identification and repurposing opportunities. In addition to GNNs, integrating network inference methodologies

represents another potential improvement. These methodologies can identify upstream cellular processes that drive gene expression changes, allowing to obtain better insights into the underlying biological mechanisms. By applying network inference techniques, it's possible to enhance the robustness of our models and ensure biological insight throught network contextualization.

The need to study biological network inference led to the research presented in Chapter 3. This study introduces two functions for exploring alternative solutions in network signaling inference problems. We reviewed the state of the art, discussing the issues of multiple optimal solutions and non-identifiability. We also highlighted the significance of alternative solutions in providing manifold biological explanations. The methodology and results were illustrated, beginning with a toy example and after applying it to real data. Specifically, we utilized differentially expressed genes in a pancreatic cancer cell line treated with Bafetinib to infer the biological context by linking gene expression with drug-induced perturbation. We presented a graphical representation of the metrics' distributions of the solutions, with particular attention to those of the optimal solution. The sensitivity analysis was conducted on the optimal solution nodes and edges. Furthermore, the pathway enrichment analysis was performed on two distinct solutions, underlining the importance of considering alternative solutions. The strengths of this study include the implementation using CORNETO, a versatile framework applicable not only to signaling network inference problems but also to various network inference challenges. This versatility offers potential applications in different contexts, such as flux balance analysis or Bayesian network problems. Nevertheless, the lack of different applications represents a limitation, given that the method was tested on a single problem type and data source. Future steps should involve testing the functions in different contexts to validate their performance and expanding the enrichment analysis. This strategy will allow the extraction of precise, concise, and meaningful biological insights. The standardization of enrichment analysis has the potential to guide the development of a pipeline suited for extracting significant biological insights from a pool of alternative solutions.

From a future perspective, adapting Chapter 3's network inference methodology to Chapter 1's data would be interesting. This adaptation could involve modifying the ILP problem to use phenotype genes instead of perturbations and incorporating drug target information and mechanisms of action. This approach could potentially identify a signaling pathway that begins from PGs, passes through drug targets that reverse the signal, and ends at DM with opposite signs to those observed in disease models. This integration could enhance our knowledge of the underlying biological mechanisms and lead to the prioritization of novel therapeutic targets.

Concluding, computational biology is becoming increasingly integrated with diverse mathematical and computational methodologies, which reflects the growing importance of these approaches in the field. While initially, this field primarily focused on

deterministic models (such as network analysis), incorporating stochastic methodologies has significantly transformed the approaches employed to address various biological challenges. In particular, two lines of research have emerged and developed over the last years. Firstly, there has been a proliferation of interest in studying models that incorporate stochastic optimization, encompassing both traditional linear programming techniques and contemporary advancements in machine learning (ML) and deep learning (DL). Concurrently, the research strand of complex network theory has been the subject of extensive study, contributing to the development of this research area. The research areas mentioned above, along with their respective methodological frameworks, have not only developed independently but have also converged, resulting in innovative combinations of techniques that lead to increasingly sophisticated analytical tools.

This thesis aims to illustrate how these various research elements can be applied both as standalone techniques and as part of an integrated, extended pipeline, providing solutions to diversified biological challenges. The overall intention is to highlight the growing importance of computational methodologies, becoming indispensable components of a comprehensive experimental pipeline. In particular, the computational phase is increasingly intersecting with the biological workflows, particularly in the context of drug development. It has become essential for drug discovery to move away from viewing the laboratory and data analysis phases as separate and distinct processes. Instead, it is crucial to recognize that they are interdependent and complementary. The computational moment is no longer a mere data analysis, a standalone and independent process; it is an integral part of the pipeline, where different tools correspond to each different step. Consequently, the rationale behind this thesis is to highlight how a complete pipeline for addressing any clinical or biological research question can no longer ignore either of the two moments and must consider both simultaneously. These stages are not merely sequential; rather, they are interdependent and can even occur concurrently.

In particular, the two methodologies presented in this thesis, in combination with the wet lab techniques, can delineate a blueprint for a complete drug development pipeline, as described above. The first methodology, which introduces a DR framework, can be used to facilitate the identification of compounds or drugs for experimental testing. Once preliminary wet-lab results have been obtained, these data can be analyzed using the ILP-based methodology described in Chapter 3. This analysis will be able to contextualize the biological mechanisms inferred from the experimental results using the knowledge gained from the DR pipeline. In an ideal scenario, the exchange of information between the two computational pipelines and the wet lab are continuous, thus leading to obtain the best answer to the biological research question.

Bibliography

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digital Medicine*, *4*(1), 65. https://doi.org/10.1038/s41746-021-00438-z

Agrawal, A., Balcı, H., Hanspers, K., Coort, S. L., Martens, M., Slenter, D. N., Ehrhart, F., Digles, D., Waagmeester, A., Wassink, I., Abbassi-Daloii, T., Lopes, E. N., Iyer, A., Acosta, J. M., Willighagen, L. G., Nishida, K., Riutta, A., Basaric, H., Evelo, C. T., … Pico, A. R. (2024). WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, *52*(D1), D679–D689. https://doi.org/10.1093/NAR/GKAD960

Aittokallio, T. (2022). What are the current challenges for machine learning in drug discovery and repurposing? *Expert Opinion on Drug Discovery*, *17*(5), 423–425. https://doi.org/10.1080/17460441.2022.2050694

Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414. https://doi.org/10.1093/NAR/GKW985

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology 2015 33:8*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Alsentzer, E., Li, M. M., Kobren, S. N., Network, U. D., Kohane, I. S., & Zitnik, M. (2022). Deep learning for diagnosing patients with rare genetic diseases. *MedRxiv*, 2022.12.07.22283238. https://doi.org/10.1101/2022.12.07.22283238

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), 1–12. https://doi.org/10.1186/GB-2010-11-10-R106/COMMENTS

Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery 2004 3:8*, *3*(8), 673–683. https://doi.org/10.1038/nrd1468

Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., & Jafari, M. (2018). A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*, *12*(1), 1–17. https://doi.org/10.1186/S12918-018-0598-2/FIGURES/8

Atz, K., Cotos, L., Isert, C., Håkansson, M., Focht, D., Hilleke, M., Nippa, D. F., Iff, M., Ledergerber, J., Schiebroek, C. C. G., Romeo, V., Hiss, J. A., Merk, D., Schneider, P., Kuhn, B., Grether, U., & Schneider, G. (2024). Prospective de novo drug design with

deep interactome learning. *Nature Communications 2024 15:1*, *15*(1), 1–18. https://doi.org/10.1038/s41467-024-47613-w

Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., & Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods 2017 14:6*, *14*(6), 584–586. https://doi.org/10.1038/nmeth.4263

Badam, T. V. S., de Weerd, H. A., Martínez-Enguita, D., Olsson, T., Alfredsson, L., Kockum, I., Jagodic, M., Lubovac-Pilav, Z., & Gustafsson, M. (2021). A validated generally applicable approach using the systematic assessment of disease modules by GWAS reveals a multi-omic module strongly associated with risk factors in multiple sclerosis. *BMC Genomics*, *22*(1), 1–13. https://doi.org/10.1186/S12864-021-07935-1/FIGURES/5

Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, *2*(1). https://doi.org/10.1093/bioadv/vbac016

Badia-i-Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., & Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, *24*(11), 739–754. https://doi.org/10.1038/s41576-023-00618-5

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2010). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics 2011 12:1*, *12*(1), 56–68. https://doi.org/10.1038/nrg2918

Barabási, A.-L. (2016). *Network Science*.

Barata, J. T., & Oliveira, M. L. (2019). Cell Signaling in Cancer. In *Molecular and Cell Biology of Cancer* (pp. 31–49). https://doi.org/10.1007/978-3-030-11812-9_3

Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., & Snyder-Cappione, J. E. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications 2019 10:1*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-13055-y

Berquez, M., Chen, Z., Festa, B. P., Krohn, P., Keller, S. A., Parolo, S., Korzinkin, M., Gaponova, A., Laczko, E., Domenici, E., Devuyst, O., & Luciani, A. (2023). Lysosomal cystine export regulates mTORC1 signaling to guide kidney epithelial cell fate specialization. *Nature Communications 2023 14:1*, *14*(1), 1–21. https://doi.org/10.1038/s41467-023-39261-3

Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., & Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods 2023 20:9*, *20*(9), 1355–1367. https://doi.org/10.1038/s41592-023-01938-4

Brewer, G. J. (2009). Drug development for orphan diseases in the context of personalized medicine. *Translational Research*, *154*(6), 314–322. https://doi.org/10.1016/J.TRSL.2009.03.008

Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *International Journal of Molecular Sciences*, *10*(6), 2763. https://doi.org/10.3390/IJMS10062763

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., Prlić, A., Sastry, A., Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M. T., Nielsen, J., … Palsson, B. O. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272–281. https://doi.org/10.1038/nbt.4072

Buphamalai, P., Kokotovic, T., Nagy, V., & Menche, J. (2021). Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature Communications 2021 12:1*, *12*(1), 1–15. https://doi.org/10.1038/s41467-021-26674-1

Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *IScience*, *25*(2), 103798. https://doi.org/10.1016/j.isci.2022.103798

Carlson, M. (2019). *org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.*

Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H., & Yuan, S. (2019). Advancing Drug Discovery via Artificial Intelligence. *Trends in Pharmacological Sciences*, *40*(8), 592–604. https://doi.org/10.1016/j.tips.2019.06.004

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R* (R package version 1.7.4). https://shiny.posit.co/

Chen, H., King, F. J., Zhou, B., Wang, Y., Canedy, C. J., Hayashi, J., Zhong, Y., Chang, M. W., Pache, L., Wong, J. L., Jia, Y., Joslin, J., Jiang, T., Benner, C., Chanda, S. K., & Zhou, Y. (2024). Drug target prediction through deep learning functional representation of gene signatures. *Nature Communications 2024 15:1*, *15*(1), 1–15. https://doi.org/10.1038/s41467-024-46089-y

Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A. L., & Loscalzo, J. (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications 2018 9:1*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-05116-5

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., Natoli, T., Narayan, R., Aicheler, F., Amoroso, N., Arenas, A., Azhagesan, K., Baker, A., Banf, M., Batzoglou, S., … Marbach, D. (2019). Assessment of network module identification across complex diseases. *Nature Methods 2019 16:9*, *16*(9), 843–852. https://doi.org/10.1038/s41592-019-0509-5

Chowdhury, H. A., Bhattacharyya, D. K., & Kalita, J. K. (2019). (Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/TCBB.2019.2893170

Cong, Y., & Endo, T. (2022). Multi-Omics and Artificial Intelligence-Guided Drug Repositioning: Prospects, Challenges, and Lessons Learned from COVID-19. *OMICS: A Journal of Integrative Biology*, *26*(7), 361–371. https://doi.org/10.1089/omi.2022.0068

Coté, T. R., Xu, K., & Pariser, A. R. (2010). Accelerating orphan drug development. *Nature Reviews Drug Discovery 2010 9:12*, *9*(12), 901–902. https://doi.org/10.1038/nrd3340

Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics 2017 18:9*, *18*(9), 551–562. https://doi.org/10.1038/nrg.2017.38

Csabai, L., Fazekas, D., Kadlecsik, T., Szalay-Bek, M., azs BohárBoh, B., Madgwick, M., Gul, L., Sudhakar, P., Kubisch, J., James Oyeyemi, O., Liska, O., Ari, E., Hotzi, B., Billes, V. A., MolnárMoln, E., CsályiCs, K., Demeter, A., Koltai, M., Varga, até, … KorcsmárosKorcsm, T. (2021). SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Research*, *50*. https://doi.org/10.1093/nar/gkab909

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. https://igraph.org

Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., & Noom, D. (2023). igraph for R: R interface of the igraph library for graph theory and network analysis. *Zenodo. Available Online at: Https://CRAN. R-Project. Org/Package= Igraph (Accessed March 30, 2023)*.

De Domenico, M. (2023). More is different in real-world multilayer networks. *Nature Physics*, *19*(9), 1247–1262. https://doi.org/10.1038/s41567-023-02132-1

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Dimitrakopoulos, G. N., Klapa, M. I., & Moschonas, N. K. (2022). How Far Are We from the Completion of the Human Protein Interactome Reconstruction? *Biomolecules*, *12*(1), 140. https://doi.org/10.3390/biom12010140

Dolgalev, I. (2022). *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format* (7.5.1.9001).

Douglass, E. F., Allaway, R. J., Szalai, B., Wang, W., Tian, T., Fernández-Torras, A., Realubit, R., Karan, C., Zheng, S., Pessia, A., Tanoli, Z., Jafari, M., Wan, F., Li, S., Xiong, Y., Duran-Frigola, M., Bertoni, M., Badia-i-Mompel, P., Mateo, L., … Califano, A. (2022). A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. *Cell Reports Medicine*, *3*(1), 100492. https://doi.org/10.1016/j.xcrm.2021.100492

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 2009 4:8*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Ehrhart, F., Willighagen, E. L., Kutmon, M., van Hoften, M., Curfs, L. M. G., & Evelo, C. T. (2021). A resource to explore the discovery of rare diseases and their causative genes. *Scientific Data 2021 8:1*, *8*(1), 1–8. https://doi.org/10.1038/s41597-021-00905-y

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., & Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials*, *18*(5), 435–441. https://doi.org/10.1038/s41563-019-0338-z

Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. *Drug Discovery Today*, *26*(3), 769–777. https://doi.org/10.1016/j.drudis.2020.12.003

Elmonem, M. A., Veys, K. R., Soliman, N. A., Van Dyck, M., Van Den Heuvel, L. P., & Levtchenko, E. (2016). Cystinosis: a review. *Orphanet Journal of Rare Diseases 2016 11:1*, *11*(1), 1–17. https://doi.org/10.1186/S13023-016-0426-Y

Emma, F., Nesterova, G., Langman, C., Labbé, A., Cherqui, S., Goodyer, P., Janssen, M. C., Greco, M., Topaloglu, R., Elenberg, E., Dohil, R., Trauner, D., Antignac, C., Cochat, P., Kaskel, F., Servais, A., Wühl, E., Niaudet, P., Van't Hoff, W., … Levtchenko, E. (2014). Nephropathic cystinosis: an international consensus document. *Nephrology Dialysis Transplantation*, *29*(suppl_4), iv87–iv94. https://doi.org/10.1093/NDT/GFU090

Erb, I., & Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, *135*(1–2), 21–36. https://doi.org/10.1007/S12064-015-0220-8/FIGURES/9

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655. https://doi.org/10.1093/NAR/GKX1132

Fang, J., Zhang, P., Zhou, Y., Chiang, C. W., Tan, J., Hou, Y., Stauffer, S., Li, L., Pieper, A. A., Cummings, J., & Cheng, F. (2021). Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease. *Nature Aging 2021 1:12*, *1*(12), 1175–1188. https://doi.org/10.1038/s43587-021-00138-z

Ferrero, E., Dunham, I., & Sanseau, P. (2017). In silico prediction of novel therapeutic targets using gene-disease association data. *Journal of Translational Medicine*, *15*(1), 1–16. https://doi.org/10.1186/S12967-017-1285-6/FIGURES/6

Festa, B. P., Chen, Z., Berquez, M., Debaix, H., Tokonami, N., Prange, J. A., Hoek, G. Van De, Alessio, C., Raimondi, A., Nevo, N., Giles, R. H., Devuyst, O., & Luciani, A. (2018). Impaired autophagy bridges lysosomal storage disease and epithelial dysfunction in the kidney. *Nature Communications 2018 9:1*, *9*(1), 1–17. https://doi.org/10.1038/s41467-017-02536-7

Fukushima. (1969). Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements. *IEEE Transactions on Systems Science and Cybernetics*, *5*(4), 322–333. https://doi.org/10.1109/TSSC.1969.300225

Gahl, W. A., Thoene, J. G., & Schneider, J. A. (2002). Cystinosis. *New England Journal of Medicine*, *347*(2). https://doi.org/10.1056/NEJMRA020552

Gansner, E. R., & North, S. C. (1999). An open graph visualization system and its applications to software engineering. *Pract. Exper*, *S1*, 1–5.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375. https://doi.org/10.1101/GR.240663.118

Gargano, M. A., Matentzoglu, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, A. V., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G., Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., … Robinson, P. N. (2024). The Human Phenotype Ontology in 2024:

phenotypes around the world. *Nucleic Acids Research*, *52*(D1), D1333–D1346. https://doi.org/10.1093/NAR/GKAD1005

Gjerga, E. (2020). *Modelling and analysis of large-scale models of signalling networks.* [RWTH Aachen University]. https://doi.org/https://doi.org/10.18154/RWTH-2021-00617

Gjerga, E., Dugourd, A., Tobalina, L., Sousa, A., & Saez-Rodriguez, J. (2021). PHONEMeS: Efficient Modeling of Signaling Networks Derived from Large-Scale Mass Spectrometry Data. *Journal of Proteome Research*, *20*(4), 2138–2144. https://doi.org/10.1021/acs.jproteome.0c00958

Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., & Zhang, X. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, *33*(13), 2010–2019. https://doi.org/10.1093/BIOINFORMATICS/BTX069

Greene, D., Richardson, S., & Turro, E. (2017). ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*, *33*(7), 1104–1106. https://doi.org/10.1093/BIOINFORMATICS/BTW763

Gurobi Optimization, L. (2023). *Gurobi Optimizer Reference Manual*. https://www.gurobi.com

Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., & Saez-Rodriguez, J. (2013). Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. *Bioinformatics*, *29*(18), 2320–2326. https://doi.org/10.1093/bioinformatics/btt393

Hagberg, A., Swart, P., & Chult, D. (2008, July). Exploring Network Structure, Dynamics, and Function Using NetworkX. *Proceedings of the 7th Python in Science Conference*.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/10.1016/J.CELL.2021.04.048

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hernández-Camacho, J. D., Bernier, M., López-Lluch, G., & Navas, P. (2018). Coenzyme Q10 supplementation in aging and disease. *Frontiers in Physiology*, *9*(FEB), 316577. https://doi.org/10.3389/FPHYS.2018.00044/BIBTEX

Hossain, S. M. M., Halsana, A. A., Khatun, L., Ray, S., & Mukhopadhyay, A. (2021). Discovering key transcriptomic regulators in pancreatic ductal adenocarcinoma using Dirichlet process Gaussian mixture model. *Scientific Reports*, *11*(1), 7853. https://doi.org/10.1038/s41598-021-87234-7

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science*, *292*(5518), 929–934. https://doi.org/10.1126/science.292.5518.929

Jackson, J. D., Smith, F. G., Litman, N. N., Yuile, C. L., & Latta, H. (1962). The Fanconi syndrome with cystinosis. *The American Journal of Medicine*, *33*(6), 893–910. https://doi.org/10.1016/0002-9343(62)90221-8

Jamalpoor, A., Eerde, A. M. van, Lilien, M. R., Gelder, C. A. van, Zaal, E. A., Valentijn, F. A., Broekhuizen, R., Zielhuis, E., Egido, J. E., Altelaar, M., Berkers, C. R., Masereeuw, R., & Janssen, M. J. (2024). The lysosomal V-ATPase B1 subunit in renal proximal tubule is linked to nephropathic cystinosis. *BioRxiv*, 2020.07.24.219808. https://doi.org/10.1101/2020.07.24.219808

Jamalpoor, A., van Gelder, C. A., Yousef Yengej, F. A., Zaal, E. A., Berlingerio, S. P., Veys, K. R., Pou Casellas, C., Voskuil, K., Essa, K., Ammerlaan, C. M., Rega, L. R., van der Welle, R. E., Lilien, M. R., Rookmaaker, M. B., Clevers, H., Klumperman, J., Levtchenko, E., Berkers, C. R., Verhaar, M. C., … Janssen, M. J. (2021). Cysteamine–bicalutamide combination therapy corrects proximal tubule phenotype in cystinosis. *EMBO Molecular Medicine*, *13*(7). https://doi.org/10.15252/EMMM.202013067/SUPPL_FILE/EMMM202013067-SUP-0004-DATASETEV2.XLSX

Janyasupab, P., Suratanee, A., & Plaimas, K. (2021). Network diffusion with centrality measures to identify disease-related genes. *Mathematical Biosciences and Engineering : MBE*, *18 3*(3), 2909–2929. https://doi.org/10.3934/MBE.2021147

Jensen, K., Gudmundsson, S., & Herrgård, M. J. (2018). *Enhancing Metabolic Models with Genome-Scale Experimental Data* (pp. 337–350). https://doi.org/10.1007/978-3-319-92967-5_17

Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., Moffat, J., & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, *6*(7), 57. https://doi.org/10.1186/s13073-014-0057-7

Jeon, M., Jagodnik, K. M., Kropiwnicki, E., Stein, D. J., & Ma'ayan, A. (2021). Prioritizing pain-associated targets with machine learning. *Biochemistry*, *60*(18), 1430–1446. https://doi.org/10.1021/ACS.BIOCHEM.0C00930/SUPPL_FILE/BI0C00930_SI_001.PDF

Jha, M., Roy, S., & Kalita, J. K. (2020). Prioritizing disease biomarkers using functional module based network analysis: A multilayer consensus driven scheme. *Computers in Biology and Medicine*, *126*. https://doi.org/10.1016/J.COMPBIOMED.2020.104023

Ji, Z., Wang, B., Yan, K., Dong, L., Meng, G., & Shi, L. (2017). A linear programming computational framework integrates phosphor-proteomics and prior knowledge to predict drug efficacy. *BMC Systems Biology*, *11*(S7), 127. https://doi.org/10.1186/s12918-017-0501-6

Jiang, P., & Singh, M. (2010). SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, *26*(8), 1105–1111. https://doi.org/10.1093/bioinformatics/btq078

Jin, Z., Sato, Y., Kawashima, M., & Kanehisa, M. (2023). KEGG tools for classification and analysis of viral proteins. *Protein Science*, *32*(12), e4820. https://doi.org/10.1002/PRO.4820

Joblib Development Team. (2020). *Joblib: running Python functions as pipeline jobs*. https://joblib.readthedocs.io/

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, *614*(7949), 742–751. https://doi.org/10.1038/s41586-022-05688-9

Kamya, P., Ozerov, I. V., Pun, F. W., Tretina, K., Fokina, T., Chen, S., Naumov, V., Long, X., Lin, S., Korzinkin, M., Polykovskiy, D., Aliper, A., Ren, F., & Zhavoronkov, A. (2023). PandaOmics: An AI-Driven Platform for Therapeutic Target and Biomarker Discovery. *Journal of Chemical Information and Modeling*, *64*, 3961–3969. https://doi.org/10.1021/ACS.JCIM.3C01619/ASSET/IMAGES/LARGE/CI3C01619_0002.JPEG

Kandasamy, K., Sujatha Mohan, S., Raju, R., Keerthikumar, S., Sameer Kumar, G. S., Venugopal, A. K., Telikicherla, D., Navarro, D. J., Mathivanan, S., Pecquet, C., Gollapudi, S. K., Tattikota, S. G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H. K. C., Zhong, J., Sekhar, R., … Pandey, A. (2010). NetPath: A public resource of curated signal transduction pathways. *Genome Biology*, *11*(1). https://doi.org/10.1186/GB-2010-11-1-R3

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. *Academic.Oup.ComM Kanehisa, M*

Furumichi, Y Sato, M Ishiguro-Watanabe, M TanabeNucleic Acids Research, 2021•academic.Oup.Com, *49*. https://doi.org/10.1093/nar/gkaa970

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., … Pandey, A. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Research*, *37*(Database issue). https://doi.org/10.1093/NAR/GKN892

Khan, M. A., Wang, X., Giuliani, K. T. K., Nag, P., Grivei, A., Ungerer, J., Hoy, W., Healy, H., Gobe, G., & Kassianos, A. J. (2020). Underlying Histopathology Determines Response to Oxidative Stress in Cultured Human Primary Proximal Tubular Epithelial Cells. *International Journal of Molecular Sciences 2020, Vol. 21, Page 560*, *21*(2), 560. https://doi.org/10.3390/IJMS21020560

Kim, D., Tran, A., Kim, H. J., Lin, Y., Yang, J. Y. H., & Yang, P. (2023). Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *Npj Systems Biology and Applications*, *9*(1), 51. https://doi.org/10.1038/s41540-023-00312-6

Kimura, S., Naito, H., Segawa, H., Kuroda, J., Yuasa, T., Sato, K., Yokota, A., Kamitsuji, Y., Kawata, E., Ashihara, E., Nakaya, Y., Naruoka, H., Wakayama, T., Nasu, K., Asaki, T., Niwa, T., Hirabayashi, K., & Maekawa, T. (2005). NS-187, a potent and selective dual Bcr-Abl/Lyn tyrosine kinase inhibitor, is a novel agent for imatinib-resistant leukemia. *Blood*, *106*(12), 3948–3954. https://doi.org/10.1182/blood-2005-06-2209

Kingma, D. P., & Ba, J. L. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1412.6980v9

Klamt, S., & von Kamp, A. (2009). Computing paths and cycles in biological interaction graphs. *BMC Bioinformatics*, *10*(1), 181. https://doi.org/10.1186/1471-2105-10-181

Kochen, M. A., Andrews, S. S., Wiley, H. S., Feng, S., & Sauro, H. M. (2022). Dynamics and Sensitivity of Signaling Pathways. *Current Pathobiology Reports*, *10*(2), 11–22. https://doi.org/10.1007/s40139-022-00230-y

Koschützki, D., & Schreiber, F. (2008). Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regulation and Systems Biology*, *2*(2), 193. https://doi.org/10.4137/GRSB.S702

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1–13. https://doi.org/10.1186/1471-2105-9-559/FIGURES/4

Leclerc, R. D. (2008). Survival of the sparsest: Robust gene networks are parsimonious. *Molecular Systems Biology*, *4*. https://doi.org/10.1038/msb.2008.52

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, *17*(1), 2–12. https://doi.org/10.1093/BIB/BBV020

Li, M. M., Huang, K., & Zitnik, M. (2022). Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering 2022 6:12*, *6*(12), 1353–1369. https://doi.org/10.1038/s41551-022-00942-x

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/BIOINFORMATICS/BTR260

Ligtenberg, W. (2019). *reactome.db: A set of annotation maps for reactome* (R package version 1.68.0).

Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen - van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, *6*(1), 26286. https://doi.org/10.1038/srep26286

Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., & Saez-Rodriguez, J. (2019). From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Systems Biology and Applications*, *5*(1), 40. https://doi.org/10.1038/s41540-019-0118-z

Lo Surdo, P., Iannuccelli, M., Contino, S., Castagnoli, L., Licata, L., Cesareni, G., & Perfetto, L. (2023). SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Research*, *51*(D1), D631–D637. https://doi.org/10.1093/nar/gkac883

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Computational Biology*, *11*(3), e1004075. https://doi.org/10.1371/JOURNAL.PCBI.1004075

Luciani, A., & Devuyst, O. (2024). The CTNS-MTORC1 axis couples lysosomal cystine to epithelial cell fate decisions and is a targetable pathway in cystinosis. *Autophagy*, *20*(1), 202–204. https://doi.org/10.1080/15548627.2023.2250165

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., … Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *István A. Kovács 1*, *580*, 402. https://doi.org/10.1038/s41586-020-2188-x

Lund-Johansen, F., Tran, T., & Mehta, A. (2021). Towards reproducibility in large-scale analysis of protein–protein interactions. *Nature Methods 2021 18:7*, *18*(7), 720–721. https://doi.org/10.1038/s41592-021-01202-7

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., & Zhavoronkov, A. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in Genetics*, *9*(JUL), 378508. https://doi.org/10.3389/FGENE.2018.00242/BIBTEX

Marappan, G., & Marappan, G. (2020). Coenzyme Q10: Regulators of Mitochondria and beyond. *Apolipoproteins, Triglycerides and Cholesterol*. https://doi.org/10.5772/INTECHOPEN.89496

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., & Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature Communications*, *5*(1), 3083. https://doi.org/10.1038/ncomms4083

Martínez, V., Navarro, C., Cano, C., Fajardo, W., & Blanco, A. (2015). DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artificial Intelligence in Medicine*, *63*(1), 41–49. https://doi.org/10.1016/j.artmed.2014.11.003

McGowan, E., Sanjak, J., Mathé, E. A., & Zhu, Q. (2023). Integrative rare disease biomedical profile based network supporting drug repurposing or repositioning, a case study of glioblastoma. *Orphanet Journal of Rare Diseases*, *18*(1). https://doi.org/10.1186/S13023-023-02876-2

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426

McKinney, W., & others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Medic, G., van der Weijden, M., Karabis, A., & Hemels, M. (2017). A systematic literature review of cysteamine bitartrate in the treatment of nephropathic cystinosis. *Current Medical Research and Opinion*, *33*(11), 2065–2076. https://doi.org/10.1080/03007995.2017.1354288

Melas, I. N., Sakellaropoulos, T., Iorio, F., Alexopoulos, L. G., Loh, W. Y., Lauffenburger, D. A., Saez-Rodriguez, J., & Bai, J. P. F. (2015). Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integrative Biology* , *7*(8), 904–920. https://doi.org/10.1039/c4ib00294f

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A. L. (2015). Uncovering disease-disease relationships through the incomplete

interactome. *Science*, *347*(6224), 841. https://doi.org/10.1126/SCIENCE.1257601/SUPPL_FILE/MENCHE_SM.PDF

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Academic.Oup.ComH Mi, D Ebert, A Muruganujan, C Mills, LP Albou, T Mushayamaha, PD ThomasNucleic Acids Research, 2021•academic.Oup.Com*, *49*. https://doi.org/10.1093/nar/gkaa1106

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Stephan, R., Tiwari, K., Varusai, T., Weiser, J., … D'Eustachio, P. (2024). The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, *52*(D1), D672–D678. https://doi.org/10.1093/NAR/GKAD1025

Misselbeck, K., Parolo, S., Lorenzini, F., Savoca, V., Leonardelli, L., Bora, P., Morine, M. J., Mione, M. C., Domenici, E., & Priami, C. (2019). A network-based approach to identify deregulated pathways and drug effects in metabolic syndrome. *Nature Communications 2019 10:1*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-13208-z

Mitsos, A., Melas, I. N., Siminelakis, P., Chairakaki, A. D., Saez-Rodriguez, J., & Alexopoulos, L. G. (2009). Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. *PLoS Computational Biology*, *5*(12), e1000591. https://doi.org/10.1371/journal.pcbi.1000591

Morselli Gysi, D., do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., & Barabási, A.-L. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, *118*(19). https://doi.org/10.1073/pnas.2025581118

Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., & Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, *51*(20), 10934–10949. https://doi.org/10.1093/nar/gkad841

Nair, A., Chauhan, P., Saha, B., & Kubatzky, K. F. (2019). Conceptual Evolution of Cell Signaling. *International Journal of Molecular Sciences 2019, Vol. 20, Page 3292*, *20*(13), 3292. https://doi.org/10.3390/IJMS20133292

Nazarieh, M., & Helms, V. (2019). TopControl: A Tool to Prioritize Candidate Disease-associated Genes based on Topological Network Features. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-55954-6

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nielsen, J. (2017). Systems Biology of Metabolism. *Annual Review of Biochemistry*, *86*(1), 245–275. https://doi.org/10.1146/annurev-biochem-061516-044757

Nosengo, N. (2016). Can you teach old drugs new tricks? *Nature*, *534*(7607), 314–316. https://doi.org/10.1038/534314A

Oellrich, A., Hoehndorf, R., Gkoutos, G. V., & Rebholz-Schuhmann, D. (2012). Improving Disease Gene Prioritization by Comparing the Semantic Similarity of Phenotypes in Mice with Those of Human Diseases. *PLOS ONE*, *7*(6), e38937. https://doi.org/10.1371/JOURNAL.PONE.0038937

Olsen, T., Jackson, B., Feeser, T., Kent, M., Moad, J., Krishnamurthy, S., Lunsford, D., & Soans, R. (2018). Diagnostic Performance of Deep Learning Algorithms Applied to Three Common Diagnoses in Dermatopathology. *Journal of Pathology Informatics*, *9*(1), 32. https://doi.org/10.4103/JPI.JPI_31_18

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science : A Publication of the Protein Society*, *30*(1), 187. https://doi.org/10.1002/PRO.3978

Ozen, M., Emamian, E. S., & Abdi, A. (2022). Learning feedback molecular network models using integer linear programming. *Physical Biology*, *19*(6), 066004. https://doi.org/10.1088/1478-3975/ac920d

Pache De Faria Guimaraes, L., Seguro, A. C., Shimizu, M. H. M., Lopes Neri, L. A., Sumita, N. M., De Bragança, A. C., Aparecido Volpini, R., Cunha Sanches, T. R., Macaferri Da Fonseca, F. A., Moreira Filho, C. A., & Vaisbich, M. H. (2014). N-acetyl-cysteine is associated to renal function improvement in patients with nephropathic cystinosis. *Pediatric Nephrology*, *29*(6), 1097–1102. https://doi.org/10.1007/S00467-013-2705-3/METRICS

Paganin, M., Tebaldi, T., Lauria, F., & Viero, G. (2023). Visualizing gene expression changes in time, space, and single cells with expressyouRcell. *IScience*, *26*(6), 106853. https://doi.org/10.1016/j.isci.2023.106853

Parolo, S., Mariotti, F., Bora, P., Carboni, L., & Domenici, E. (2023). Single-cell-led drug repurposing for Alzheimer's disease. *Scientific Reports 2023 13:1*, *13*(1), 1–14. https://doi.org/10.1038/s41598-023-27420-x

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/abs/1912.01703v1

Paton, V., Ramirez Flores, R. O., Gabor, A., Badia-i-Mompel, P., Tanevski, J., Garrido-Rodriguez, M., & Saez-Rodriguez, J. (2024). Assessing the impact of transcriptomics data analysis pipelines on downstream functional enrichment results. *Nucleic Acids Research*, *52*(14), 8100–8111. https://doi.org/10.1093/NAR/GKAE552

Pavan, S., Rommel, K., Marquina, M. E. M., Höhn, S., Lanneau, V., & Rath, A. (2017). Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLOS ONE*, *12*(1), e0170365. https://doi.org/10.1371/JOURNAL.PONE.0170365

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, *4*(1), 1–27. https://doi.org/10.1186/1756-0381-4-10/FIGURES/11

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), Art. 1.0.2. http://jmlr.org/papers/v12/pedregosa11a.html

Peng, Y., Yuan, M., Xin, J., Liu, X., & Wang, J. (2020). Screening novel drug candidates for Alzheimer's disease by an integrated network and transcriptome analysis. *Bioinformatics*, *36*(17), 4626–4632. https://doi.org/10.1093/BIOINFORMATICS/BTAA563

Pons, P., & Latapy, M. (2006). Computing Communities in Large Networks Using Random Walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. https://doi.org/10.7155/JGAA.00124

Porras Millan, P., Ochoa, D., Rogon, M., Garcia-Alonso, L., & Turei, D. (2022). *Network analysis of protein interaction data*. EMBL-EBI Training.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, *17*(2), 147–154. https://doi.org/10.1038/s41592-019-0690-6

Priya, S., Tripathi, G., Singh, D. B., Jain, P., & Kumar, A. (2022). Machine learning approaches and their applications in drug discovery and design. *Chemical Biology & Drug Design*, *100*(1), 136–153. https://doi.org/10.1111/cbdd.14057

Pushpakom, S., Iorio, F., Eyers, P. A., Jane Escott, K., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., & Pirmohamed, M. (2018). *Drug repurposing: progress, challenges and recommendations*. https://doi.org/10.1038/nrd.2018.168

Qiu, Y., Jiang, H., Ching, W.-K., & Cheng, X. (2018). Discovery of Boolean metabolic networks: integer linear programming based approach. *BMC Systems Biology*, *12*(S1), 7. https://doi.org/10.1186/s12918-018-0528-3

Quinn, T. P., Richardson, M. F., Lovell, D., & Crowley, T. M. (2017). propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Scientific Reports 2017 7:1*, *7*(1), 1–9. https://doi.org/10.1038/s41598-017-16520-0

Rajpal, S., Agarwal, M., Kumar, V., Gupta, A., & Kumar, N. (2021). Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification. *IEEE Access*, *9*, 103347–103364. https://doi.org/10.1109/ACCESS.2021.3093616

Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T. J., Goodman, R. P., Grabarek, Z., Haas, M. E., Hung, W. H. W., Joshi, P. R., Jourdain, A. A., Kim, S. H., Kotrys, A. V, Lam, S. S., McCoy, J. G., Meisel, J. D., Miranda, M., Panda, A., … Mootha, V. K. (2021). MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research*, *49*(D1), D1541–D1547. https://doi.org/10.1093/nar/gkaa1011

Ratto, E., Chowdhury, S. R., Siefert, N. S., Schneider, M., Wittmann, M., Helm, D., & Palm, W. (2022). Direct control of lysosomal catabolic activity by mTORC1 through regulation of V-ATPase assembly. *Nature Communications 2022 13:1*, *13*(1), 1–15. https://doi.org/10.1038/s41467-022-32515-6

Revilla Sancho, L. (2024). *BioCor: Functional similarities, R package version 1.28.0*. https://bioconductor.org/packages/release/bioc/html/BioCor.html

Richards, A. L., Eckhardt, M., & Krogan, N. J. (2021). Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Mol Syst Biol*, *17*, 8792. https://doi.org/10.15252/msb.20188792

Richardson, L. (2007). Beautiful soup documentation. *April*.

Richelle, A., Chiang, A. W. T., Kuo, C.-C., & Lewis, N. E. (2019). Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. *PLOS Computational Biology*, *15*(4), e1006867. https://doi.org/10.1371/journal.pcbi.1006867

Riniker, S., Wang, Y., Jenkins, J. L., & Landrum, G. A. (2014). Using information from historical high-throughput screens to predict active compounds. *Journal of Chemical Information and Modeling*, *54*(7), 1880–1891. https://doi.org/10.1021/CI500190P/SUPPL_FILE/CI500190P_SI_002.ZIP

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P. E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., Limeta, A., Hedin, A., Gustafsson, J., Kerkhoven, E. J., Svensson, L. T., Palsson, B. O., Mardinoglu, A., Hansson, L., Uhlén, M., & Nielsen, J. (2020). An atlas of human metabolism. *Science Signaling*, *13*(624), 1482. https://doi.org/10.1126/SCISIGNAL.AAZ1482/SUPPL_FILE/AAZ1482_SM.PDF

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), 1–9. https://doi.org/10.1186/GB-2010-11-3-R25/FIGURES/3

Rodríguez Mier, P., Gabor, A., & Saez-Rodriguez, J. (2024). *CORNETO: A Unified Omics-Driven Framework for Network Inference*. https://saezlab.github.io/corneto/main/index.html

Rodríguez Mier, P., & Türei, D. (2023). *PerMedCoE summer school: from pathway modelling tools to cell-level simulations*.

Rodríguez-Mier, P., Poupin, N., de Blasio, C., Le Cam, L., & Jourdan, F. (2021). DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks. *PLOS Computational Biology*, *17*(2), e1008730. https://doi.org/10.1371/journal.pcbi.1008730

Roessler, H. I., Knoers, N. V. A. M., van Haelst, M. M., & van Haaften, G. (2021). Drug Repurposing for Rare Diseases. *Trends in Pharmacological Sciences*, *42*(4), 255–267. https://doi.org/10.1016/j.tips.2021.01.003

Röhl, A., & Bockmayr, A. (2017). A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. *BMC Bioinformatics*, *18*(1), 2. https://doi.org/10.1186/s12859-016-1412-z

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., … Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, *159*(5), 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. https://doi.org/10.1140/epjst/e2010-01179-1

Sadegh, S., Skelton, J., Anastasi, E., Bernett, J., Blumenthal, D. B., Galindez, G., Salgado-Albarrán, M., Lazareva, O., Flanagan, K., Cockell, S., Nogales, C., Casas, A. I., Schmidt,

H. H. H. W., Baumbach, J., Wipat, A., & Kacprowski, T. (2021). Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nature Communications 2021 12:1*, *12*(1), 1–12. https://doi.org/10.1038/s41467-021-27138-2

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., & Sorger, P. K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, *5*(1). https://doi.org/10.1038/msb.2009.87

Sahasrabudhe, S. A., Terluk, M. R., & Kartha, R. V. (2023). N-acetylcysteine Pharmacology and Applications in Rare Diseases—Repurposing an Old Antioxidant. *Antioxidants 2023, Vol. 12, Page 1316*, *12*(7), 1316. https://doi.org/10.3390/ANTIOX12071316

Sansanwal, P., Li, L., Hsieh, S., & Sarwal, M. M. (2010). Insights into novel cellular injury mechanisms by gene expression profiling in nephropathic cystinosis. *Journal of Inherited Metabolic Disease*, *33*(6), 775–786. https://doi.org/10.1007/s10545-010-9203-6

Sardana, D., Zhu, C., Zhang, M., Gudivada, R. C., Yang, L., & Jegga, A. G. (2011). Drug repositioning for orphan diseases. *Briefings in Bioinformatics*, *12*(4), 346–356. https://doi.org/10.1093/BIB/BBR021

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, *7*(1), 1–16. https://doi.org/10.1186/1471-2105-7-302/FIGURES/13

Siegenthaler, C., & Gunawan, R. (2014). Assessment of Network Inference Methods: How to Cope with an Underdetermined Problem. *PLoS ONE*, *9*(3), e90481. https://doi.org/10.1371/journal.pone.0090481

Skinnider, M. A., & Foster, L. J. (2021). Meta-analysis defines principles for the design and analysis of co-fractionation mass spectrometry experiments. *Nature Methods 2021 18:7*, *18*(7), 806–815. https://doi.org/10.1038/s41592-021-01194-4

Skinnider, M. A., Squair, J. W., & Foster, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nature Methods 2019 16:5*, *16*(5), 381–386. https://doi.org/10.1038/s41592-019-0372-4

Smith, C. L., & Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *1*(3), 390–399. https://doi.org/10.1002/WSBM.44

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., & Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Molecular Systems Biology*, *11*(12), 848. https://doi.org/10.15252/MSB.20156351

Solano Noriega, J. J., Leyva López, J. C., Browne, F., & Liu, J. (2021). An Outranking Approach for Gene Prioritization Using Multinetworks. *Int. J. Comput. Intell. Syst.*, *14*(1), 1728–1741. https://doi.org/10.2991/IJCIS.D.210608.003

Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., & Altman, R. B. (2020). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *25*, 463–474.

Stewart, B. J., Ferdinand, J. R., Young, M. D., Mitchell, T. J., Loudon, K. W., Riding, A. M., Richoz, N., Frazer, G. L., Staniforth, J. U. L., Braga, F. A. V., Botting, R. A., Popescu, D. M., Vento-Tormo, R., Stephenson, E., Cagan, A., Farndon, S. J., Polanski, K., Efremova, M., Green, K., … Clatworthy, M. R. (2019). Spatiotemporal immune zonation of the human kidney. *Science*, *365*(6460), 1461–1466. https://doi.org/10.1126/SCIENCE.AAT5031/SUPPL_FILE/AAT5031_DATA_S9.CSV

Stolfi, P., Mastropietro, A., Pasculli, G., Tieri, P., & Vergni, D. (2023). NIAPU: network-informed adaptive positive-unlabeled learning for disease gene identification. *Bioinformatics*, *39*(2). https://doi.org/10.1093/BIOINFORMATICS/BTAC848

Su, Y., Wu, J., Li, X., Li, J., Zhao, X., Pan, B., Huang, J., Kong, Q., & Han, J. (2023). DTSEA: A network-based drug target set enrichment analysis method for drug repurposing against COVID-19. *Computers in Biology and Medicine*, *159*, 106969. https://doi.org/10.1016/j.compbiomed.2023.106969

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, *49*(D1), D605–D612. https://doi.org/10.1093/NAR/GKAA1074

Takano, K., Yatabe, M. S., Abe, A., Suzuki, Y., Sanada, H., Watanabe, T., Kimura, J., & Yatabe, J. (2014). Characteristic Expressions of GABA Receptors and GABA Producing/Transporting Molecules in Rat Kidney. *PLOS ONE*, *9*(9), e105835. https://doi.org/10.1371/JOURNAL.PONE.0105835

Tamura, T., Muto-fujita, A., Tohsato, Y., & Kosaka, T. (2023). Gene Deletion Algorithms for Minimum Reaction Network Design by Mixed-Integer Linear Programming for Metabolite Production in Constraint-Based Models: gDel_minRN. *Journal of Computational Biology*, *30*(5), 553–568. https://doi.org/10.1089/cmb.2022.0352

Tang, Y.-Y., Wei, P.-J., Zhao, J., Xia, J., Cao, R.-F., & Zheng, C.-H. (2021). Identification of driver genes based on gene mutational effects and network centrality. *BMC Bioinformatics*, *22*(S3), 457. https://doi.org/10.1186/s12859-021-04377-0

Teng, D., Gong, Y., Wu, Z., Li, W., Tang, Y., & Liu, G. (2022). In Silico Prediction of Potential Drug Combinations for Type 2 Diabetes Mellitus by an Integrated Network and Transcriptome Analysis. *ChemMedChem*, *17*(3). https://doi.org/10.1002/cmdc.202100620

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., Iersel, M. van, Lauffenburger, D. A., & Saez-Rodriguez, J. (2012). CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Systems Biology*, *6*(1), 133. https://doi.org/10.1186/1752-0509-6-133

The Gene Ontology Consortium. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1). https://doi.org/10.1093/GENETICS/IYAD031

The MAQC Consortium. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, *28*(8), 827–838. https://doi.org/10.1038/nbt.1665

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bölling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., … Palsson, B. Ø. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, *31*(5), 419–425. https://doi.org/10.1038/nbt.2488

Tognetti, M., Sklodowski, K., Müller, S., Kamber, D., Muntel, J., Bruderer, R., & Reiter, L. (2022). Biomarker Candidates for Tumors Identified from Deep-Profiled Plasma Stem Predominantly from the Low Abundant Area. *Journal of Proteome Research*, *21*(7), 1718–1735. https://doi.org/10.1021/ACS.JPROTEOME.2C00122/ASSET/IMAGES/LARGE/PR2C00122_0007.JPEG

Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, *13*(12), 966–967. https://doi.org/10.1038/nmeth.4077

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T., & Saez-Rodriguez, J. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, *17*(3). https://doi.org/10.15252/msb.20209923

Valdeolivas, A., Türei, D., & Gabor, A. (2016). *OmnipathR: client for the OmniPath web service*.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine

learning in drug discovery and development. *Nature Reviews Drug Discovery 2019 18:6*, *18*(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Väremo, L., Nielsen, J., & Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, *41*(8), 4378–4391. https://doi.org/10.1093/nar/gkt111

Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell*, *144*(6), 986–998. https://doi.org/10.1016/j.cell.2011.02.016

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vitali, F., Cohen, L. D., Demartini, A., Amato, A., Eterno, V., Zambelli, A., & Bellazzi, R. (2016). A Network-Based Data Integration Approach to Support Drug Repurposing and Multi-Target Therapies in Triple Negative Breast Cancer. *PLOS ONE*, *11*(9), e0162407. https://doi.org/10.1371/journal.pone.0162407

Walakira, A., Rozman, D., Režen, T., Mraz, M., & Moškon, M. (2021). Guided extraction of genome-scale metabolic models for the integration and analysis of omics data. *Computational and Structural Biotechnology Journal*, *19*, 3521–3530. https://doi.org/10.1016/j.csbj.2021.06.009

Wang, Y. Bin, You, Z. H., Yang, S., Yi, H. C., Chen, Z. H., & Zheng, K. (2020). A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Medical Informatics and Decision Making*, *20*(2), 1–9. https://doi.org/10.1186/S12911-020-1052-0/TABLES/4

Wang, H., Robinson, J. L., Kocabas, P., Gustafsson, J., Anton, M., Cholley, P.-E., Huang, S., Gobom, J., Svensson, T., Uhlen, M., Zetterberg, H., & Nielsen, J. (2021). Genome-scale metabolic network reconstruction of model animals as a platform for translational research. *Proceedings of the National Academy of Sciences*, *118*(30). https://doi.org/10.1073/pnas.2102344118

Wang, J., Luo, L., Ding, Q., Wu, Z., Peng, Y., Li, J., Wang, X., Li, W., Liu, G., Zhang, B., & Tang, Y. (2021). Development of a Multi-Target Strategy for the Treatment of Vitiligo via Machine Learning and Network Analysis Methods. *Frontiers in Pharmacology*, *12*, 754175. https://doi.org/10.3389/FPHAR.2021.754175/BIBTEX

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274–1281. https://doi.org/10.1093/BIOINFORMATICS/BTM087

Wang, R. S., & Loscalzo, J. (2018). Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. *Journal of Molecular Biology*, *430*(18), 2939–2950. https://doi.org/10.1016/J.JMB.2018.05.016

Wang, Y., Guo, M., Ren, Y., Jia, L., & Yu, G. (2019). Drug repositioning based on individual bi-random walks on a heterogeneous network. *BMC Bioinformatics*, *20*(S15), 547. https://doi.org/10.1186/s12859-019-3117-6

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-Learning-Based Drug–Target Interaction Prediction. *Journal of Proteome Research*, *16*(4), 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org

Wickham, H. (2024). *rvest: Easily Harvest (Scrape) Web Pages*. https://rvest.tidyverse.org/

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org

Wickham, H., Hester, J., & Ooms, J. (2023). *xml2: Parse XML*. https://xml2.r-lib.org/

Wickham, H., Vaughan, D., & Girlich, M. (2024). *tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org

Wildman, S. S., Dunn, K., Van Beusecum, J. P., Inscho, E. W., Kelley, S., Lilley, R. J., Cook, A. K., Taylor, K. D., & Peppiatt-Wildman, C. M. (2023). A novel functional role for the classic CNS neurotransmitters, GABA, glycine, and glutamate, in the kidney: potent and opposing regulators of the renal vasculature. *American Journal of Physiology - Renal Physiology*, *325*(1), F38–F49. https://doi.org/10.1152/AJPRENAL.00425.2021/ASSET/IMAGES/MEDIUM/F-00425-2021R01.PNG

Wilmer, M. J., Emma, F., & Levtchenko, E. N. (2010). The pathogenesis of cystinosis: Mechanisms beyond cystine accumulation. *American Journal of Physiology - Renal Physiology*, *299*(5), 905–916. https://doi.org/10.1152/AJPRENAL.00318.2010/ASSET/IMAGES/LARGE/ZH20111060650003.JPEG

Wilmer, M. J., Kluijtmans, L. A. J., van der Velden, T. J., Willems, P. H., Scheffer, P. G., Masereeuw, R., Monnens, L. A., van den Heuvel, L. P., & Levtchenko, E. N. (2011). Cysteamine restores glutathione redox status in cultured cystinotic proximal tubular epithelial cells. *Biochimica et Biophysica Acta*, *1812*(6), 643–651. https://doi.org/10.1016/J.BBADIS.2011.02.010

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., … Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037

Wishart, D. S., Guo, A. C., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., … Gautam, V. (2022). HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, *50*(D1), D622–D631. https://doi.org/10.1093/NAR/GKAB1062

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Wu, Z., Li, W., Liu, G., & Tang, Y. (2018). Network-Based Methods for Prediction of Drug-Target Interactions. *Frontiers in Pharmacology*, *9*. https://doi.org/10.3389/fphar.2018.01134

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., & Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics*, *41*(3), 316–323. https://doi.org/10.1038/ng.337

Yepes, S., Tucker, M. A., Koka, H., Xiao, Y., Jones, K., Vogt, A., Burdette, L., Luo, W., Zhu, B., Hutchinson, A., Yeager, M., Hicks, B., Freedman, N. D., Chanock, S. J., Goldstein, A. M., & Yang, X. R. (2020). Using whole-exome sequencing and protein interaction networks to prioritize candidate genes for germline cutaneous melanoma susceptibility. *Scientific Reports*, *10*(1). https://doi.org/10.1038/S41598-020-74293-5

Youssef, I., Law, J., & Ritz, A. (2019). Integrating protein localization with automated signaling pathway reconstruction. *BMC Bioinformatics*, *20*(S16), 505. https://doi.org/10.1186/s12859-019-3077-x

Yu, G. (2020). Gene Ontology Semantic Similarity Analysis Using GOSemSim. *Methods in Molecular Biology*, *2117*, 207–215. https://doi.org/10.1007/978-1-0716-0301-7_11

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, *26*(7), 976–978. https://doi.org/10.1093/BIOINFORMATICS/BTQ064

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, Z., Wu, Z., Wang, Z., Wang, Y., Zhou, M., Li, W., Liu, G., & Tang, Y. (2024). Network-Based Methods and Their Applications in Drug Discovery. *Journal of Chemical Information and Modeling*, *64*(1), 57–75. https://doi.org/10.1021/acs.jcim.3c01613

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128/MACHINEREADABLECITATION/RIS

Zhang, Y., Li, W., Zhang, Y., Hu, E., Rong, Z., Ge, L., Deng, G., He, Y., Lv, J., Chen, L., & He, W. (2020). Network-based integration method for potential breast cancer gene identification. *Journal of Cellular Physiology*, *235*(11), 7960–7969. https://doi.org/10.1002/JCP.29450

Zhu, Q., Nguyễn, Đ.-T., Sheils, T., Alyea, G., Sid, E., Xu, Y., Dickens, J., Mathé, E. A., & Pariser, A. (2021). Scientific evidence based rare disease research discovery with research funding data in knowledge graph. *Orphanet Journal of Rare Diseases*, *16*(1), 483. https://doi.org/10.1186/s13023-021-02120-9

Zitnik, M., Li, M. M., Wells, A., Glass, K., Deisy, #, Gysi, M., Krishnan, A., Murali, T. M., Radivojac, P., Roy, S., Ana¨ıs, #, Baudot, A., Bozdag, S., Chen, D. Z., Cowen, L., Devkota, K., Gitter, A., Gosline, S., Gu, P., … Milenkovi´cmilenkovi´c, T. (2024). Current and future directions in network biology ISCB: International Society for Computational Biology LLM: large language model PPI: protein-protein interaction TF: transcription factor TGFβ: transforming growth factor-beta. *Ziynet Nesibe Kesimoglu*, *14*, 27.

List of Figure

List of Table

Appendix

Table A7 - List of Phenotype genes from the literature

ENTREZID	Symbol	Source
182	JAG1	HPO
185	AGTR1	MGI
229	ALDOB	HPO
355	FAS	MGI
359	AQP2	MGI
364	AQP7	MGI
463	ZFHX3	MGI
525	ATP6V1B1	HPO, MGI
540	ATP7B	HPO
617	BCS1L	HPO
624	BDKRB2	MGI
760	CA2	HPO, MGI
790	CAD	HPO
1184	CLCN5	HPO
1188	CLCNKB	HPO
1244	ABCC2	MGI
1282	COL4A1	MGI
1364	CLDN4	MGI
1374	CPT1A	HPO
1497	CTNS	HPO
1585	CYP11B2	HPO
1758	DMP1	MGI
1816	DRD5	MGI
1907	EDN2	MGI
1962	EHHADH	HPO
2067	ERCC1	MGI
2099	ESR1	MGI
2108	ETFA	HPO
2109	ETFB	HPO
2110	ETFDH	HPO
2184	FAH	HPO
2299	FOXI1	MGI
2335	FN1	HPO
2581	GALC	MGI
2625	GATA3	HPO
2628	GATM	HPO
2645	GCK	HPO
2778	GNAS	HPO
2828	GPR4	MGI

ENTREZID	Symbol	Source
5159	PDGFRB	MGI
5251	PHEX	HPO
5256	PHKA2	HPO
5261	PHKG2	HPO
5373	PMM2	HPO
5641	LGMN	MGI
5660	PSAP	MGI
5824	PEX19	HPO
5828	PEX2	MGI
6514	SLC2A2	HPO
6521	SLC4A1	HPO, MGI
6550	SLC9A3	MGI
6557	SLC12A1	MGI
6559	SLC12A3	HPO
6561	SLC13A1	MGI
6565	SLC15A2	MGI
6569	SLC34A1	HPO
6584	SLC22A5	MGI
6648	SOD2	MGI
6774	STAT3	HPO
6775	STAT4	MGI
6833	ABCC8	HPO
6834	SURF1	HPO
6928	HNF1B	HPO
7021	TFAP2B	MGI
7809	BSND	MGI
8111	GPR68	MGI
8671	SLC4A4	HPO, MGI
8879	SGPL1	MGI
8942	KYNU	HPO
9075	CLDN2	MGI
9429	ABCG2	MGI
9451	EIF2AK3	HPO
9498	SLC4A8	MGI
9997	SCO2	HPO
10686	CLDN16	HPO, MGI
10723	SLC12A7	MGI
11234	HPS5	MGI
11275	KLHL2	MGI

ENTREZID	Symbol	Source
2896	GRN	MGI
2981	GUCA2B	MGI
3043	HBB	HPO
3172	HNF4A	HPO
3240	HP	MGI
3250	HPR	MGI
3257	HPS1	MGI
3263	HPX	MGI
3565	IL4	MGI
3630	INS	HPO
3651	PDX1	HPO
3767	KCNJ11	HPO
3913	LAMB2	MGI
3958	LGALS3	MGI
4036	LRP2	HPO
4043	LRPAP1	MGI
4240	MFGE8	MGI
4512	COX1	HPO
4513	COX2	HPO
4514	COX3	HPO
4535	ND1	HPO
4536	ND2	HPO
4537	ND3	HPO
4538	ND4	HPO
4540	ND5	HPO
4541	ND6	HPO
4558	TRNF	HPO
4564	TRNH	HPO
4567	TRNL1	HPO
4570	TRNN	HPO
4572	TRNQ	HPO
4574	TRNS1	HPO
4575	TRNS2	HPO
4578	TRNW	HPO
4594	MMUT	HPO, MGI
4694	NDUFA1	HPO
4700	NDUFA6	HPO
4709	NDUFB3	HPO
4714	NDUFB8	HPO
4715	NDUFB9	HPO
4716	NDUFB10	HPO
4719	NDUFS1	HPO
4720	NDUFS2	HPO

ENTREZID	Symbol	Source
22824	HSPA4L	MGI
25915	NDUFAF3	HPO
26235	FBXL4	HPO
26258	BLOC1S6	MGI
26276	VPS33B	HPO
27235	COQ2	HPO
29078	NDUFAF4	HPO
50484	RRM2B	HPO, MGI
50507	NOX4	MGI
50617	ATP6V0A4	HPO, MGI
51103	NDUFAF1	HPO
51300	TIMMDC1	HPO
51458	RHCG	MGI
53405	CLIC5	MGI
54539	NDUFB11	HPO
55005	RMND1	HPO
55151	TMEM38B	MGI
55243	KIRREL1	MGI
55327	LIN7C	MGI
55330	BLOC1S4	MGI
55572	FOXRED1	HPO
55863	TMEM126B	HPO
56606	SLC2A9	MGI
56670	SUCNR1	MGI
57107	PDSS2	HPO
57511	COG6	HPO
57724	EPG5	HPO
57835	SLC4A5	MGI
63894	VIPAS39	HPO
65010	SLC26A6	MGI
65082	VPS33A	MGI
79133	NDUFAF5	HPO
79783	SUGCT	MGI
80224	NUBPL	HPO
83697	SLC4A9	MGI
84062	DTNBP1	MGI
84159	ARID5B	MGI
84300	UQCC2	HPO
84334	COA8	HPO
91942	NDUFAF2	HPO
115111	SLC26A7	MGI
116085	SLC22A12	MGI
116228	COX20	HPO

ENTREZID	Symbol	Source
4722	NDUFS3	HPO
4723	NDUFV1	HPO
4724	NDUFS4	HPO
4726	NDUFS6	HPO
4728	NDUFS8	HPO
4729	NDUFV2	HPO
4842	NOS1	MGI
4853	NOTCH2	HPO
4868	NPHS1	MGI
4952	OCRL	HPO
5019	OXCT1	MGI
5091	PC	HPO
5136	PDE1A	MGI

ENTREZID	Symbol	Source
125206	SLC5A10	MGI
126328	NDUFA11	HPO
133686	NADK2	HPO
137682	NDUFAF6	HPO
142680	SLC34A3	HPO
284184	NDUFAF8	HPO
348932	SLC6A18	MGI
374291	NDUFS7	HPO
389840	MAP3K15	MGI
441432	AQP7P3	MGI
100131801	PET100	HPO
107436076	CYP2C23P	MGI

Table A8 - GO cellular component enrichment results.

Modules	GO ID	Description	P-value	P.adjust	General
M_12	GO:0005796	Golgi lumen	0.001	0.038	Golgi apparatus
M_122	GO:0034702	Ion channel complex	<0.0001	<0.0001	membrane
M_122	GO:1902495	Transmembrane transporter complex	<0.0001	<0.0001	membrane
M_130	GO:0005759	Mitochondrial matrix	<0.0001	<0.0001	mitochondrion
M_130	GO:0030062	Mitochondrial tricarboxylic acid cycle enzyme complex	0.001	0.009	cytoplasm
M_131	GO:0017101	Aminoacyl-trna synthetase multienzyme	<0.0001	<0.0001	extracellular region
M_131	GO:0017101	Aminoacyl-trna synthetase multienzyme	<0.0001	<0.0001	nucleoplasm
M_132	GO:0016324	Apical plasma membrane	<0.0001	0.002	membrane
M_132	GO:0031526	Brush border membrane	<0.0001	0.002	membrane
M_134	GO:0005759	Mitochondrial matrix	<0.0001	<0.0001	mitochondrion
M_134	GO:0098803	Respiratory chain complex	<0.0001	<0.0001	membrane
M_137	GO:0005778	Peroxisomal membrane	<0.0001	<0.0001	cytoplasm
M_137	GO:0031903	Microbody membrane	<0.0001	<0.0001	cytoplasm
M_143	GO:0005811	Lipid droplet	<0.0001	<0.0001	lipid droplet
M_143	GO:0035577	Azurophil granule membrane	0.001	0.016	lysosome
M_151	GO:0005759	Mitochondrial matrix	<0.0001	<0.0001	mitochondrion
M_151	GO:1990204	Oxidoreductase complex	<0.0001	<0.0001	extracellular region
M_26	GO:0005759	Mitochondrial matrix	<0.0001	<0.0001	mitochondrion
M_26	GO:0005743	Mitochondrial inner membrane	<0.0001	0.001	mitochondrion
M_265	GO:1902495	Transmembrane transporter complex	<0.0001	<0.0001	membrane
M_265	GO:1990351	Transporter complex	<0.0001	<0.0001	extracellular region
M_33	GO:0033176	Proton-transporting V-type atpase	<0.0001	<0.0001	membrane
M_33	GO:0016471	Vacuolar proton-transporting V-type atpase complex	<0.0001	<0.0001	cytoplasm
M_34	GO:0005743	Mitochondrial inner membrane	<0.0001	0.001	mitochondrion
M_34	GO:0005743	Mitochondrial inner membrane	<0.0001	0.001	membrane
M_49	GO:0005759	Mitochondrial matrix	<0.0001	0.007	mitochondrion
M_49	GO:0031514	Motile cilium	<0.0001	0.012	extracellular region
M_5	GO:0101002	Ficolin-1-rich granule	<0.0001	<0.0001	cytoplasm
M_5	GO:1904813	Ficolin-1-rich granule lumen	<0.0001	0.002	cytoplasm
M_54	GO:0005743	Mitochondrial inner membrane	<0.0001	<0.0001	mitochondrion
M_54	GO:0005743	Mitochondrial inner membrane	<0.0001	<0.0001	membrane
M_72	GO:0005776	Autophagosome	<0.0001	<0.0001	cytoplasm
M_72	GO:0036452	ESCRT complex	<0.0001	<0.0001	endosome
M_89	GO:0045171	Intercellular bridge	<0.0001	<0.0001	extracellular region
M_89	GO:0045171	Intercellular bridge	<0.0001	<0.0001	nucleoplasm
M_9	GO:0005747	Mitochondrial Respiratory Chain Complex I	<0.0001	<0.0001	mitochondrion
M_9	GO:0005747	Mitochondrial Respiratory Chain Complex I	<0.0001	<0.0001	membrane
M_95	GO:0005811	Lipid droplet	<0.0001	0.003	lipid droplet

Table A9 - Hallmark enrichment on modules

Module	Description	P-value	P.adjust	Gene symbol
M_12	Glycolysis	0.000	0.000	B4GALT7/B3GALT6/GLCE/B3GAT3/ B3GAT1/SDC2/XYLT2/HS2ST1
M_12	Bile acid metabolism	0.001	0.004	DIO1/DIO2/SULT1B1/SLC35B2
M_12	Hypoxia	0.001	0.004	B3GALT6/NDST1/SDC2/NDST2/HS3ST1
M_122	Inflammatory response	0.000	0.004	KCNMB2/GABBR1/KCNA3/ KCNJ2/BEST1/SLC4A4
M_130	Oxidative phosphorylation	0.000	0.000	GLUD1/GOT2/IDH1/IDH2/IDH3A/ IDH3B/OAT/POR/SLC25A11
M_130	Bile acid metabolism	0.002	0.014	IDH1/IDH2/PECR/BBOX1
M_130	Xenobiotic metabolism	0.003	0.014	CDO1/IDH1/POR/KYNU/NFS1
M_130	Fatty acid metabolism	0.009	0.035	IDH1/IDH3B/AADAT/D2HGDH
M_131	Mtorc1 signaling	0.000	0.001	EPRS1/GCLC/GOT1/PSAT1/ASNS/ BCAT1/CYB5B/GMPS/EEF1E1
M_134	Oxidative phosphorylation	0.000	0.000	FH/SDHA/SDHB/SDHC/SDHD
M_134	Fatty acid metabolism	0.000	0.000	FH/SDHA/SDHC/SDHD
M_137	Fatty acid metabolism	0.000	0.001	ACSL1/ACSL4/HMGCS1/ ACAT2/ACSL5/CD36
M_137	Cholesterol homeostasis	0.000	0.003	HMGCR/HMGCS1/ACAT2/ACSS2
M_137	Peroxisome	0.001	0.005	SLC27A2/ACSL1/ACSL4/ACSL5
M_137	Bile acid metabolism	0.001	0.005	SLC27A5/SLC27A2/ACSL1/ACSL5
M_137	Mtorc1 signaling	0.002	0.007	ACSL3/HMGCR/HMGCS1/ACLY/NMT1
M_137	Androgen response	0.010	0.032	ACSL3/HMGCR/HMGCS1
M_14	Hypoxia	0.000	0.001	GBE1/GYS1/PPP1R3C/UGP2
M_14	Glycolysis	0.002	0.009	GALE/GYS1/UGP2
M_151	Oxidative phosphorylation	0.000	0.000	CS/DLAT/DLD/DLST/ALAS1/ACAT1/ ALDH6A1/OGDH/PDHA1/PDHB/ MPC1/BCKDHA/PDHX
M_151	Fatty acid metabolism	0.000	0.000	DLD/DLST/HIBCH/HMGCL/HMGCS2/ PDHA1/PDHB/BCKDHB/ACSS1
M_151	Adipogenesis	0.001	0.004	CS/DBT/DLAT/DLD/HIBCH/BCKDHA
M_26	Mtorc1 signaling	0.002	0.024	MTHFD2/DHFR/MTHFD2L/SHMT2
M_265	Myogenesis	0.000	0.000	SORBS3/MYL6B/FKBP1B/MYL11/ MYL3/MYL4/FXYD1/SLC6A8/SPARC/ TNNC2/TNNC1/TNNI1/TNNI2/ TNNT2/TPM2/TPM3/CAMK2B/ SORBS1/ITGB5/DES/DMD/ACTN2/ MYH3/MYH8
M_33	Oxidative phosphorylation	0.000	0.000	TCIRG1/ATP6V1D/ATP6V1H/ATP6V0C/ ATP6V1C1/ATP6V1E1/ATP6AP1/ ATP6V0E1/ATP6V1F/ATP6V1G1/ATP6V0B
M_33	Uv response up	0.002	0.020	GCH1/AQP3/ATP6V1C1/CA2/ATP6V1F
M_47	Dna repair	0.000	0.000	ADA/DGUOK/GUK1/NT5C/ITPA/

				PNP/NT5C3A/CDA
M_47	Il2 stat5 signaling	0.006	0.047	DCPS/SLC29A2/PNP/NT5E
M_49	Dna repair	0.000	0.000	CMPK2/AK1/NME1/NME3/ NME4/RRM2B/TK2
M_5	Hypoxia	0.000	0.000	FBP1/SLC37A4/GPI/HK1/ PFKFB3/PGM1/PGM2/ SLC2A1/SLC2A3/SLC2A5
M_5	Glycolysis	0.000	0.000	GNPDA1/AGL/G6PD/SLC37A4/ MPI/PFKFB1/PMM2/PGM2/PRPS1
M_5	Mtorc1 signaling	0.001	0.011	G6PD/SLC37A4/GPI/HPRT1/ PGM1/SLC2A1/SLC2A3
M_5	Xenobiotic metabolism	0.004	0.027	FBP1/G6PC1/HPRT1/ DCXR/PMM1/UPP1
M_54	Oxidative phosphorylation	0.000	0.000	PMPCA/FXN/NDUFB8/ PHB2/GRPEL1/TIMM50/ AFG3L2/TIMM17A
M_54	Myc targets v2	0.010	0.036	HSPD1/HSPE1
M_54	Myc targets v1	0.018	0.042	PHB2/HSPD1/HSPE1
M_89	Xenobiotic metabolism	0.001	0.009	GSR/GSS/GSTA3/ GSTM4/GSTO1
M_89	Reactive oxygen species pathway	0.001	0.009	GSR/MGST1/PRDX2
M_9	Oxidative phosphorylation	0.000	0.000	NDUFS7/NDUFA1/NDUFA2/ NDUFA3/NDUFA4/NDUFA5/ NDUFA6/NDUFA7/NDUFA8/ NDUFA9/NDUFAB1/NDUFB1/ NDUFB2/NDUFB3/NDUFB4/ NDUFB5/NDUFB6/NDUFB7/ NDUFC1/NDUFC2/NDUFS1/ NDUFS2/NDUFS3/NDUFV1/ NDUFS4/NDUFS6/NDUFS8/ NDUFV2/RETSAT
M_9	Adipogenesis	0.007	0.043	GPD2/NDUFA5/ NDUFAB1/NDUFB7/NDUFS3/RETSAT
M_9	Reactive oxygen species pathway	0.011	0.046	NDUFA6/NDUFB4/NDUFS2

Table A10 - Consistency analysis results.

Module	Drug ID	Drug Name	Status	MoA	Symbol	DM up	DM down
M_122	DB00241	Butalbital	approved	potentiator	GABRA2	3	4
M_122	DB00306	Talbutal	approved	potentiator	GABRA2	3	4
M_122	DB00312	Pentobarbital	approved	potentiator	GABRA2	3	4
M_122	DB00350	Minoxidil	approved	inducer	KCNJ1	3	4
M_122	DB00371	Meprobamate	approved	agonist	GABRA2	3	4
M_122	DB00418	Secobarbital	approved	potentiator	GABRA2	3	4
M_122	DB00463	Metharbital	withdrawn	potentiator	GABRA2	3	4
M_122	DB00534	Chlormerodrin	approved	inducer	SLC12A1	3	4
M_122	DB00592	Piperazine	approved	agonist	GABRB3	3	4
M_122	DB00599	Thiopental	approved	potentiator	GABRA2	3	4
M_122	DB00794	Primidone	approved	potentiator	GABRA2	3	4
M_122	DB00818	Propofol	approved	potentiator	GABRB2	3	4
M_122	DB00818	Propofol	approved	potentiator	GABRB3	3	4
M_122	DB00837	Progabide	experimental	agonist	GABBR1	3	4
M_122	DB00849	Methylphenobarbital	approved	potentiator	GABRA2	3	4
M_122	DB00867	Ritodrine	approved	activator	KCNJ1	3	4
M_122	DB00898	Ethanol	approved	agonist	GLRA2	3	4
M_122	DB01046	Lubiprostone	approved	inducer	CLCN2	3	4
M_122	DB01198	Zopiclone	approved	potentiator	GABRA2	3	4
M_122	DB01346	Quinidine barbiturate	approved	potentiator	GABRA2	3	4
M_122	DB01351	Amobarbital	approved	potentiator	GABRA2	3	4
M_122	DB01352	Aprobarbital	experimental	potentiator	GABRA2	3	4
M_122	DB01353	Butobarbital	approved	potentiator	GABRA2	3	4
M_122	DB01354	Heptabarbital	experimental	potentiator	GABRA2	3	4
M_122	DB01355	Hexobarbital	experimental	potentiator	GABRA2	3	4
M_122	DB01483	Barbital	illicit	potentiator	GABRA2	3	4
M_122	DB01496	Dihydro-2-thioxo-5-((5-(2-(trifluoromethyl)phenyl)-2-furanyl)methyl)-4,6(1H,5H)-pyrimidinedione	experimental	potentiator	GABRA2	3	4
M_122	DB01567	Fludiazepam	experimental	potentiator	GABRA2	3	4
M_122	DB01567	Fludiazepam	experimental	potentiator	GABRB2	3	4
M_122	DB01567	Fludiazepam	experimental	potentiator	GABRB3	3	4
M_122	DB06716	Fospropofol	approved	potentiator	GABRB2	3	4
M_122	DB06716	Fospropofol	approved	potentiator	GABRB3	3	4
M_122	DB08891	Arbaclofen	investigational	agonist	GABBR1	3	4
M_122	DB08892	Arbaclofen Placarbil	investigational	agonist	GABBR1	3	4

M_122	DB13994	AZD-7325	investigational	positive allosteric modulator	GABRA2	3	4
M_122	DB16733	Rimtuzalcap	investigational	positive allosteric modulator	KCNN2	3	4
M_130	DB09092	Xanthinol	approved	cofactor	IDH3A	3	5
M_131	DB00115	Cyanocobalamin	approved	cofactor	MTR	4	5
M_131	DB00200	Hydroxocobalamin	approved	cofactor	MTR	4	5
M_131	DB03614	Mecobalamin	approved	cofactor	MTR	4	5
M_132	DB00360	Sapropterin	approved	cofactor	PAH	7	9
M_134	DB09270	Ubidecarenone	approved	cofactor	SDHA	1	4
M_137	DB00175	Pravastatin	approved	inhibitor	HMGCR	3	1
M_137	DB00227	Lovastatin	approved	inhibitor	HMGCR	3	1
M_137	DB00439	Cerivastatin	approved	inhibitor	HMGCR	3	1
M_137	DB00641	Simvastatin	approved	inhibitor	HMGCR	3	1
M_137	DB01076	Atorvastatin	approved	inhibitor	HMGCR	3	1
M_137	DB01095	Fluvastatin	approved	inhibitor	HMGCR	3	1
M_137	DB01098	Rosuvastatin	approved	inhibitor	HMGCR	3	1
M_137	DB06693	Mevastatin	experimental	inhibitor	HMGCR	3	1
M_137	DB08860	Pitavastatin	approved	inhibitor	HMGCR	3	1
M_137	DB11936	Bempedoic acid	approved	inhibitor	ACLY	3	1
M_151	DB09092	Xanthinol	approved	cofactor	OGDH	0	5
M_72	DB00893	Iron Dextran	approved	activator	HBB	1	4
M_72	DB06154	Pentaerythritol tetranitrate	approved	agonist	HBB	1	4
M_89	DB06151	Acetylcysteine	approved	stimulator	GSS	1	8
M_9	DB09270	Ubidecarenone	approved	cofactor	NDUFV3	1	3