Universität Rostock

Traditio et Innovatio

DISSERTATION

zur

zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)

Accessible Tools for Biomolecular Data Analysis using Large-Scale Knowledge Graphs

Promotionsgebiet Systembiologie und Bioinformatik
Fakultät für Informatik und Elektrotechnik

UNIVERSITÄT ROSTOCK

vorgelegt von

MATTI VAN WELZEN

geboren am 30. Juni, 1995, Schwerin, Deutschland

Betreuer: Professor Olaf Wolkenhauer
Universität Rostock

Rostock, Mai 2024

https://doi.org/10.18453/rosdok_id00005086

Gutachter: Prof. Dr. Olaf Wolkenhauer
Lehrstuhl für Systembiologie und Bioinformatik
Universität Rostock
Email: olaf.wolkenhauer@uni-rostock.de

Prof. Dr. Dirk Repsilber
School of Medical Sciences
Örebro University
Email: dirk.repsilber@oru.se

Prof. Dr. Anna Niarakis
Centre de Biologie Intégrative de Toulouse
Université Paul Sabatier
Email: anna.niaraki@univ-tlse3.fr

Verteidigung: 23. September 2024

Acknowledgements

First and foremost, I would like to thank all my colleagues in the Department of Systems Biology and Bioinformatics (SBI) for the wonderful working environment and scientific exchanges. My deep gratitude goes to Olaf Wolkenhauer for his invaluable guidance and mentorship, both during my scientific journey and personal growth. His encouragement and efforts have been crucial for the experiences and achievements throughout my journey. I would also like to express my appreciation to Shailendra Gupta for his crucial contribution to my PhD projects. His support and our many insightful discussions have not only been essential to the scientific work but have also contributed greatly to my own development. Many thanks to Markus Wolfien, who mentored me during my first years at the SBI, gave me helpful advice and sparked my interest in systems biology. Much appreciation also goes to Martin Scharm and Kristian Schultz for their technical support in software development and implementation.

I further want to show my gratitude towards the Disease Maps Community. Already at my first community meeting in 2019, which was also the first conference of my PhD, I was greeted with an excellent scientific exchange in a welcoming and open-hearted environment. Every interaction with the community continues to generate new ideas and brings great excitement and joy. Much appreciation also goes to the MINERVA team, Piotr Gawron and Marek Ostaszeweski, for their help with our projects and fantastic support with any issues I was facing, especially during the plugin development.

I acknowledge funding received by Biologische Heilmittel Heel GmbH for research projects TIRIP and NIRP that are part of this thesis. Apart from the financial support from HEEL, I would especially like to express my gratitude to our colleagues Konstantin Cesnulevicius, Myron Schultz, and David Lescheid for their great collaboration on both professional and personal levels. I am very thankful for their full support of my scientific career during our projects and the opportunities and helpful advice they provided.

My deepest gratitude is reserved for my family, whose unwavering encouragement and backing have allowed me to follow my interests and passion in life. Above all, I would like to thank my wife, Annika, from the bottom of my heart for her love, support, and patience during the many years of my studies.

Abstract

Biomolecular experiments generate data to improve our understanding of biological systems and their response to external stimuli. In the context of diseases, the goal is to use this information to modulate such systems in a desired fashion. However, data interpretation proves challenging due to the heterogeneity of data types across multiple levels of structural and functional organization.

With a substantial number of potentially significant molecules, cell types, processes, and disease phenotypes, there is a need for tools to explore data from various sources and types. To this day, there is no single all-encompassing multi-level approach to analyzing such heterogeneous biomolecular data.

In this PhD research project, I constructed knowledge graphs (KGs) to represent disease-related multi-level processes in a standardized format. I then developed tools to explore these knowledge graphs together with experimental data. The KGs are accessible through public, interactive, and community-driven platforms, referred to as "Disease Maps".

I demonstrate the approach with three Disease Maps. First, I introduce the Atlas of Inflammation Resolution (AIR) as a Disease Map of the molecular and cellular processes involved in acute inflammation and inflammation resolution. I present a novel enrichment-based analytical approach called 2DEA, which was integrated into the AIR. The approach facilitates *in silico* perturbation experiments and inferences from experimental data. I demonstrated the applicability of the AIR and the 2DEA through two studies in which I evaluated the mode of action of multi-target anti-inflammatory drugs and investigated cell type-specific gene regulation of lipid mediator synthesis. Secondly, I present the Sarcopenia Map that links molecular processes of food intake, gastrointestinal diseases, and sarcopenia through Boolean modeling. Finally, I describe the MASLD Map, which combines multi-compartmental and Boolean approaches to study spatiotemporal mechanisms in steatotic liver diseases. The results of my work have been developed in several interdisciplinary collaborations with experimental, clinical, and industry partners.

With the outcomes presented in this thesis, I provide accessible tools for biomolecular data analysis using large-scale knowledge graphs. They support the integrative analysis of diverse data types across multiple levels of functional and structural organization in biological systems.

Zusammenfassung

Biomolekulare Experimente liefern Daten, die unser Verständnis biologischer Systeme und deren Reaktion auf externe Reize verbessern. Im Kontext von Erkrankungen besteht das Ziel darin, dieses Verständnis zu nutzen, um die Systeme in gewünschter Weise zu modulieren. Die Dateninterpretation erweist sich aufgrund der Heterogenität der Datentypen auf mehreren Ebenen der strukturellen und funktionellen Organisation jedoch als schwierig. Angesichts einer beträchtlichen Anzahl potenziell signifikanter Moleküle, Zelltypen, Prozesse und Krankheitsphänotypen besteht ein Bedarf an Methoden zur Untersuchung biomedizinischer Daten aus verschiedenen Quellen und Typen. Bis heute gibt es keinen allumfassenden Ansatz zur Analyse solcher heterogenen Daten.

In diesem Forschungsprojekt habe ich *Knowledge Graphs* (KGs) erstellt, um Krankheitsprozesse auf mehreren biologischen Ebenen in einem standardisierten Format darzustellen. Anschließend habe ich Methoden entwickelt, um diese KGs zusammen mit biomolekularen Daten zu untersuchen. Die KGs sind über interaktive und öffentlich-zugängliche Plattformen veröffentlicht, die als *Disease Maps* bezeichnet werden.

Ich demonstriere den Ansatz anhand von drei *Disease Maps*. Zunächst stelle ich den *Atlas of Inflammation Resolution (AIR)* als *Disease Map* der molekularen und zellulären Prozesse in akuten Entzündungen vor. Ich habe einen neuen, auf *Enrichment* Methoden basierenden Ansatz namens 2DEA entwickelt und in den AIR integriert. Der Ansatz ermöglicht *in silico* Perturbationsexperimente und erlaubt es Rückschlüsse aus komplexen, experimentellen Daten zu ziehen. Ich habe die Anwendung des AIR und des 2DEA in zwei Studien demonstriert, in denen ich die Wirkungsweise von entzündungshemmenden Medikamenten analysiert und zusätzlich die zellspezifische Genregulation der Lipidmediatorsynthese untersucht habe. Zweitens stelle ich die *Sarcopenia Map* vor, die molekulare Prozesse der Nahrungsaufnahme, Magen-Darm-Erkrankungen und Sarkopenie durch Boolesche Modellierung verbindet. Schließlich beschreibe ich die *MASLD Map*, die multikompartimentelle und boolesche Ansätze kombiniert, um räumlich-zeitliche Mechanismen bei steatotischen Lebererkrankungen zu untersuchen.

Mit den in dieser Arbeit vorgestellten Ergebnissen stelle ich zugängliche Tools für die biomolekulare Datenanalyse mit Hilfe von großen Wissensgraphen zur Verfügung. Sie unterstützen die integrative Analyse verschiedener Datentypen auf mehreren Ebenen der funktionalen und strukturellen Organisation biologischer Systeme.

Theses

- Disease Maps as Knowledge Graphs (KG) representations of disease mechanisms support inflammation research through context-specific knowledge exploration and data visualization.
- 2DEA enhances the informative value of enrichment analyses by combining experimental data with information extracted from KGs.
- Topological analysis of KGs supports process-specific clustering in unsupervised machine learning from single-cell data to identify cell type-specific gene regulation.
- Immune process-specific KGs and the 2DEA improve the inference of drug mechanisms from experimental data of medical products.
- Boolean models of modularized KGs allow the simulation and prediction of potential mechanisms underlying systemic disease processes.
- The multi-compartmental design of logic KG models enables the simulation of spatiotemporal processes.

Contents

List of Figures

List of Tables

List of Abbreviations

2DEA	Two-Dimensional Enrichment Analysis
ABM	Agent-Based Model
aCaBs	Antibiotic Selected Cardiac Bodies
AF	Activity Flow
AIR	Atlas Of Inflammation Resolution
CRN	Core Regulatory Network
DAMP	Damage-Associated Molecular Pattern
DCE	Differentially Changed Element
DEG	Differentially Expressed Gene
DMC	Disease Map Community
EDA	Exploratory Data Analysis
ES	Enrichment Score
FC	log2 Fold Change
GO	Gene Ontology
GRN	Gene Regulatory Network
GSEA	Gene Set Enrichment Analysis
ID	Intestinal Dysfunctions
iSABs	Induced Sinoatrial Bodies
KG	Knowledge Graph
KGML	Kegg Markup Language
LC	Liver Cirrhosis
LM	Lipid Mediator
MASLD	Metabolic Dysfunction-Associated Steatotic Liver Disease
MASH	Metabolic Dysfunction-Associated Steatohepatitis
NSAID	Non-Steroidal Anti-Inflammatory Drug
ODE	Ordinary Differential Equation
ORA	Overrepresentation Analysis
PAMP	Pathogen-Associated Molecular Pattern
PCA	Principal Component Analysis
PD	Process Description
PIM	Pro-Inflammatory Mediator
PPI	Protein Protein Interaction
ROS	Reactive Oxygen Species
Rt	Retention Time
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SLD	Steatotic Liver Disease
SPM	Specialized Pro-Resolving Mediator
TF	Transcription Factor
UI	User Interface
VEP	Variant Effect Predictor

Thesis Outline

Chapter 1 addresses the existing challenges in biomedical data analysis and introduces the reader to the principle of modeling biological systems. It introduces knowledge graphs as graph-structured models, particularly Disease Maps as publicly available and community-driven knowledge graph resources for diseases. The chapter provides a foundation for the thesis by defining the terminology and giving an overview of existing approaches, software, and resources. Ideas presented in this chapter were discussed in a review article Hoch *et al.*, 2024 [1].

Chapter 2 discusses the application of systems biology approaches to the field of inflammation to investigate multi-level processes and perform context-specific data visualization. The chapter introduces the "Atlas of Inflammation Resolution" (AIR) Disease Map in terms of modeling, curation, community building, and development of tools for map exploration. The AIR is published in Serhan & Gupta *et al.*, 2020 [2].

Chapter 3 introduces the two-dimensional enrichment analysis (2DEA) as a novel enrichment-based approach to perform data analysis on large-scale KGs. It describes the underlying methodology and compares the 2DEA to established approaches, such as GSEA. The chapter additionally depicts the development of Disease Map tools that employ the 2DEA for data integration and knowledge derivation. The methodology is published in Hoch *et al.*, 2022 [3].

Chapter 4 presents a methodology for investigating cell type-specific gene regulation using knowledge graph approaches. It describes the application to a single-cell RNA-Seq dataset of immune cell types, analyzing their specific lipid mediator synthesis using subgraphs from the AIR. The methods and results described in this chapter are published in Hoch *et al.*, 2023a [4].

Chapter 5 covers an industry project in which I investigated a multi-component natural product and compared the pharmacological principles of single and multi-target treatment. It describes how the 2DEA and other KG approaches were employed to infer drug mechanisms from RNA-Seq and interactome data. The results of the chapter are published in Hoch *et al.*, 2023b [5].

Chapter 6 presents the "Sarcopenia Map", a Disease Map created in collaboration with the gastroenterology department of the University Rostock Medical Center that links nutrition, gastrointestinal diseases, and sarcopenia. It describes creating a Boolean model from the Disease Map's KG to simulate the effects of dietary changes and disease disorders on muscle growth and function. The Sarcopenia Map is published in Hoch & Ehlers *et al.*, 2022 [6].

Chapter 7 first discusses the challenges of spatial modeling in the context of liver disease. It presents an approach that combines agent-based and Boolean modeling methods to investigate spatial disease mechanisms in metabolic-associated steatotic liver disease (MASLD). The chapter discusses the development and curation of the MASLD Map, which transforms this approach into a publicly available tool for studying processes in MASLD pathology.

The chapters may contain verbatim references from passages of their respective publications.

Chapter 1

Introduction

1.1 The Systems Biology Approach

Biomolecular research has evolved considerably, from measuring single molecules to generating data on single cells and whole organisms [7]. A key motivation behind this shift is the desire to understand not just isolated molecular processes but how these processes interact within larger biological systems. A **biological system** is understood as any association of biological entities, from molecules to organisms, functioning together to sustain (biological) functions in a self-organized manner [8]. The mentioned shift is supported by the emergence of omics technologies that quantify molecules of whole systems, enabling studies on much larger scales [9]. This capability allows researchers to compare the system's responses to stimuli, monitor changes over time, or observe its behavior under different conditions. These systems and the data generated from them are characterized by a high degree of heterogeneity at several levels of structural and functional organization [10]. Consequently, given the inherent complexity of even simple biological systems, no single experiment can capture their full extent. Research is therefore progressing incrementally, with each experiment contributing to a larger, more comprehensive picture (Figure 1). This iterative process is supported by statistical analysis to ensure observed differences are meaningful and not due to chance, requiring careful experimental design and sufficient sample sizes to avoid biases.

Even small systems, such as of few molecular reactions, often exhibit complex non-linear behavior [11], [12]. While the interpretation of these processes might appear to be relatively straightforward, feedback mechanisms, threshold effects, and steady-state behavior emerge as phenomena that can not be inferred from the behavior of the individual components alone. Additionally, the complexity of systems scales with size, amplifying the complexity and non-linearity and leading to amounts of information that exceed human capabilities. Consequently, interpreting biomolecular data requires a thorough understanding of the experimental context and the biological system under investigation. While new experimental methods such as omics technologies are effective in generating new insights, they can exacerbate this challenge due to the huge amounts of data produced in a very short time.

Computational methods have become indispensable in addressing the complexity of the problem. They facilitate the processing of vast amounts of data and the comparison of new

findings with existing information. This process has led to the emergence of **systems biology**, a field dedicated to understanding biological systems through mathematical and computational approaches [13]. They help to uncover hidden patterns and offer deeper insights into the functioning of biological systems through simplified, more comprehensible, and more tangible representations, known as **models**.

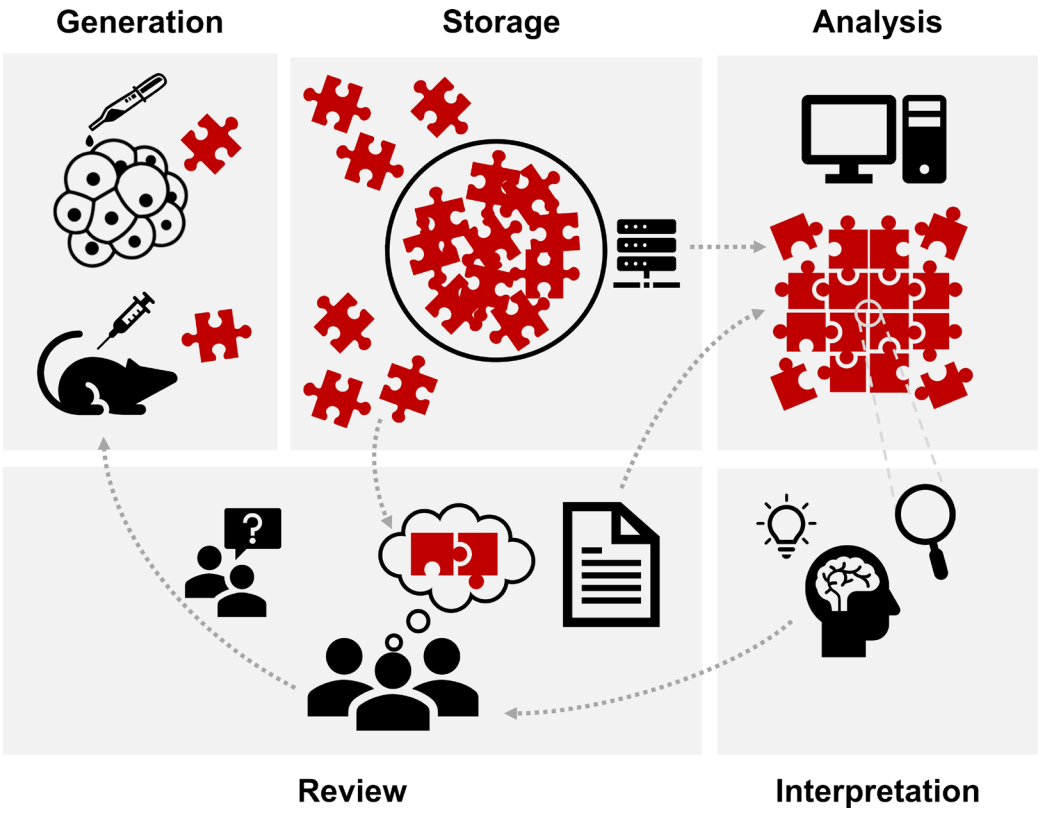

Figure 1: Conceptual workflow of knowledge inference from experimental data.

1.1.1 Modeling in Biology

Modeling seeks to create a reduced representation of reality, a process referred to as **abstraction**, by using minimal information from selectively simplifying or reinterpreting real-world phenomena to achieve as accurate predictions and insights as possible (Figure 2). This information reduction is crucial for maintaining independence from data availability and for reducing computational costs [14].

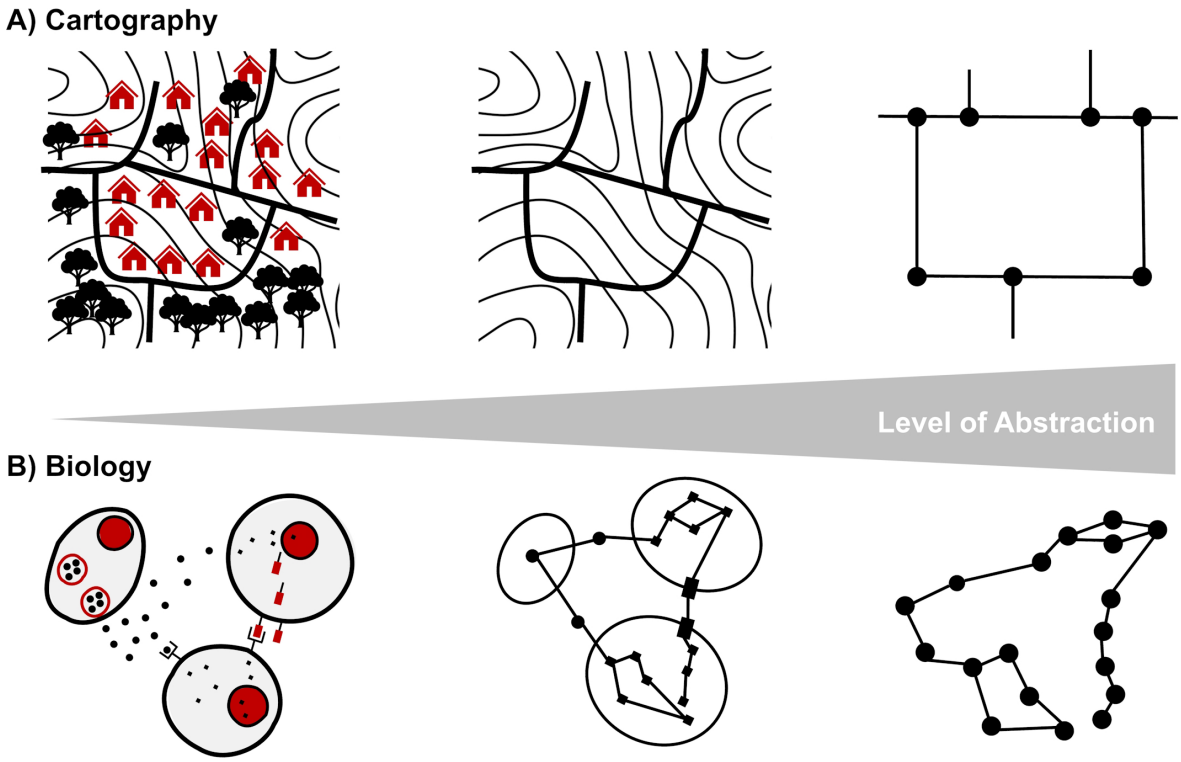

Figure 2: Concepts of abstraction in modeling. Comparison of landscape models in cartography to extract relevant information of interest, such as for pathfinding (A), with modeling biological systems as graph representations (B). Adapted from Hoch *et al.* 2024.

The functions of biological systems are orchestrated by a continuous exchange of matter and energy between multiple levels of subsystems, such as organs, tissues, cells, and cellular organelles. This exchange is mediated and tightly controlled through physicochemical interactions of molecules. These molecular interactions are highly complex, influenced by factors such as the three-dimensional structure of the molecules, their movement in extracellular and intracellular fluids, and the physical conditions of the environment, among others [15]. Even if all the necessary information were available, a detailed computational representation of a single cell would probably exceed the computing capacity of the next few decades, not to mention multicellular tissues with billions of cells, many unique in function and structure [16]. Biological models abstract these processes to allow feasible computations.

Abstraction can be performed on multiple scales, either functional (from molecular signaling pathways to clinical symptoms), temporal (from microseconds of molecular signaling to days or weeks in a clinical context), or spatial (from single cells to whole tissues). On these scales, the level of detail, referred to as **granularity** or **resolution**, increases from the coarse-grained (low-resolution or higher level) representations at the top to the fine-grained (high-resolution) at the bottom. In informational or horizontal abstraction, details within the same granularity level might be omitted when such details have less relevance to the research question. When modeling metabolic processes, for example, the reaction chains of interest can

be decoupled from other processes, which, in reality, form a highly interconnected system. On the other side, conceptual or vertical abstraction is based on the idea that emergent properties at higher levels can often be modeled more feasible without accounting for the complexity at the lower level. In the context of the "law of large numbers," the mean value of the sample approaches the mean value of the population as the sample size increases [17]. Translated to biology, an enzymatic rate equation determined by experimental measurements over minutes at the cellular level aggregates potentially millions of enzyme molecules. The physicochemical properties of the nanoscale can be ignored because their collective behavior averages individual variances and noise. Similarly, describing cellular movements does not require knowledge of molecular events, and social models on the level of individual patients neglect the cellular processes in each patient.

Historically, systems biology has been applied to very specific research questions, mostly focused on mechanisms of molecular interactions, aiming to identify the functions of particular molecules. As those require a high degree of granularity, such approaches are associated with computational limitations and high curation efforts, restricting models to small scales. However, with the increasing availability of omics data and the application of computational approaches in clinical research, large-scale solutions have become of great interest [18], [19]. Modeling diseases usually requires understanding the communication processes between cells and tissues or even at a systemic level rather than highly specific molecular mechanisms. Consequently, data analysis often does not have a fixed hypothesis and predefined results, a process commonly referred to as **exploratory data analysis** (EDA). Experiments generate comprehensive and often heterogeneous datasets such as genomics, proteomics, and metabolomics, and analyses are then performed to uncover new patterns, trends, or relationships, thus developing new hypotheses.

In response to the shift from small-scale to such large-scale approaches, many efforts aim to combine models of different scales or resolutions, commonly referred to as multi-scale modeling [20]. The challenge lies in defining the coupling between different levels of abstraction [21], [22], [23]. Although the molecular mechanisms that trigger reactions at the tissue level can be described, the physical processes, such as the organization of microfilaments and 3D conformations, are of near-infinite complexity. The differences in temporal scales introduce an even more significant issue. Molecular models designed to run at the millisecond scale would need to be run for minutes up to hours, exponentially magnifying any uncertainties. Nevertheless, multi-scale models show great potential for translating systems biology into clinical practice, enabling the investigation of molecular

perturbations on phenotypic levels. Multiple approaches to their design have already been developed, differing in the investigated biological systems, level of abstraction, spatial or temporal scales, and their deterministic or stochastic nature [20].

1.1.2 Knowledge Graphs

The events underlying the function of biological systems can be modeled through graph-structured representations of the relationships between their entities (Figure 2B) [24]. The resulting structures are referred to as **networks** or **graphs**, with both terms often being used interchangeably. However, networks are more commonly used to describe concepts in specific applications such as social sciences, information technology, or biology, while graphs refer to the mathematical representation that can be analyzed computationally using graph-theoretic approaches (Figure 3A) [25], [26], [27]. Iñiguez *et al*. described graph theory as being "focused on providing rigorous proofs for graph properties" and network science as "more akin to phenomenological physics […] with the goal of gaining intuition of their underlying generative mechanisms" [27]. The term **knowledge graphs (KG)** is used to describe graphs with a focus on the evidence-based semantics and relationships between entities, especially when these are of different resolutions and scales [28], [29]. In this thesis, KG is the preferred terminology to do justice to the graph-theoretical perspective and the heterogeneity of biological systems. Nevertheless, both terms, network, and graph, are treated as interchangeable, as established methods and tools are using both. Experience dealing with people from different backgrounds has shown that one should not commit to a specific terminology but adapt to the audience and the means of communication.

Depending on whether a KG is created from newly generated data or prior knowledge, one can distinguish between reverse and forward modeling, respectively [30]. An example of reverse or data-driven modeling is the creation of KGs from scratch using purely new experimental data, a process referred to as network inference [31]. The research question is the inference of causalities between the measured molecules. A prominent example is the analysis of co-expression from transcriptome data to generate KGs of interactions between TFs and gene targets, referred to as gene regulatory networks (GRNs) [32]. Forward or knowledge-driven modeling uses prior knowledge to construct models, which is why, in these approaches, KGs are sometimes referred to as prior knowledge networks (PKNs) [33].

In a KG, mathematically denoted as G, the entities at any level of biological granularity (e.g., cells or molecules) can be represented as **nodes** (or vertices, $V(G)$). A node represents all copies of an entity, with properties that uniquely identify it depending on the context and

model granularity. The interactions, either of conceptual, physicochemical, or functional nature, between two nodes are depicted as **edges** $E(G)$ defined as:

$$E(G) = \{(u, v) | u, v \in V\} \tag{1.1}$$

The graph can be further parameterized to include additional details about the characteristics of the nodes and edges, which vary depending on the context and the availability of data. Properties of a node can, for example, include qualitative (**states**) and quantitative (**concentrations**) parameters. Additionally, the direction and sign of edges are crucial in many modeling approaches, as they represent **causality**, indicating processes such as activation, deactivation, upregulation, or downregulation between the nodes. In such case, an edge $e \in E(G)$ can associated with a type τ_e, which could be positive (1) or negative (-1). The type $\tau(u, v)$ thus describes the causality between a source u und a target v, changing the definition of the edges to:

$$E_{dir}(G) = \{(u, \tau_e, v) | u, v \in V\} \tag{1.2}$$

As all KGs described in this thesis are directed, E is also used for E_{dir} for the sake of simplicity. The adjacent nodes or **neighbors** $N(v)$ of a node v that are connected to v either as a source or as a target are defined as:

$$N(v) = \{u \in V \mid (u, v) \in E \vee (v, u) \in E\} \tag{1.3}$$

A path P in a graph G, denoted as $P \in \mathcal{P}(v_0, v_n)$, is a sequence of vertices $P = (v_0, v_1, \dots, v_n)$ such that each consecutive pair of vertices (v_i, v_{i+1}) forms an edge in G, i.e., $(v_i, v_{i+1}) \in E$ for all i from 0 to $n - 1$. Consequently, a path P can be written as a sequence of edges connecting v_0 and v_n as $P = (e_1, e_2, \dots, e_n)$. The length of the path P, denoted as $\ell(P)$, is the number of edges in P, n in this case, where $n \in \mathbb{N}$. In a directed graph, the type of a path P is defined as $\tau(P) = \prod_{i=1}^{\ell(P)} \tau_e(e_i)$. The length of a path thus equals the number of edges in the path. The shortest path $\sigma(u, v)$ between two nodes $u, v \in V$ is defined as an existing path between u and v with minimized length.

$$\sigma(u, v) = \underset{P \in \mathcal{P}(u,v)}{arg\,min}\, \ell(P) \tag{1.4}$$

1.1.3 Modularization of Biological Systems

Exploring biological systems has historically centered on identifying key molecules presumed to be drivers of higher-level processes. However, the role of these molecules as primary drivers is sometimes questioned, as they are potentially influenced by study biases [34]. Similarly, modularity in graph theory, defined as subgraphs with an overrepresented number of edges, might be influenced by study bias such that more studied nodes show high

degree centralities (see Section 1.4.2) [35]. The gradual accumulation and reinterpretation of data have divided biological systems into apparent modules commonly referred to as **pathways** in the context of biomolecular interaction**s**. These pathways, e.g., metabolic pathways, signaling cascades, or structural complexes, are considered as discrete modules performing specific functions within the cell or organism (see Section 1.1.4). In KGs, pathways are represented as **subgraphs** (Figure 3B). The higher-level processes described by a pathway are usually integrated into KGs as single-node representations, often referred to as **phenotype** nodes, with edges connecting them to lower-level node representations (i.e., molecules). Nodes within the pathway with a path toward a phenotype are referred to as its **modulators**. The set of all phenotypes in a KG will be denoted as $V_p(G) \subset V(G)$ and the set of phenotypes modulated by a node u defined as $N_p(u) = \{v \in V_p(G) \mid \mathcal{P}(u, v) \neq \emptyset\}$.

While modularization allows researchers to dissect and study biological processes in a more manageable way, it can obscure the inherent interconnectedness of biological systems. For example, while each pathway is traditionally viewed as a distinct module in metabolic pathways, they are interconnected through shared metabolites and regulatory mechanisms. This interconnectedness means that changes in one module can have cascading effects on others, significantly impacting our understanding of diseases and the development of treatments. In the context of disease, targeting a specific module might have unforeseen consequences due to the interconnected nature of the system. This understanding leads to more holistic approaches in drug development, focusing on understanding and targeting large-scale interactions rather than isolated modules [18]. Computational methods help overcome these limitations by combining subgraphs of individual pathways with pathway-unspecific KGs from public databases to create large-scale comprehensive KGs (Figure 3B). Modularization can refer to a functional organization of KGs and describe the spatial organization of biological systems, i.e., cellular organelles, cells, or tissues. Modules in a spatial context are often referred to as **compartments**.

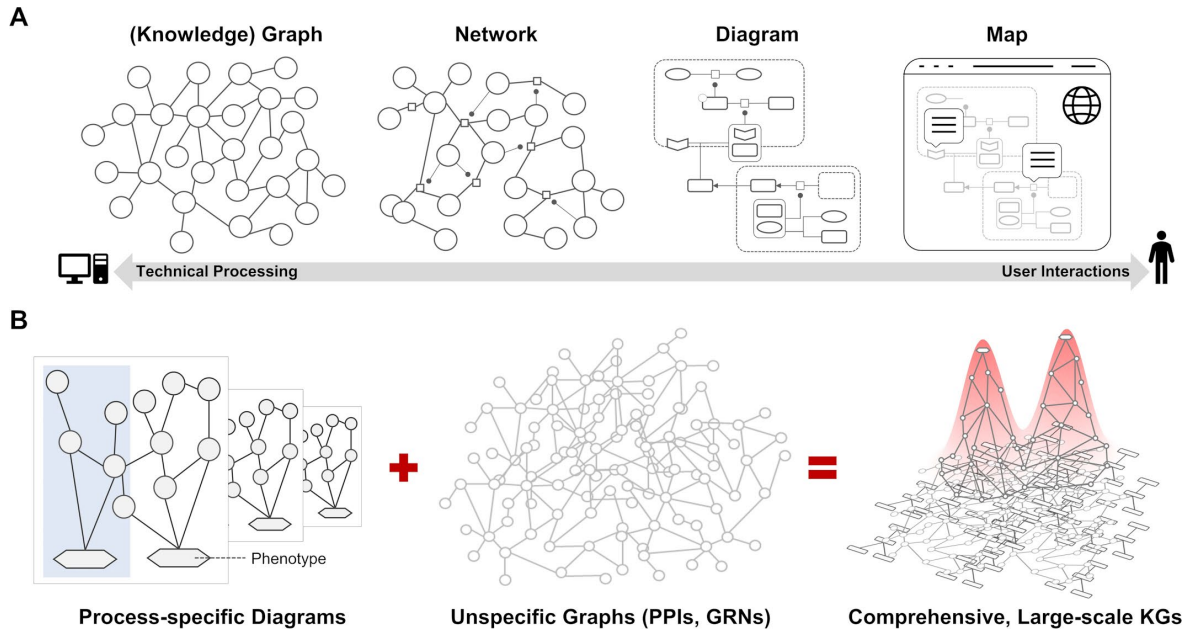

Figure 3: Concepts of knowledge graph (KG) designs in computational models of biological systems. (A) While "network" is a general term for graph-structured models, "graph" is usually reserved for structures mathematically analyzable using graph theory. Standardizing networks, e.g., through the Systems Biology Mark-up Language (SBML), enables reproducible and visually appealing representations in "diagrams." "Map" is a term the Disease Map community employs to describe web-accessible and interactive diagram presentations. **(B)** Curation of large-scale KGs by combining manually curated KGs of specific processes with unspecific KGs generated from large databases. In large-scale KGs, pathways are not treated as isolated subgraphs but as integrated components modulated by collective signals from the underlying KGs. PPIs - Protein-Protein-Interactions; GRNs - Gene Regulatory Networks. Adapted from Hoch *et al*. 2024.

1.1.4 Signal Transduction in Biological Systems

Directionality in KGs describes how information is conveyed throughout the system. Biological systems are characterized by high molecular density, where chemical reactions predominantly occur between directly adjacent molecules. Such reactions, facilitated through various mechanisms like chemical modifications or transport across compartments, alter the states or concentrations of molecules, thereby modifying their functions. Due to the differences in the type of the molecules and their interactions, pathways of biological systems are usually divided into three major types, which require different modeling approaches and data [36].

(i) **Metabolic models** describe catalytic reactions in which small molecules, i.e., metabolites, are processed to produce energy, nutrients, mediators, and structural molecules for cellular homeostasis. Concentrations of metabolites and their changes on a large scale can be quantified through metabolomics.

(ii) **Gene regulation models** refer to how specialized proteins, called transcription factors (TFs), bind to DNA and modulate the accessibility of the transcription apparatus. In this way, TFs can adapt the expression of genes and, subsequently, protein levels to cellular requirements. Gene expression is assessed using

transcriptomics, which measures the quantities of gene products such as messenger RNA (mRNA) or microRNA (miRNA).

(iii) **Signaling models** describe how the transmission of information is mediated by the activity of proteins and their modulation through chemical modifications, usually phosphorylations. A cascade design in which one protein modifies several others, e.g., after receptor stimulation, amplifies the signals. The endpoints of signaling pathways can be the modulation of enzymes in metabolic reactions or TFs in gene regulation. Conclusions on the signaling processes can be drawn from proteomics, which measures the abundance of proteins and their modified states.

Approaches focusing on limited aspects or parts of systems may only consider one of these model types. However, due to the heterogeneity of biological systems, their KG representations, and the generated data, the research question might require exceeding their boundaries. For example, one might want to draw conclusions from perturbations in protein signaling about the effects on transcriptional regulation or link changes in gene expression to metabolic effects. An example is the close coupling of the carbohydrate system with protein signaling and gene expression through proteins such as AMPK [37]. In biological systems, any perceived effect, be it a change in concentration, localization, or alteration, can trigger new signals in different signaling pathways as a kind of feedback control that induces new responses. Consequently, a clear distinction between the cause and effect of signals, especially between the different signaling pathways, often proves difficult. Under these considerations, the term **signal** in the following refers to any change in the abundance or properties of biomolecular entities that convey information within the system, triggering new signals by itself.

The sequence of successively inducing signals is referred to as **signal transduction**, **signal flow,** or **signaling**. In KGs, signals are represented as numerical properties assigned to nodes or edges. In quantitative models, these values typically represent actual molecular concentrations, while in qualitative contexts, they might be more arbitrary, e.g., representing the nodes' **states** on ordinal scales. The signal flow in KGs can be defined by mathematical functions that describe how the signal value of a node is calculated based on the signals of other nodes. In quantitative models, the change in signal value is calculated over an infinitesimally small-time interval, allowing for a continuous and dynamic representation of signal flow, e.g., concentration changes over time ($\frac{dC}{dt}$). These changes can be solved mathematically using ordinary differential equations (ODEs) [38]. Qualitative models usually employ an incremental approach, updating the signal or state in discrete steps. This method

involves iteratively recalculating the signal values based on a set of rules or logical conditions, reflecting the state changes of the nodes over distinct time intervals. While the signal transduction of specific pathways can be modeled in great detail, a unified approach must be inherently more non-specific to account for the heterogeneity of type-specific mechanisms. As a solution, many approaches use a multi-layered design where the specifications and details of each KG model are maintained in one layer, thus emphasizing the connectivity between the layers [39], [40]. The multi-layer structure is particularly useful for multi-scale modeling, where the layers describe different scales and, therefore, require very different approaches.

1.1.5 Knowledge-Driven Data Analysis

In biomolecular experiments, one cannot directly measure actual molecular signaling, at least not on a larger scale, but only the system's response in a given state (further discussed in Section 1.3.3). Consequently, the understanding of biological systems is based on combining data from such "snapshots" with prior knowledge of already identified mechanisms described in KGs. The interpretation of data values as input signals for the KG and the subsequent calculation of the response signals of other nodes in the KG, a process referred to as **simulation**, enables **predictions** to be made. In this context, 'predictions' are understood as the projections of the data along spatial or temporal scales or onto different biological levels, thereby identifying patterns that infer relationships or hint at underlying mechanisms. Barsi and Szali (2021) referred to this process as “causal reasoning” [20], distinguishing it from pure knowledge- or data-driven models. Dugourd & Saez-Rodriguez described such an analysis as "footprint-based," in which empirical data are regarded as the footprint of a biological process, whose functioning can thus be inferred [41].

Given a directed KG, simulations can be performed in both directions: (i) simulating backward or **upstream** by backtracking incoming edges to predict potential underlying mechanisms or contributing factors (**causes**) in the past, or (ii) simulating forward or **downstream**, following outgoing edges to predict future consequences (**effects**). Considering the ideas from Section 1.1.1, higher-level predictions, such as the impact of molecular changes on cellular phenotypes, benefit from aggregation and generalization, making them more feasible. In contrast, making forward predictions into more fine-grained levels is often more challenging due to the inherent complexity and variability at these detailed levels. Conversely, predicting causes tends to be more straightforward at lower, more detailed levels, where specific interactions and relationships can be more directly observed and analyzed.

In essence, the computer-aided data analysis using large-scale KGs is comparable to the analytical thought process of a researcher conducting molecular experiments (schematically visualized in Figure 4). Drawing associations between disturbances and their observed effects can be straightforward in small-scale systems, such as individual pathways. Considering an example where the activity of an enzyme in a metabolic pathway, such as glycolysis, is increased, the direct outcome, a rise in the concentration of the products, is relatively easy to predict. Even before conducting the experiment, the researcher already had an idea of the outcomes in their head by connecting the type of perturbation with subsequent enzymatic steps, i.e., combining the input data with prior knowledge. Computational KG approaches follow the same principle but on a larger scale for systems whose intricate relationships become incredibly complex and exceed the limits of human intuition. For example, it is much less intuitive to determine the broader implications of increased glycolysis, such as how it might influence other metabolic pathways or even affect cellular activities and intercellular communication within a tissue.

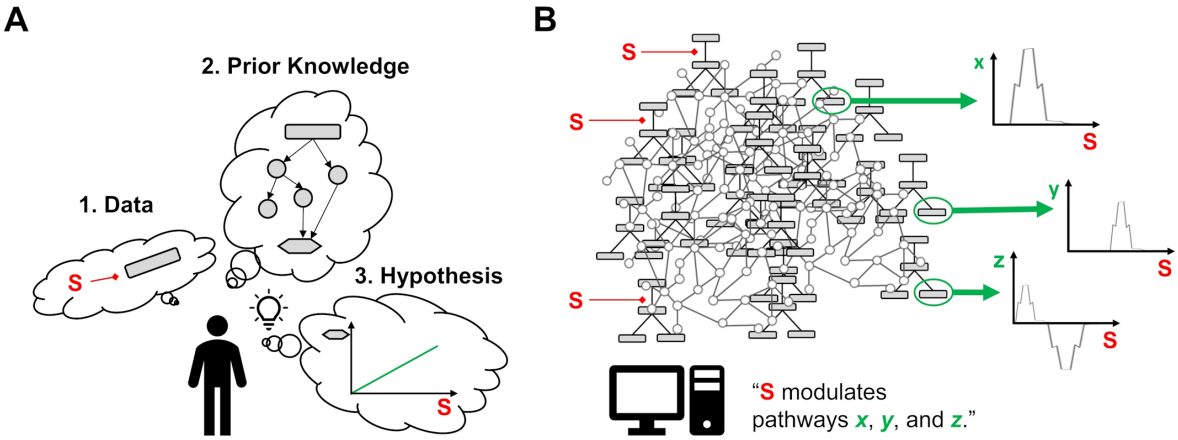

Figure 4: Illustrative example of how computer-aided approaches support the analysis of biomolecular data. Linking prior knowledge with data from biomolecular experiments is the basis of data analysis. With these associations in mind, researchers conducting biomolecular experiments have a certain prior expectation of the results and can ultimately interpret them. Computer-aided knowledge graph analysis enables the integration and analysis of large-scale data that supports understanding non-linear processes.

1.2 Knowledge Graph Resources

Figure 5 shows an exemplary workflow for the knowledge-based analysis of biomolecular data. The first step is the curation of a KG based on prior knowledge, e.g., from scientific literature or databases. This section summarizes the different resources, databases, and standards that are being utilized to develop comprehensive KGs. The next step is the integration of empirical or hypothetical data into KGs, which will be discussed in Section 1.3. Finally, KG-driven simulations (as described in Section 1.1.5) enable predictions to be made

from the integrated data. Different analysis approaches are reviewed in Section 1.4. The design of the KG substantially influences the scope of knowledge that can be inferred from it. Therefore, the process of KG curation should be guided by a clear vision of the analytical objectives. Their design should reflect the biological scale and resolution that best serves these goals.

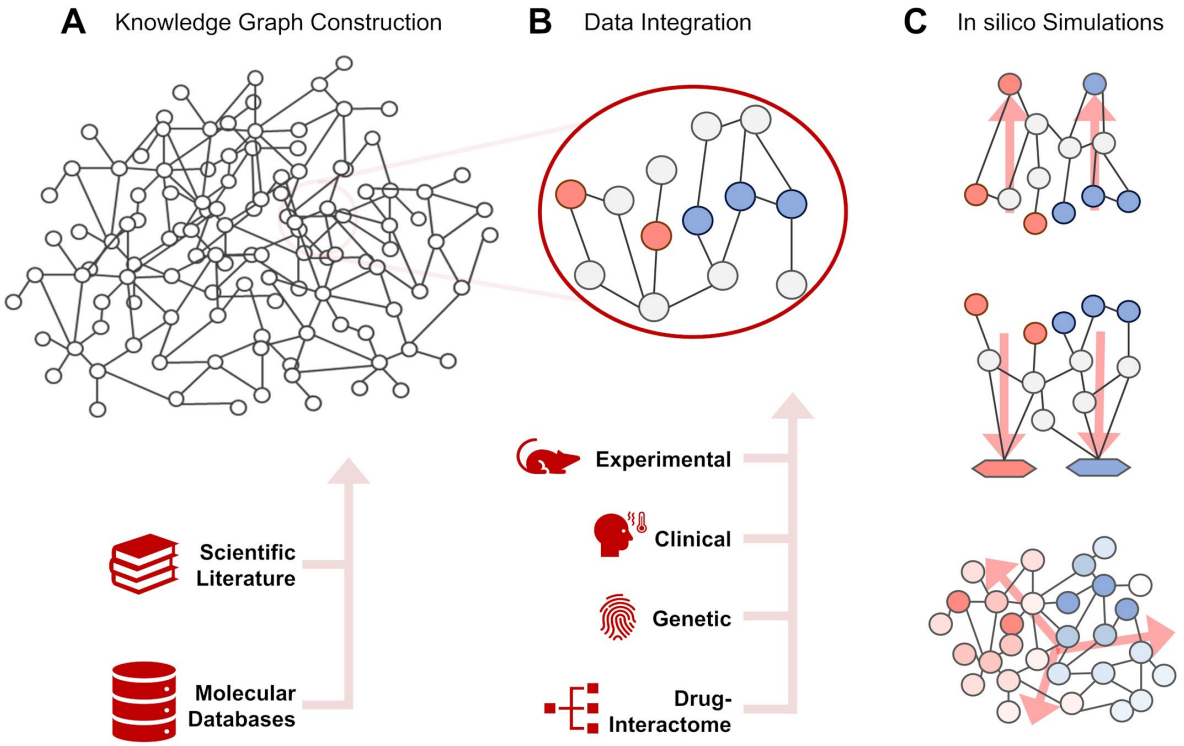

Figure 5: An exemplary systems biology workflow using knowledge graphs (KG) for biomolecular data analysis.

1.2.1 Interaction Databases

Databases play a crucial role in biomolecular data science, offering a wealth of pre-existing knowledge about molecular interactions. The information within these databases can be extracted from various sources, including scientific literature and resources with experimental evidence. However, the information's scale, detail, and confidence can vary drastically between databases. Some databases are primarily based on validated experimental data. They may focus on direct physical interactions between molecules, which are validated by techniques such as crystallography. Other databases use text-mining approaches to extract information from scientific literature. These tend to have a higher level of uncertainty, as they can misinterpret the syntax and context in which molecules are mentioned. Due to their low specificity, they are primarily used when knowledge quantity is prioritized over quality. The STRING database is an example of a rather unspecific protein-protein interaction (PPI) database [42]. STRING annotates interactions with a confidence score and allows users to filter

information based on evidence. Generally, these large-scale resources are functionally unspecific, being guided mostly by the source of their data and, consequently, the molecular types of entries. Many databases, for example, focus on describing the regulation of gene targets by TFs (e.g., TRRUST [43]), miRNAs (e.g., mirTarBase [44], miRDB [45]), or long non-coding RNAs (lncRNAs, e.g., LncRNA2Target [46], LncTarD [47]). Specifically, KGs of TF-gene interactions are referred to as gene regulatory networks (GRNS). Some databases, such as BioGrid [42] or STRING [36], also provide heterogeneous information. However, such large databases are often subject to redundancies and uncertainties in the data [48], [49], [50], [51]. In the projects presented in this thesis, I specifically use such large databases to enhance process-specific KGs with regulatory information, as described above in Section 1.1.3.

In addition to the interactions between two molecules, some resources aim to capture associations between biological levels. Examples of such a database type are the so-called ontologies, which classify higher-level processes such as biological functions or diseases as standardized terms. The **Gene Ontology** (GO) database defines biological processes and their connections in a hierarchical, tree-like structure [52]. In contrast, the Human Phenotype Ontology (HPO) focuses on pathological phenotypes in human disease and associated genetic variants [53]. The identifiers from the ontology databases are typically used to annotate higher-level nodes in KGs. In addition, ontology databases curate the genes for which there is experimental evidence that they are involved in the processes, all referred to as the gene set of the ontology term.

Given this wide variety of resources, tools have been developed to provide a framework for accessing and connecting information from multiple databases, such as OmniPath [54], which provides Python and R packages for seamless integration into data analysis workflows.

1.2.2 Knowledge Graph Designs

KGs can be curated with varying degrees of detail, primarily distinguished between Activity Flow (AF) and Process Description (PD) [55]. Both terms originate from the Systems Biology Graphical Notation (SBGN) standard [56] (see next section) but have since been adopted generally for KG approaches. AF represents general causality between two nodes, often with high abstraction [57]. It is thus equivalent to the format of directed KGs processable by graph theory in which edges are minimally defined by a source, target, and type (as defined in Section 1.1.2). In contrast, PD was developed to describe reactions as direct physicochemical interactions between the participating nodes [58]. The target nodes are the product of a

physicochemical modification or transport of the source nodes, with a modifier node being the converting enzyme or transporter. The PD format can thus describe biomolecular interactions in great detail. In PD, edges can include multiple nodes as sources and targets. With $\mathbf{S}_u(e) \subseteq V$ and $\mathbf{S}_v(e) \subseteq V$ denoting source and target node sets of an edge e, respectively Therefore, an edge $e \in E_{PD}$ in PD format is described as:

$$E_{PD}(G) = \{(\boldsymbol{S}_u, \tau_e, \boldsymbol{S}_v) \mid \boldsymbol{S}_u, \boldsymbol{S}_v \subseteq V\} \tag{1.5}$$

Furthermore, PD integrates mechanistic information, describing edges with a set of **modifications** M where each is defined by a set of **modifier** nodes $\boldsymbol{S}_m$ and a modification type $\tau_m \in \{-1,1\}$ mapped from the definition of the modification in the PD KG standard. Every modification is specific to an edge, and thus, the set of all modifications in a graph G is defined as $M(G) = \{(\boldsymbol{S}_{\mathrm{m}}, \tau_m, e) | \boldsymbol{S}_{\mathrm{m}} \subseteq V, \tau_m \in \{-1,1\}, e \in E_{PD}\}$. The specifications of the PD format result in the following new definitions:

- The set of modifications that modify an edge e is defined as:

$$M_e(e) = \{(\boldsymbol{S}_m, \tau_m) | (\boldsymbol{S}_m, \tau_m, e) \in M(G)\} \tag{1.6}$$

- The set of modifications $M_v(v)$ that modify all edges in which v is a target is defined as:

$$M_v(v) = \bigcup_{e \in E_{PD} : v \in \boldsymbol{S}_v(e)} M_e(e) \tag{1.7}$$

- The set of modifiers $N_m(v)$ that modify any edges in which v is a target is defined as:

$$N_m(v) = \bigcup_{(\boldsymbol{S}_m, \tau_m) \in M_v(v)} \boldsymbol{S}_m \tag{1.8}$$

- The set of nodes $N_s(v)$ that are source nodes in edges where a node v is a target is defined as:

$$N_s(v) = \bigcup_{e \in E_{PD} : v \in \boldsymbol{S}_v(e)} \boldsymbol{S}_u(e) \tag{1.9}$$

- The incoming neighbors $N_{in}(v)$ of a node v is defined as the set of modifier nodes and source nodes of edges where v is a target.

$$N_{in}(v) = N_m(v) \cup N_s(v) \tag{1.10}$$

- The outgoing neighbors $N_{out}(v)$ of a node v is defined as the set of target nodes of edges where v is a source.

$$N_{out}(v) = N_t(v) = \bigcup_{e \in E_{PD}: v \in \boldsymbol{S}_u(e)} \boldsymbol{S}_v(e) \tag{1.11}$$

- the neighbors $N(v)$ of a node v is defined as the combined set of incoming neighbors $N_{in}(v)$ and outgoing neighbors $N_{out}(v)$ of v.

$$N(v) = N_{in}(v) \cup N_{out}(v) \tag{1.12}$$

Given the varying levels of detail, PD is backward compatible with AF, as it involves abstracting detailed reactions into a simpler format [59]. Conversely, converting AF to PD requires additional effort to add mechanistic details, i.e., integrating modifications as direct edges. Converting PD to AF often results in a loss of information, and there is no universally defined method for this transformation. Given a KG $G_{PD} = (V_{PD}, E_{PD})$ in PD format, the transformed KG in AF format is defined as $G_{AF} = (V_{AF}, E_{AF})$ with $V_{AF} = V_{PD}$ and

$$E_{AF} = \bigcup_{e \in E_{PD}} f_{PD \to AF}(e) \tag{1.13}$$

$$f_{PD \to AF}(e) = \begin{cases} \{\{(u, \tau_e, v) | u \in \boldsymbol{S}_u, v \in \boldsymbol{S}_v\}\} & \text{if } M(e) = \emptyset \\ f_{mod \to AF}(e) & \text{otherwise} \end{cases} \tag{1.14}$$

A common strategy is to integrate modifier nodes by adding a new edge to each target and customize the edge type according to the type of modification:

$$f_{mod \to AF}(e) = \bigcup_{(\boldsymbol{S_m}, \tau_m) \in M(e)} \{(u, \tau_e(e), v), (m, \tau_m \cdot \tau_e(e), v) |\, u \in \boldsymbol{S}_u(e), v \in \boldsymbol{S}_v(e), m \in \boldsymbol{S}_m\} \tag{1.15}$$

In catalytic reactions, i.e., where the modifier is essential for the reaction to occur, modifiers are sometimes integrated as intermediate nodes between each pair of source and target nodes. In addition, especially in metabolic reactions, a negative feedback loop is added from the modifier to the source to represent the consumption of the substrate by the reaction. Under these aspects, the definition changes to:

$$f_{mod \to AF}(e) = \bigcup_{(\boldsymbol{S_m}, \tau_m) \in M(e)} \{(u, 1, m), (m, -1, u), (m, \tau_m \cdot \tau_e(e), v) |\, u \in \boldsymbol{S}_u(e), v \in \boldsymbol{S}_v(e), m \in \boldsymbol{S}_m\} \tag{1.16}$$

The methodologies presented in the following sections and chapters are built on AF representations. Thus, if not stated otherwise, the graph-theoretical notations $G, V,$ and E refer to their activity flow representations G_{AF}, V_{AF}, and E_{AF}.

1.2.3 Curation Standards

The molecular interaction databases mentioned in Section 1.2.1 are usually curated in AF format. Gene regulatory databases, for example, present their information in a tabular format, with the TFs, gene target, and the type of interaction, either positive or negative, in separate columns. Given the large scale of these interaction databases, a tabular format is convenient because it provides searchable, filterable, and easily processable data. KGs in PD format, however, contain a great variety of information, where each edge can contain a different number of nodes and changes, which is unfavorable for a standardized tabular format. Secondly, this complexity makes it difficult to link the information when browsing, making a visual representation preferable for curation and exploration. Building KG models under the FAIR (Findable, Accessible, Interoperable, Reusable) principles [60], therefore, requires a careful and standardized curation process [61]. Various standards have been established to ensure these principles, differing in their emphasis on visualization, syntax, or both. Given the complexity of the content and visualization, the KGs curated in these standards are referred to as **diagrams** (Figure 3A) or **pathway diagrams**, as they are very often modularized in the context of pathways (see Section 1.1.3).

Among the standards, SBGN is one of the earliest, offering a visual representation of different molecule classes and unique arrow-shaped depictions for their interactions [56]. On the other hand, the Systems Biology Markup Language (SBML) is a separate but complementary standard used for computational modeling of biological processes [62], [63]. It includes features for describing complex reactions, kinetic parameters, and additional annotations, enhancing the representation beyond aspects from SBGN. There are many more curation standards available, such as CellDesigner-SBML [64], KEGG Markup Language (KGML) [65], Graphical Pathway Markup Language (GPML) [66], Biological Pathway Exchange (BioPAX) [67], and Biological Connection Markup Language (BCML) [68], with varying levels of curation detail, visual style and, most important, interoperability [68], [69], [70]. Apart from general diagram standards, SBGN Bricks is another curation standard, offering an ontology of standardized PD representations for recurring biological processes to facilitate and further standardize the creation of pathway diagrams [63].

Several tools have been developed to create biological KGs within these standards. One of the earliest and most established tools, CellDesigner, offers a user-friendly interface to construct KG models via simple drag-and-drop mechanisms using pre-designed shapes and forms [64]. CellDesigner introduced its own standard, known as CellDesigner-SBML. As the name suggests, this is an extension of SBML, incorporating more visualization and layout

specifications while maintaining compatibility with other tools that use the SBML standard. All of the KG diagrams that I developed and that are presented in this thesis were curated in CellDesigner-SBML using the CellDesigner software. The curation processes are described in more detail in Sections 2.3, 6.2, and 7.2.1.

The WikiPathways resource [71] (see next section) developed PathVisio to create Pathway representations in their GPML standard, which can be integrated seamlessly into WikiPathways [66]. The relatively new Newt editor, built on the SBGNViz tool [72], is becoming increasingly popular due to its intuitive, modern design and web-based accessibility [73]. Newt supports CellDesigner-SBML, SBML, GPML, and SBGN Bricks.

1.2.4 Pathway Resources

In contrast to databases that curate unspecific molecular interactions, some resources contain manually curated KG diagrams, e.g., of pathways or cell types. Such databases offer standardized representations of data, detail process- or cell-specific interactions, and provide higher accuracy through manual annotation. KEGG (Kyoto Encyclopedia of Genes and Genomes) [74], Reactome [75], and WikiPathways [71] are three central pathway databases in the field of bioinformatics, each with unique features such as curation styles, file formats, APIs, functionalities, and integration with other tools. KEGG is characterized by primarily integrating data from multiple sources such as genomic, chemical, and systemic functional information. KEGG emphasizes molecular interactions within small pathways, linking to related databases like KEGG Genes or KEGG Enzymes. It supports standard bioinformatics formats, including KGML and FASTA. Its REST-style APIs enable programmatic access for bioinformatics pipelines. The functions of KEGG are diverse and include genome mapping, pathway mapping, and the analysis of relationships between diseases, genes, and drugs. Reactome focuses on an expert-driven curation approach with a targeted update cycle by a team of scientists, primarily covering human biology pathways. It offers data in formats like SBML and BioPAX and a multi-level organizational approach with various resolution scales. Reactome incorporates many nodes in their graphs representing biological processes and nested, encapsulated pathways, not just individual molecules. The Reactome Content Service API provides access to the critical data repository for computational analysis. WikiPathways is characterized by a community-curated model that encourages a variety of contributors and, consequently, a wide range of pathways. It supports GPML and BioPAX. The web service API provided by WikiPathways promotes a collaborative environment for accessing and

contributing to pathway data. Generally, these databases emphasize collaborative editing and curation and focus on the accessibility and diversity of their content.

1.2.5 Disease Maps

To ensure the relevance of KG models for disease research, they must first and foremost describe the most important disease-specific processes. The idea of tailoring KG models to specific diseases has led to the development of **Disease Maps**. These are community-built, comprehensive, and publicly accessible resources that collect validated knowledge about a disease, its molecules, phenotypes, and processes in KG formats [76], [77]. Given the interconnected nature of biological pathways that are not exclusively disease-specific, the concept of modularity is necessary for Disease Maps. Modularity enables the encapsulation of core disease mechanisms in higher-level representations while establishing comprehensive KGs that link all underlying processes. The standardized diagrams of modularized subgraphs in Disease Maps are often called (**sub-)maps** (Figure 3A). The ambition of Disease Maps extends past the technological facets of KG design. The aim is to provide publicly accessible, interactive models that can be utilized by bioinformaticians, laboratory researchers, and clinicians. Encoding this knowledge in a standardized format enables established analytical tools to extract information from complex interactions or perform *in silico* experiments on integrated experimental data. Examples of published Disease Maps include the Parkinson's Disease Map [78], the Rheumatoid Arthritis Map [79], the AsthmaMap [80], the Atherosclerosis Map [81], the Atlas of Inflammation Resolution (AIR, Chapter 2) [2], the Sarcopenia Map (Chapter 6) [6], The MASLD Map (Chapter 7), and the COVID-19 Disease Map [82].

Visualization is critical for exploring and understanding Disease Maps as an intuitive and interactive gateway, facilitating the interpretation and application of these resources. Platforms like MINERVA host many published Disease Maps, offering a sophisticated environment for visualizing and exploring their submaps. **MINERVA** was developed as a web-based platform for curating and interactively visualizing Disease Maps in SBML, CellDesigner-SBML, and SBGN formats [83]. It further provides an API for conversion between these modeling standards [84]. MNERVA's features include automated annotation with multiple databases, search capabilities for map content and drug, miRNA, and chemical targets using external APIs, and tools to support community-driven projects such as account management and commenting capabilities. MINERVA allows data upload and colored visualization of map elements. Another functionality is the visualization of genome and

protein sequences with a 3D protein structure in an interactive tool called MolArt, developed by the MINERVA team [85], utilizing the ProtVista JavaScript package [86]. Utilizing these features, many currently published Disease Maps, including the AIR or the Parkinson's Disease Map, are hosted on MINERVA.

The Disease Map Community (**DMC**) has been established to connect research groups working on Disease Maps projects. It has significantly contributed to the standardization of Disease Map curation and the reproducibility of existing models. The COVID-19 Disease Map illustrates the usefulness of Disease Maps in advancing our understanding of diseases and promoting collaborative research [82].

1.3 Data Integration on Knowledge Graphs

The first step in knowledge-driven data analysis is translating data into compatible formats and integrating them into the KG. **Data integration** refers to the association (=**mapping**) of entries and their values from a dataset, denoted as D, to nodes in the graph. The set of nodes with any mapped data that is either quantitative (level-based) or qualitative (activity-based) I refer to as **differentially change elements (DCE)** [3], denoted as V_d:

$$V_d = \{f_m(d) \mid d \in D\} \tag{1.17}$$

with the mapping function

$$f_m: D \to V \tag{1.18}$$

The definition of f_m depends on the KG curation, data type, and research question. Commonly, identifiers in the data are mapped to names or attributes of nodes in the KG. Values from the data, such as changes in the mRNA read counts or concentrations, are then integrated as a signal $s(u)$ for every $u \in V_d$ (as described in Section 1.1.2). The DCEs, thus, are characterized by a numerical value representing a relative change between samples or absolute values such as concentrations, either derived from experimental data (data-dependent, **empirical** DCEs) or arbitrarily assumed by the user (data-independent, **hypothetical** DCEs). DCEs from transcriptomics data are referred to as differentially expressed genes (DEGs) and are most often defined by an adjusted p-value < 0.05. Relative changes are usually represented as $log2$ fold change values (FC), resulting in negative values for a reduction in measurements (**downregulation, inhibition**) or positive values for an increase (**upregulation, activation**). DCEs can also be phenotypes referring to increased (positive value) or decreased (negative value) activities of measurable biological processes or

clinical features. The nodes to which entries from the data are being mapped depend on the data and granularity of the KG. The data integration and the level of the data on the biological scale finally determine the scope of the analysis. Figure 6 schematically shows the causalities between different scales of biological data defined by their KG representations.

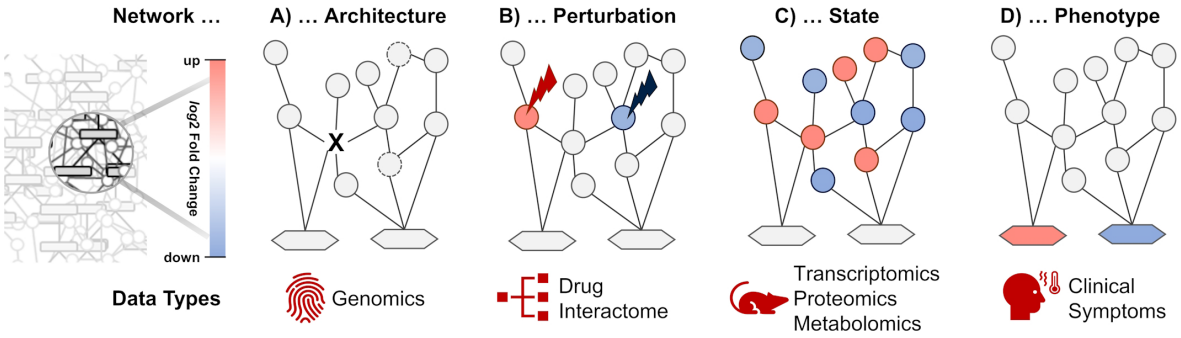

Figure 6: Biomolecular data types and their representation in knowledge graphs (KGs). (A) Population-wide references of molecular interactions form the foundational architecture of knowledge graphs. However, individual mutations can introduce alterations, necessitating adjustments to the graph's topology. (B) External stimuli, physiological or pathological, serve as perturbations in the biological system, inducing specific changes to graph nodes often considered as an input signal. (C) Data from biological experiments, notably from omics technologies, capture snapshots of a system's state at specific moments. Dynamic behavior predictions can be formulated by analyzing differences between successive system states. (D) Clinical traits, or observable characteristics, are integrated as higher-level nodes in the graph, linked to the molecular interactions they arise from. Adapted from Hoch *et al*. 2024.

1.3.1 Data Defining the Architecture

The interactions of proteins with other molecules are determined by their chemical properties. Throughout the population, the molecular structures of proteins are primarily fixed, so there are a finite number of possible interactions, most of which have already been identified. Genetic variations usually result in increased or decreased activity of individual gene products with minor effects on biological function but generally do not alter the underlying topology of KG representations. However, some mutations can result in complete loss or gain of function of gene products and cause severe disease. Similarly, chemical changes in proteins, such as those caused by drugs (acetylation of COX1 and COX2 by aspirin) or pathogens (ADP-ribosylation of G proteins by *Vibrio cholerae*), can affect their binding properties. In these cases, the KG topology changes to a state that differs from its default one. Identifying the resulting differences in signal transduction and incorporating them into the analysis of other data of the same origin is the principle of targeted and personalized medicine.

1.3.2 Data Representing Perturbations

Any stimulus that introduces a signal to the system by altering a molecule's state or concentration is called **perturbation**. Theoretically, any physiological action, including muscle

movements or digestion, can be considered a perturbation that triggers a nonautonomous response in related cells. Clinically relevant, however, are pathological perturbations such as injuries, pathogenic infections, or toxins, whereas perturbations caused by therapeutic interventions aim to improve the outcomes. These data describing the interactions between drugs and their target in the KGs are often referred to as **interactomes**. Spontaneous mutations can also be considered perturbations that cause diseases through pathological alterations in signaling. In the KG, a perturbation is integrated as a signal $\mathrm{s}(v,t)$ on a perturbed node v, usually at time point $t = 0$. In some applications, perturbations are integrated at several time points, e.g., when simulating successive administrations of the same or different drugs. In Chapter 4, I developed approaches and tools that allow *in silico* perturbation experiments on Disease Maps. In Chapter 5, specifically Section 5.3, I employ these methodologies to analyze large-scale interactome data of a multi-component drug.

1.3.3 Data Measuring the State

Accurately measuring signal transduction in its entirety within biological systems is still challenging. While technologies such as fluorescence resonance energy transfer (FRET) or bioluminescence resonance energy transfer (BRET) can measure single molecular interaction events in real-time, they are far away from large-scale applications on entire biological systems [87]. To conclude on processes happening in the system, individual snapshots must be recorded at individual points in time and space. This can be achieved, for example, using omics methods, which can simultaneously analyze comprehensive sets of biological data, such as measuring all mRNAs in a sample. However, they destroy the measured sample, requiring each omics analysis to come from separate samples, causing high variances in the data.

While consequent differences between samples on a larger scale can still be assessed by increasing the number of samples and statistical analysis, detailed signal transduction events become undetectable. Even if omics measurements were available at infinitesimally small intervals, since they come from different samples with inherent variance, they would introduce noise that makes detecting correlations impossible. Moreover, the functionality of a state cannot be directly inferred from the data of a single sample. A particular molecule's mere abundance or expression level does not directly translate to its activity or role in biological processes [88]. Therefore, biological experiments are designed as comparative analyses by contrasting conditions or observing shifts over time. Careful selection of the samples to be compared is crucial to ensure that the observed differences can be attributed to the research question. Even without external stimuli, the system is in an active state of physiological

processes such as circadian rhythm, heartbeat, or metabolism. An attempt is made to remove such information through statistical tests or as a form of abstraction from the KG model. Physiological noise on a low temporal scale, such as metabolism, can be removed from the data by statistically analyzing a sufficient sample size. Fluctuations on a higher scale, such as circadian rhythms, must be considered in the experimental design [89]. Samples may show significant differences at different times of the day, leading to an increase in false positives if both sample groups were taken at different times or false negatives due to increased noise if all samples are spread throughout the day.

From the state of a system, researchers usually either aim to predict effects on non-accessible higher-level processes, e.g., when translating data from *in vitro* experiments, or identify possible perturbations causing or reversing the observed state. When looking purely at the state's observations without prior knowledge, those cannot be directly correlated to underlying perturbations. For example, a drug activating a protein with negative feedback on its expression results in an observed decrease in mRNA levels, resulting in misinterpretation [90]. KG data integration can solve these issues by including prior knowledge of the analysis's underlying transcriptional regulation. When both the perturbation and state data are integrated, e.g., when studying the effects of a drug with known targets, KG approaches can be used to simulate the signal transduction mechanisms and get deeper insights into the mode of action [91].

Some factors can also distort experimental data. One example is bulk data (RNA-Seq), where the aggregated amounts of mRNA are measured in a complete tissue sample and not in individual cells. Consequently, the measurements are influenced by the cellular composition [92]. Cells that express a gene particularly strongly or weakly change the total mRNA levels of the gene in the entire tissue sample in proportion to their number. It is, therefore, difficult to deduce whether the observed difference is due to a change in actual gene expression or the overall composition of the cells.

1.3.4 Data Describing Phenotypes

A challenge in systems biology is deriving higher-level information, i.e., recognizable biological processes or clinical features, from lower-level data (molecular data) or vice versa. Rather than describing the molecular state of a system, such phenotypic data represent how the state manifests itself at higher biological levels, some of which can be observed with the human eye, e.g., clinical symptoms. Although some databases attempt to standardize the terminology of phenotypes and collect the associated molecules to ensure reproducibility and

interoperability, these are not yet established in the clinical context. While -omics measurements are usually performed via platforms that aggregate probes into standardized formats during data pre-processing, clinical measurements can differ. They are often measured using different approaches and individuals, and even the same parameters can be assessed using different methods or scoring systems.

As described in Section 1.1.3, in KGs, higher-level processes are integrated as individual phenotype nodes modulated by process-specific subgraphs. However, accurate representations may be lacking or not accurately represented without standardization. Furthermore, the curation of phenotypes in the KG may be more or less detailed than the clinical measurements. For example, "edema" in KGs could be represented by multiple phenotypes, such as "increased vascular permeability" or "vasodilation." Consequently, mapping phenotype data to KGs requires manual curation, ideally in consultation with clinical experts. Only then can clinical scores be individually transformed into a standard numerical format. These standardized scores could then be integrated into the Knowledge Graph to the corresponding phenotype node or multiple phenotype nodes. In automated approaches, heterogeneous KGs of symptom-disease, symptom-phenotype, and phenotype-gene associations are generated from ontology databases, but these are not targeted and do not contain functional information [93].

1.4 *In silico* Analysis Approaches

1.4.1 Machine Learning

KG-independent data-driven analysis involves identifying properties within the data itself, often through dimensionality reduction approaches using machine learning approaches. This process simplifies complex datasets to expose underlying patterns or relationships. Supervised machine learning approaches require a predefined set of labels or classes for the data. They learn to predict these labels from the input data, making them suitable for tasks where the classes are known in advance. In contrast, unsupervised machine learning approaches aim to learn from new data without prior knowledge of labels or classes, thus the term “unsupervised.” These unsupervised methods focus on organizing the data by reducing the high dimensionality of the input data to a lower dimensional representation [94]. In this representation, each point corresponds to an individual sample in the data. Such visualization enables a feasible assessment, as samples with similar values are positioned close to each other, while separated samples indicate considerable differences. Unsupervised machine learning

becomes particularly useful when identifying cell types from large-scale single-cell data, detecting outliers that may disrupt statistical analyses, or providing insights into broader data patterns in clinical data. They effectively assess whether there are underlying populations in the data that may influence further data analysis. Principal Component Analysis (PCA) [95] is a mathematical approach that transforms high-dimensional data into a set of orthogonal components, where the first components capture most of the variance in the data. Uniform Manifold Approximation and Projection (UMAP) [96] uses manifold learning and topological data analysis. It can effectively preserve both the local and global structures of the data. In Chapter 4, I utilize UMAP to analyze single-cell RNA-Seq data from immune cell types to identify cell clusters with similar process-specific expression profiles.

1.4.2 Topological Analysis

Graph theory is an expansive field with many applications, extending beyond systems biology into areas such as social science, infrastructure, and others. Many of these applications have established methodologies for quite some time. Because deriving information from KGs requires understanding their topological structure, graph theory approaches have found large applications in systems biology [97], [98]. One application is the identification of (shortest) paths, described in Section 1.1.2, which is an essential step to assess potential indirect interactions across KGs. Several well-established algorithms have been developed to derive paths from KGs, which differ in their computational complexity depending on the KG properties, such as whether edges are directed and signed, weightings, or the presence of feedback loops [99]. In Section 2.6 and Section 6.6, I use these algorithms to filter, select, and highlight paths on Disease Maps for interactive KG exploration. Identifying all paths in KGs, i.e., all connections between each pair of nodes, has a computational complexity that grows exponentially with the size of the KG. However, algorithms that focus instead on determining only the shortest paths between two nodes in the KG can be highly biased by (i) misestimating the length of interactions that lack intermediates, (ii) neglecting the biochemical relevance of longer pathways, and (iii) overrepresenting more intensively studied molecules.

Topological analysis further includes calculating centrality measures that numerically represent the local or global interconnection of nodes and edges in the graph [100], [101]. Examples of centrality measures include the following.

The in-degree of a node v, denoted as $c_{d_{in}}$, is the number of edges directed towards the node, defined as:

$$c_{d_{in}}(v) = |\{u \in V(G) \mid (u,v) \in E(G)\}| \tag{1.19}$$

The out-degree of a node v, denoted as $c_{d_{out}}$, is the number of edges that the node directs towards other nodes, defined as:

$$c_{d_{out}}(v) = |\{u \in V(G) \mid (u,v) \in E(G)\}| \tag{1.20}$$

The betweenness centrality c_B of a node v is based on the number of shortest paths that pass through v (denoted as $\sigma_{\rightarrow v \rightarrow}$):

$$c_B(v) = \sum_{s,t \in V(G), s \neq v \neq t} \frac{|\sigma_{\rightarrow v \rightarrow}(s,t)|}{|\sigma(s,t)|} \tag{1.21}$$

The closeness centrality c_C of a node v is a measure of how close the node is to all other nodes in the graph. It is defined as the inverse of the sum of the shortest path distances from v to all other nodes, given by:

$$c_C(v) = \sum_{u \in V(G), u \neq v} \frac{1}{\ell(\sigma(v,u))} \tag{1.22}$$

Centrality measures as properties of KGs can be combined with other methods to extend their informative value and possibly improve their accuracy. Similarly, in Chapter 2, I improve the enrichment analyses, which are reviewed in the next section, with a novel approach incorporating topological data. In Chapter 3, I use topological weightings in UMAPs of RNA-Seq data to identify functional cell clusters.

Modern tools now offer interactive visualizations of large-scale KGs and provide functionalities to analyze them with well-established topological algorithms. In the realm of systems biology, **CytoScape** has emerged as a popular tool for topological analyses [102]. CytoScape allows the integration of third-party plugins to enhance its functionality. CentiScape and KeyPathwayMiner are among the most frequently used tools for topological analysis in this platform and were employed in the analysis of experimental data in Section 1.5 [103], [104]. Other plugins like cy3sbml enable importing SBML or SBGN files, further expanding the platforms' utility [105]. In data analysis pipelines, the NetworkX Python package provides a variety of functionalities for KG creation, topological analysis, visualization, and export in various file formats, such as Graph Modelling Language (GML) [106].

1.4.3 Pathway Analysis

The introduction of omics technology and the significant increase in measurements changed the research question. Experiments shifted from a targeted investigation of a few processes by specific molecules to an EDA of comprehensive datasets. To make sense of the data, it must be transformed into interpretable information, e.g., by assessing the effects on phenotypes, cell types, and diseases. As outlined at the beginning of Section 1.4, the information curated in KGs provides an excellent basis for such assessments. Particularly in the case of modularized, process-specific KGs, the data mapped to nodes can be projected onto the entire module or a phenotype node, a method commonly referred to as pathway analysis [107]. A variety of approaches have been explored, utilizing signal transduction simulations, such as in HiPathia [108] or TieDie [109], or statistical enrichment with or without consideration of topological features [110]. Garrido-Rodriguez *et al*. recently extensively reviewed many of these methods [107]. My thesis will focus primarily on enrichment-based analysis approaches, described in more detail in this section. Over the years, many approaches have been developed using different data integration and statistical methods, which are reviewed extensively in [111], [112].

One of the earliest computational methods to infer knowledge from large-scale data is the overrepresentation analysis (ORA) [112]. In general, ORA calculates the probability of entries from an input list occurring more or less in another set than expected using Fisher's exact test to assess the significance of the overlap. Applied to biology, the input list can be empirical data, e.g., DEGs from a transcriptomics experiment, and the set can be a predefined group of genes associated with a superordinate term. This term can refer to any disease, cell type, tissue, compartment, or pathway. It can even represent a single gene; in this case, the set could consist of regulatory TFs or miRNAs. A common standardized format for the description of gene sets is the Gene Matrix Transposed File Format (GMT) [113]. GMT files have a tabular structure, each line representing a separate set. The first line contains the name of the set (e.g., the pathway), the second a more detailed description of the set, and a separate column for each gene in the set.

Many ontology resources, such as HPO or GO, have already curated gene sets for the ontology terms in their database. Since ORA does not require parameterization or information about the type of relationship in the sets, they can be curated from diverse resources with a certain degree of abstraction. Which sets are selected for analysis depends largely on the experimental design and the research question and has to be carefully chosen by the user. ORA is generally applicable to any data type as long as fitting sets are available. The applicability

and simple specifications make ORA an attractive method to assess the biological relevance of information in large-scale data. In the past, ORA was and still is mainly applied to transcriptome data due to its accessibility and abundant knowledge of gene functions to create the sets. ORA analysis of transcriptome data has, therefore, shaped the terminology, with the sets for the analysis being referred to as gene sets and the enrichment approach generally referred to as "gene set enrichment analysis" (GSEA). However, the abbreviation GSEA became better known as a name for a separate enrichment approach that differs from the original ORA [114].

While ORA evaluates the overlap of data with the curated sets, it does not provide any information on the strength and direction of data nor specific the relations in the sets. Second, the statistics in ORA are biased by the size of the gene sets, which can be mitigated to some extent if more information is included in the analysis [115]. Several approaches have been developed to solve this issue by integrating more information into the analysis. GSEA extends the ORA approach by ranking the input genes by their FC or p-values and analyzing whether their high- or low-ranked genes are overrepresented [114]. Approaches like GSEA are often referred to as Functional Class Scoring (FCS), which distinguishes them from the original ORA. Several commonly used analysis tools, such as GeneTrail [116], have integrated the GSEA approach.

Still, GSEA does not evaluate the relationship between the genes and the enriched node. Numerous enrichment approaches have addressed these limitations to broaden their scope for specific purposes, commonly referred to as topology-based approaches. A study from 2019 validating enrichment analyses with disease-specific data, found that TB approaches improve upon others [110]. Topology-based enrichment distinguishes between up- and down-regulated edges between genes (BD-Func) or integrates KG topology information to weight their algorithms (network-weighted GSEA) [117], [118], [119]. The “Reverse Causal Reasoning approach” (further referred to as RCRA) integrates KG information of upstream nodes and statistically analyzes whether their regulatory directions correspond to the FC directions [120]. However, RCRA does not include FC values of genes in the list and only considers direct upstream regulations, restricting applications of the approach.

With numerous extensively curated KG databases available, enrichment-based methods and data analysis platforms are being developed to utilize their knowledge. PANTHER (Protein ANalysis THrough Evolutionary Relationships) [121] and DAVID (Database for Annotation, Visualization, and Integrated Discovery) [122] are well-established examples of web-accessible data analysis tools. The Enrichr platform performs ORA on

automatically curated gene sets from ontologies, pathway resources, cell type data, miRNA and drug interaction databases, and more [123]. Like PANTHER and DAVID, Enrich is available on a ready-to-use website that allows quick analysis of large data for all the gene sets available. Such platforms allow researchers to analyze their data efficiently in multiple contexts and are an excellent example of high accessibility. ORA is also employed by many other data analysis tools, such as the CytoScape plugins Biological Networks Gene Ontology (BiNGO) [124] and ClueGO [125]. The knowledge graph resources also have developed analytics assessments integrated into their platform, such as ReactomeGSA [126]. In 2014, QIAGEN published the "Ingenuity Pathway Analysis" (IPA) software that provides a range of KG solutions to infer knowledge from molecular data, including their own topology-based enrichment approach [127]. Like RCRA, IPA considers only directions of gene expression regulations but additionally analyses downstream effects, including multiple steps in the KG and implementing a more sophisticated statistical analysis. A newer web platform, the Consensus Pathway Analysis (CPA), follows a combinatorial approach, enabling the simultaneous evaluation of multiple enrichment approaches and visualization of multiple data sets [128]. A recently published platform, Static and Temporal Analysis of Gene Expression Studies (STAGES), focuses on a more user-oriented approach with a streamlined interface and integrated general data formats, including Excel sheets [129]. Although most of these approaches and platforms enable the analysis of experimental data, they are either limited in their informative value due to insufficient use of topological properties or in their applicability to heterogeneous and large-scale KGs. In Chapter 3, I reflect on these limitations and tackle them by developing a novel topology-based enrichment approach.

1.4.4 Boolean Models

Topological analysis is a powerful tool in the research of graph-structured representations, capable of evaluating graph patterns and identifying unanticipated links between nodes. However, when it comes to modeling disease mechanisms, it's essential to integrate the temporal dynamics involved in signal flow. Kinetic models provide an extremely accurate way to simulate the changes in molecular concentrations over time. These models rely on ordinary differential equations (ODEs) to represent the kinetics of chemical reactions. However, the curation of these models is highly time-consuming. It necessitates the retrieval of experimental data under comparable conditions to parameterize reactions in the model. Furthermore, specific processes, such as blood flow and the distribution of molecules in the cell or bloodstream, can lead to exceedingly complicated kinetic parameters. As a result,

kinetic models tend to be restricted to small-scale applications, such as simulating a single pathway in a controlled environment. For large-scale KG models, simulations with qualitative or arbitrary numerical representations are needed.

Boolean modeling has emerged as an effective method to conduct large-scale simulations of complex models. In Boolean models, each node $v \in V$ is described by a qualitative signal called **state**, defined as $s(v) \in \{0,1\}$, referring to the node being active (ON = 1) or inactive (OFF = 0). In discrete time steps, the state of each node is then updated with the new state determined by logical rules based on the states of input nodes in the previous step. For each node v, its state at time $t+1$ is determined by a logical function f_v based on the states of its input nodes at time t, denoted as:

$$s(v, t+1) = f_v\big(s(u_1, t), s(u_2, t), \dots, s(u_k, t)\big) \tag{1.23}$$

where $\{u_1, u_2, \dots, u_k\} \subseteq N(v)$ are the input nodes for v. The logical function f_v could be simple (such as AND, OR, NOT, NAND, NOR, etc.) or more complex. For instance, if f_v is an AND function for a node v with two inputs u_1 and u_2, then $s(v, t+1) = s(u_1, t) \wedge s(u_2, t)$. The initial state $s(v, 0)$ for each node v needs to be specified to start the model. From a defined initial state, Boolean models thus simulate signal transduction mechanisms. Due to the finite number of states a Boolean model can have, the simulation eventually enters an already encountered state, provided that no disturbances are introduced during the simulation. Since Boolean models are usually deterministic, the simulation has reached a stable state, either oscillating over several steps or being fixed in one state. The steady state is often referred to as the **attractor** of the model. Analysis of the number of active states during the steady state as a function of a given input makes it possible to determine correlations between [130], [131]. Moreover, in Boolean models, the computational time increases only proportionally to the complexity of the KG, allowing efficient high-throughput analyses. In Chapters 6 and 7, I employ Boolean modeling to simulate and predict possible mechanisms underlying systemic disease processes using organ-modularized KGs.

Numerous modeling resources have been developed that enable the analysis of Boolean models in different environments, e.g., MaBoSS [132] and CellNetAnalyzer [133] in MatLab, BoolNet [134] in R, or PyBoolNet [135] in Python. The CellCollective web platform provides accessible tools for dynamic simulations and perturbation experiments of Boolean models. The platform leverages different file formats, such as Boolean expressions, GML, or SBML qual [130]. SBML qual is an extension of SBML, enabling a standardized representation of qualitative models [136]. The CaSQ tool converts SBML files to SBML qual by inferring logical rules from PD specifications, thus providing a bridge between KG resources, Disease

Maps, and qualitative modeling tools [137]. Utilizing CaSQ, Singh *et al.* generated a Boolean model from the Rheumatoid Arthritis Disease Map (RA map [138] to simulate drug effects and predict therapeutic targets [139]. In summary, Boolean models offer a relatively computationally inexpensive yet powerful method for investigating the behavior of large-scale KGs.

1.4.5 Agent-Based Modelling

Spatial dynamics must also be considered when diving into the high resolution of single cells. Processes that might seem straightforward in isolation can exhibit profound spatial complexity through interactions with neighboring cells and the surrounding cellular microenvironment, including the extracellular matrix [140]. These interactions necessitate models that can capture these spatial nuances. On a larger scale, when considering individual agents like cells and their interactions in an environment, Agent-Based Modeling (ABM) becomes valuable [141]. In this approach, the actions of independent biological units, so-called agents, are evaluated on a rule-based basis in dependence on other agents. Their collective actions can lead to emergent phenotypic changes, often observable at tissue or organ levels. Dutta-Moscato *et al*. developed an ABM of liver fibrosis progression through liver cell proliferation, reconstructing histological data *in silico* [142]. The PhysiBoSS platform was developed from the 3D ABM software PhysiCell and the MatLab Boolean modeling tool MaBoss [143]. It provides a multi-scale platform that combines agent-based simulations at the cell/tissue level with underlying molecular signaling. By alternating simulations at different scales and using the results of one scale as inputs to another, PhysiBoss efficiently handles the challenge of granularity. Pushing the computational limits of ABMs, so-called "Giga Scale Models" aim to bring cellular models up to macroscopic scales [144]. ABM and similar approaches have emerged as potential solutions, but their application and optimization for diverse disease contexts is an ongoing challenge. In Chapter 7, I build on the idea of ABMs, particularly the multi-scale approach of PhysiBoSS, by simulating spatiotemporal disease mechanisms through multi-compartmental Boolean models.

1.5 Challenges in Knowledge Graph-based Analyses

In collaboration with the Department of Cardiac Surgery at the Rostock University Medical Center, I employed KG and ORA methods to analyze gene expression in differentiated cardiac cells [145]. Specifically, I investigated molecular pathways in pacemaker

cells generated *in vivo* from induced sinoatrial bodies (iSABs) and antibiotic-selected cardiac bodies (aCaBs). Both iSABs and aCaBs are derived from murine embryonic stem cells (mESCs) but differ in the differentiation protocol. While aCaBs were differentiated from mESCs using only Myh6-antibiotic selection, the iSaBs were additionally transfected with Tbx3.

The data I analyzed consisted of DEGs, including FC values and adj. p-values comparing the gene expression between iSaBs vs. aCaBs. I applied KG- and enrichment-based approaches to identify processes and genes characteristic of the differentiation process. I identified overrepresented GO terms from the set of DEGs (adj. p-value < 0.05) in iSABs using the BiNGO and ClueGO plugins of CytoScape (Figure 7A). Both approaches are based on ORA, with Bingo providing a tree-shaped overview of related overrepresented terms, while ClueGO shows the terms in relation to overlapping groups of genes. While the statistical evaluations provide a first impression, checking the functional categories in the GO hierarchy is essential. It is highly likely that when a whole branch of the GO tree is highlighted as overrepresented, the nodes furthest down the hierarchy can be expected to be the most biologically relevant ones. In the following analysis step, I generated a PPI to identify modules of interconnected, i.e., functionally related, DEGs. Next, I extracted a PPI from the BioGrid and String databases, creating a combined KG of 8,120 nodes and 55,720 edges. I then employed another CytoScape tool called KeyPathwayMiner (KPM) to extract relevant subgraphs from the KG [104]. KPM ranks the subgraphs by the number of DEGs associated with their nodes and minimizes the number of non-DEGs connecting DEG nodes. Finally, I reapplied ClueGO to the extracted subgraphs to identify subgraph-specific overrepresented GO terms (Figure 7C). We drew conclusions on the underlying processes in the different types of stem cell-derived cardiac pacemaker models from the information obtained at each analysis step.

From today's viewpoint, the methodology applied in this study contains multiple limitations and computational bottlenecks that hinder its general applicability. The data analysis workflow is subject to several limitations that affect the interpretability of the results. First, KG files must be manually downloaded, merged, and processed to be integrated into CytoScape, requiring the installation of the software and individual plugins. In addition, the data files of the DEGs require a particular format for each tool used in the workflow. The function of the DEGs in the processes is also not considered in the enrichment analysis, which does not provide insights into whether more or less relevant genes are differentially expressed or whether they positively or negatively affect the enriched term. Moreover, the KG extracted from STRING and BioGrid is an undirected graph. Consequently, there is no information on the direction between DEGs and their role in the subgraphs included in the KG. Thus, the enrichment does not provide details on the direction and strength of the modulation.

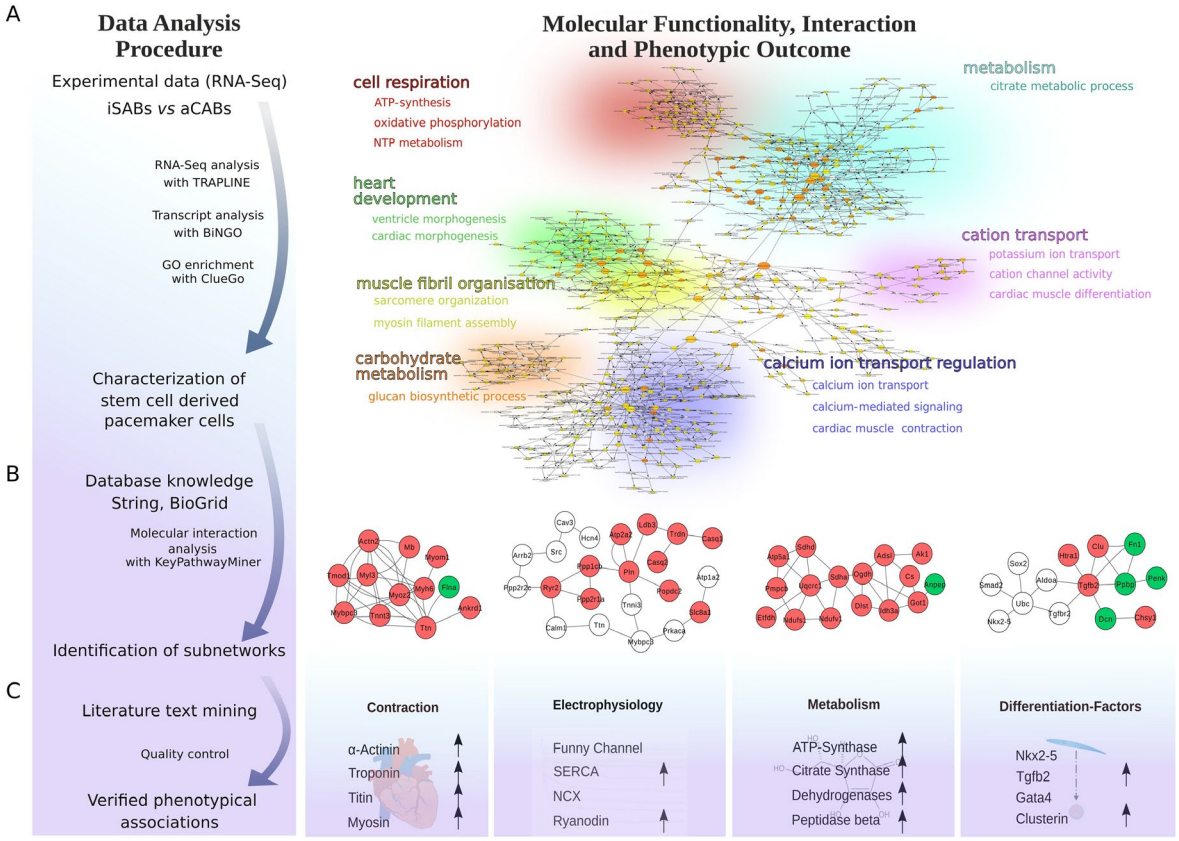

Figure 7: Systems-based data analysis procedure for identifying molecular functionalities, interactions, and phenotypic associations applied to stem cell-derived cardiac cell types using RNA-Seq data. (A) Calculation of overrepresented GO terms using the Cytoscape applications BiNGO and ClueGo. (B) Identified subgraphs obtained after KeyPathwayMiner (KPM) analysis of the former constructed interactome KG. Red represents the upregulated transcripts within iSABs, and green represents the downregulated transcripts. The edges (lines between encircled genes) are experimentally verified interactions obtained from String and BioGrid. (C) Summary of the upregulated factors identified in the data and the literature for processes within contraction, electrophysiology, metabolism, and differentiation. From Hausburg *et al.*, 2017.

1.6 Thesis Motivation

The field of systems biology has experienced enormous growth in the last few years, and new approaches are constantly being developed. Most computational approaches are often focused on a specific research question and are limited to small aspects or single scales of disease mechanisms. However, diseases affect biological systems at multiple temporal and spatial scales and disrupt biological signaling pathways at multiple scales. Consequently, the applicability of small-scale approaches to other research questions and projects may be limited, even if reproducibility is guaranteed. While this in no way diminishes the relevance of such approaches and their important scientific output, it does raise the question of whether more efforts should be made on reusable methods. The adoption of FAIR principles, promoting findability, accessibility, interoperability, and reusability of scientific data, is increasingly encouraged through initiatives like FAIRdom [60], [146]. However, accessibility

is often understood in purely technical terms, suggesting that while methods may be technically reproducible, their practical application in day-to-day research remains limited. Large parts of the data generated in laboratories and clinical practice are not adequately utilized, and their potential may not be exploited [147]. While legal and ethical considerations justifiably constrain data availability, another significant barrier might be the specialized knowledge required to leverage these tools effectively.

Addressing these challenges motivates the development of user-friendly tools that do not require extensive computational expertise. Interdisciplinarity is not only ensured through research collaborations but also through efforts that make such collaborations more feasible in the long term. Disease Maps are a great example of a step in the right direction, building publicly available and easily accessible resources for researchers to visualize and analyze their data. MINERVA continues to be an excellent framework for Disease Map projects, offering (i) an interactive, standardized, and web-based accessible interface for users, (ii) an extensive API to integrate Disease Maps into computational data analysis workflows, (iii) plugins for connecting disease-specific modeling approaches. Newly developed data analysis tools must ensure reproducibility and ease of use for researchers unfamiliar with bioinformatics. The accessibility of the methods should, therefore, be an essential feature. They should be easily accessible, if possible, via a web browser without installing any software.

In this thesis, I developed three such platforms, each specific to a field of research, specifically inflammation, gastrointestinal diseases and sarcopenia, and steatotic liver diseases. They are tailored to the individual biological contexts and users' potential experimental or clinical research questions. I created KG-based data analysis approaches integrated into these platforms, enabling disease-specific modeling. In **Serhan & Gupta, *et al.*, 2020** [2], we created the "Atlas of Inflammation Resolution" (AIR), a publicly available and interactive Disease Map describing molecular mechanisms of the innate immune response in a multi-level layout. In **Hoch *et al.*, 2022a** [3], I developed a novel enrichment-based approach utilizing KGs, labeled Two-Dimensional Enrichment Analysis (2DEA), to infer new insights on heterogeneous data across multiple biological levels. I employed this approach to investigate multi-target drug mechanisms described in **Hoch *et al.*, 2023a** [5]. In **Hoch *et al.*, 2023b** [4], I utilized the KG of the AIR to investigate gene regulatory mechanisms in immune cells using cell-specific KGs generated from single-cell RNA-Seq data. In **Hoch & Ehlers *et al.*, 2022b** [6], we created the "Sarcopenia Map," a publicly available and interactive Disease Map as a logic model of molecular interactions linking nutrition, gastrointestinal diseases, and sarcopenia. I supported the creation of the NaviCenta (**Scheel *et al.*, 2023** [148]) and COVID-19 (**Ostaszewski *et al.*, 2021** [149]) Disease Maps by providing KG curation efforts and analysis

tools. The most recent effort is the development of the MASLD Disease Map, a resource that utilizes KGs from the AIR and Sarcopenia Map to model the molecular mechanisms in liver disease under spatial aspects. The following chapters describe the design, development, and application of these resources and tools, highlighting the scientific relevance of accessible and comprehensive platforms for biomolecular data analysis.

1.7 Terminology

The following box summarizes the most important terminology introduced in the previous sections. The definitions are specific to biology and may be different in other fields.

System:
An association of biological entities, ranging from molecules to organisms, that collaboratively contribute to biological functions.

Model:
An abstracted representation of a biological system designed to facilitate the study of its behavior with minimal information to account for the limitations of data and computational resources.

Knowledge Graph:
A graph-structured model that describes the conceptual, functional, or physicochemical relationships between entities of a system across biological levels.

Phenotype:
The emergence of a system's behavior at a higher, often spatial, scale that cannot be directly attributed to a physical entity. In KGs, the term also refers to its associated node representation, usually an SBGN phenotype element.

Signal:
A qualitative or quantitative property (or change therein) of a biomolecular entity that is caused by or induces a signal in another or the same entity.

State (System):
Collective signals of all biomolecular entities in a system at a certain point in time.

State (Entity):
The signal of an entity with discrete values - common in qualitative approaches, such as Boolean models.

Simulation:
The use of computational models to emulate the behavior of a system, often through a discrete or continuous update of signals in all the system's entities (= signal flow).

Prediction:
A state of a biological system or its entities assessed through simulations.

Disease Map:
Publicly accessible and community-driven resource that supports research through interactive knowledge graph representations of disease mechanisms.

Submap:
Modularized part of a Disease Map's knowledge graph describing one or more biological processes.

Chapter 2

A Multi-Level Knowledge Graph on Inflammation

2.1 Inflammation as a Complex Biological System

Acute inflammation is a crucial defense mechanism exhibited by the immune system when faced with foreign pathogens or tissue damage. Ideally, this immune response is intended to be localized, self-limiting, and aimed at restoring physiological homeostasis [150], [151], [152]. While inflammation represents a natural and essential immune reaction to injuries or infections, its improper resolution or persistent activation can cause chronic inflammation and a magnitude of diseases [153], [154], [155], [156], [157]. The initiation phase of acute inflammation is orchestrated by pro-inflammatory mediators (**PIMs**) released from immune cells after recognizing pathogenic or damage-associated molecules [153], [158], [159]. These mediators have chemogenic functions, activating other immune cells and further stimulating their production, resulting in a positive cascade. Among these mediators are groups of lipid mediators (LMs) synthesized from arachidonic acid (AA), mainly divided into prostaglandins, thromboxane, and leukotrienes.

For a long time, the resolution of inflammation was considered a passive process, with PIM concentration declining after removing the initial stimuli. However, since then, many mediators that actively regulate inflammation through negative feedback on PIM production and immune cell activation have been identified, suggesting that inflammation is instead an active process [151], [152]. These specialized pro-resolving mediators (**SPMs**) consist of cytokines such as IL10 and IL4 as well as over 30 LMs, predominantly derived from ω-3 polyunsaturated fatty acids such as docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA) [160], [161]. Unlike PIMs, which promote inflammation, SPMs actively counteract and facilitate the resolution process. They work by stimulating the clearance of inflammatory cells, reducing the production of pro-inflammatory cytokines, and promoting tissue repair and regeneration. SPMs exhibit a variety of actions, such as enhancing phagocytosis of cellular debris, suppressing neutrophil migration, and promoting the uptake of apoptotic cells. These actions restore tissue integrity and function,

promoting the return to homeostasis. Research on SPMs has uncovered their immense therapeutic potential for a wide range of inflammatory conditions. Studies have shown that exogenous administration of SPMs can accelerate the resolution of inflammation and promote tissue repair. As a result, SPM-based therapies have emerged as a promising approach for treating chronic inflammatory diseases, including arthritis, asthma, inflammatory bowel disease, and cardiovascular diseases [162], [163], [164], [165], [166], [167]. Harnessing the pro-resolving capabilities of SPMs provides a novel and exciting strategy for developing targeted and efficient treatments for inflammatory disorders [168], [169], [170].

These findings emphasize inflammation as a non-linear process complicating the interpretation of scientific data. The loose spatial restrictions promote this complexity, as the immune response can occur almost anywhere in the body, locally or systemically. Immune cells circulate throughout the body and can invade tissues across various barriers. Additionally, the immune system shows priming behaviors [171]. Specific immune mediators can alter cellular states, preconditioning these cells to exhibit altered reactions to future encounters with stimuli. They interact with the local microenvironment and adapt their responses accordingly [172]. This higher-level diversity is inherently mediated at the molecular level, resulting in many potentially relevant molecules depending on the spatial and temporal context [173], [174]. Consequently, the interpretation of molecular data becomes highly non-intuitive. A challenge lies in discerning whether a molecule functions as a friend (beneficial) or a foe (detrimental), which largely depends on the spatiotemporal context of the immune response [175]. The clinical relevance of regulating biological processes is context-dependent as well. Although models of the immune system have been successfully developed for specific tissues and disease processes, due to this unpredictability and variability, there has been no model that describes the detailed mechanism of acute inflammation and, especially, inflammation resolution in general [176], [177], [178], [179], [180], [181], [182], [183], [184].

Computational KG approaches could support data analysis, bringing molecular variety into context through KGs of immune cell signaling [185], [186]. Tools that could detect patterns in experimental data, bring them into the context of inflammation, and link results to prior knowledge could facilitate their interpretation. Given the spatiotemporal non-specificity of inflammation, topology-based enrichment approaches could be particularly helpful in clarifying causalities and the friend-foe role of molecules in the diverse inflammatory processes. As such resources have been lacking, we aimed to develop

a large-scale KG-based approach that describes the molecular processes underlying acute inflammation and its resolution and enables data integration and analysis. This effort resulted in the development of the "Atlas of Inflammation Resolution" (AIR), which integrates such a KG into an interactive and publicly accessible Disease Map [2]. The following sections outline the roadmap of AIR, describing its development, publication, expansion, and application in more detail.

2.2 Designing the "Atlas of Inflammation Resolution"

We aimed to design the AIR as a resource that captures state-of-the-art acute inflammation and inflammation resolution research. Although inflammation is not a specific disease, the concept of AIR is consistent with that of disease maps and is therefore referred to as such in this thesis. I believe disease maps should not necessarily be understood as disease-specific but as descriptions of context-specific processes whose behavior can be pathologically altered with clinical relevance. In AIR, the context is inflammation and the resolution of inflammation.

Under the considerations from Section 1.6, the AIR should be freely accessible and provide an intuitive knowledge repository and data integration and analysis tools. To connect the resource with the scientific community, we collaborated with many scientists worldwide with many years of experience in the field of inflammation. They were involved in conceptualizing the AIR and its design, validating the content, and co-authoring the initial manuscript [187]. The following table lists the collaborators and their affiliations at the time of publishing the AIR.

Table 1: International collaborators that supported the "Atlas of Inflammation Resolution".

Name	Affiliation	Country
Mauro Perretti	The William Harvey Research Institute, Queen Mary University of London	UK
Catherine Godson	Diabetes Complications Research Centre, University College Dublin	Ireland
Yongsheng Li	Clinical Medicine Research Center, Third Military Medical University	China
Oliver Soehnlein	Department of Physiology and Pharmacology (FyFA), Karolinska Institutet	Sweden
Takao Shimizu	Department of Lipidomics, The University of Tokyo	Japan
Oliver Werz	Department of Pharmaceutical/Medicinal Chemistry, Friedrich Schiller University Jena	Germany

Name	Affiliation	Country
Valerio Chiurchiù	Department of Medicine, Campus Bio-Medico University of Rome	Italy
Angelo Azzi	JM USDA-HNRCA at Tufts University	USA
Marc Dubourdeau	Ambiotis	France
Anurag Tripathi	CSIR – Indian Institute of Toxicology Research	India

To adequately support researchers in data analysis and enable knowledge inference across multiple biological levels, the AIR needs to describe the functional links between these levels in detail. We designed a KG for the AIR in a multi-level layout of separate layers with increasing resolution of the biological levels (Figure 8): from the top level, including higher processes such as cellular interactions and clinical phenotypes (Figure 8A), to molecular pathways that modulate these top-level phenotypes (Figure 8B), to their detailed regulation at the transcriptional level, including large numbers of TF, miRNA and lncRNA interactions (Figure 8C). The bottom layer we refer to as the Molecular Interaction Map (MIM) as it contains solely interactions between molecular entities in AF format.

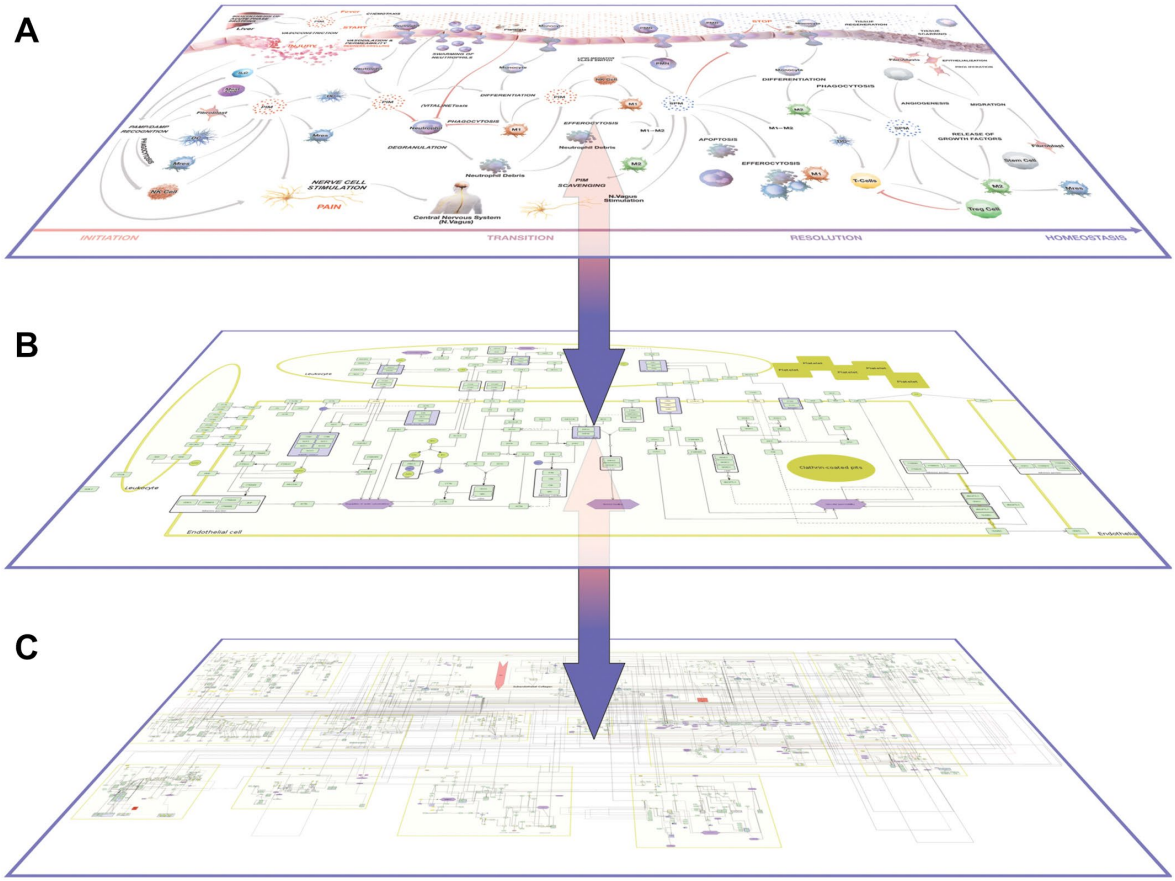

Figure 8: Hierarchical organization of the AIR. (A) The top phenotype layer contains immune cell types, cellular processes/phenotypes, and tissue-level organization. Clinicians are generally interested in connecting their patient data to this layer. (B) Each process in the top layer is connected to a respective KG diagram. The

process layer describes key molecules/pathways regulating processes in the top layer. This layer is suitable for research scientists to generate new hypotheses on the mechanistic insights of disease phenotype regulation. (C) The lower layer contains a comprehensive Molecular Interaction Map (MIM) where all the processes are merged at the molecular level. The layer is also enriched with currently available experimentally validated regulatory information. Due to the communication across multiple layers, the AIR provides a platform for integrative data analysis. From Serhan & Gupta *et al*. 2020.

The workflow for the curation of the AIR KG is depicted in Figure 9. KGs in a multi-level structure can be curated in two directions: bottom-up or top-down. The bottom-up approach starts on the lowest, most granular layer by forming KGs originating from seed molecules. In the case of the AIR, these seed molecules represent key initiators and mediators for the inflammatory response, such as damage-associated molecular patterns (DAMPs), pathogen-associated molecular patterns (PAMPs), receptor proteins that recognize them, and central TFs that regulate gene expression during the inflammatory phases. The molecules were identified from literature focusing on reviews describing the molecular signaling acute inflammatory processes in various cell types and tissues and databases such as Reactome and KEGG. Interactions connecting these molecules were then identified from public databases to create a comprehensive KG representation in AF format. The KG was created in CytoScape using the Bisogenet plugin [188] that connects information from several interaction databases, including DIP [189], BioGRID [190], HPRD [191], IntAct [192], and MINT [193]. Furthermore, regulatory interactions were fetched from public databases on miRNAs from miRbase [194], miRTarBase [195], and TriplexRNA [196]; TFs from TRNSFAC [197], TRRUST [198], TFactS [199], and HTRIdb [200]; lncRNAs from EVLncRNAs [201], lncRNADisease [202], lncRNA2Target [46], and lncTarD [47]. They are summarized in tab-delimited text files for storage and further computational KG processing, described in more detail in Section 2.5. The second method of KG curation is the top-down approach, where first, the top-level biological processes and cell types in each phase of the acute immune response were collected, summarized in an overview image (Figure 10), and associated molecular pathways manually curated in SBML standardized submaps. Finally, KGs generated from both approaches were merged to build the complete KG of the AIR, which, at the time of the thesis, includes 29,772 edges between 8,400 nodes. The following sections describe the top-down approach, the associated KG curation efforts, and the computerized processing of the standardized KG files in more detail.

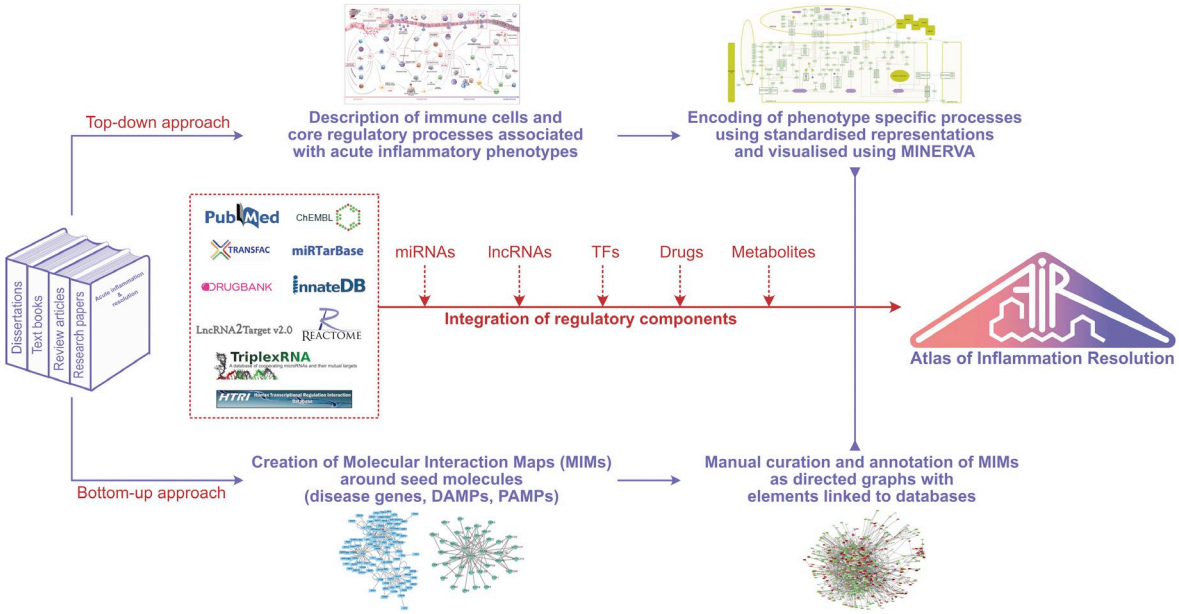

Figure 9: Workflow for constructing the Atlas of Inflammation Resolution (AIR). The AIR is constructed both bottom-up and top-down. In the case of the top-down approach, higher-level processes, phenotypes, and interplay between immune cells were identified in various stages of acute inflammation. These processes and phenotypes were extended as information flow diagrams in standard SBML notations. In the bottom-up approach, first seed molecules were identified from damage-associated molecular patterns (DAMPs), Pathogen-associated molecular patterns (PAMPs), and key disease genes associated with selected clinical phenotypes of acute inflammation. Each seed molecule is then extended with the experimentally validated interacting partners. Models generated using bottom-up and top-down approaches were later merged and integrated with experimentally validated regulatory layers, including transcription factors, miRNAs, lncRNAs, drugs, and metabolites to prepare the AIR. From Serhan & Gupta *et al.*, 2020.

2.3 Top-Down Knowledge Graph Curation

The first step in the top-down curation of the KG is the selection of the higher-level biological processes around which the KG is built. In AIR, those refer to the cellular and molecular processes involved in the initiation, transition, progression, and resolution of inflammation. A visual summary of the collected material provides curators and later users with a clear overview of higher-level information in the KG. The process of creating such an overview image for the AIR from the first sketches to the final version is depicted in Figure 10.

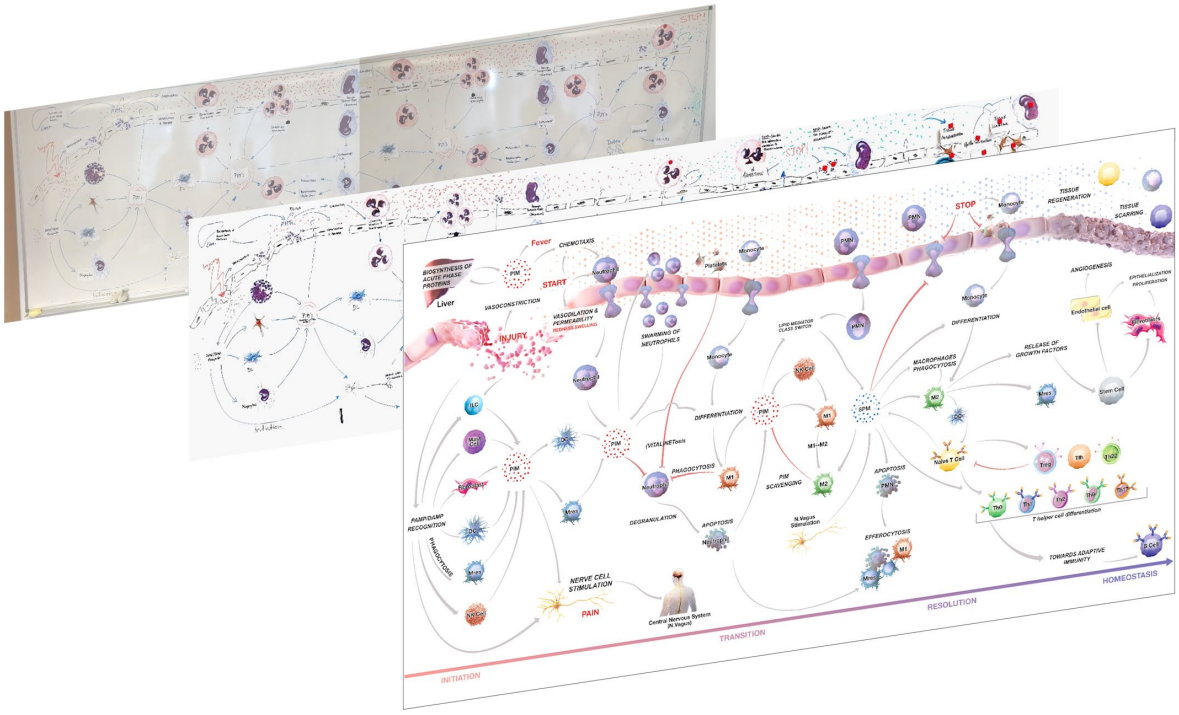

Figure 10: The development of an overview image for the Atlas of Inflammation Resolution (AIR) from the first sketches to the final version.

We performed literature research on the processes we collected in the overview image and described the molecular processes in SBML-standardized submaps using CellDesigner. The curation efforts for each submap started with screening recent literature, focusing on reviews that summarized experimental evidence on related molecular interactions into coherent pathways, then delving into experimental literature with evidence on specific reactions. In each submap, we summarized the directed interactions of biomolecules that were experimentally validated to modulate the respective process. We designed the submaps in AF format when abstraction did not lead to a loss of necessary information but retained PD formats when necessary to understand the underlying mechanisms. These mechanistic reactions include mainly the enzymatic catalysis of metabolic reactions, transmembrane transport through channels or transporters, complex formations, and essential modifications in which molecule states drastically alter their function. Connections to higher-level processes were made by representing them as an SBGN phenotype node and linking them to modulating nodes. If phenotypes are functionally related, multiple processes were included in the same submap; for example, the phenotypes “prostaglandin synthesis” and “thromboxane synthesis” both occur in the “Biosynthesis of PIMs from AA” submap since they share parts of their biosynthesis pathways. The AIR submaps were curated in CellDesigner SBML format using the CellDesigner software to create and, later, edit the files. Using the MIRIAM notation in

SBML, we provided Pubmed IDs for reactions, ChEBI IDs for metabolites, and GO IDs for phenotypes as references, if available. Proteins and genes were assigned their official HGNC symbols as names. For metabolites, their trivial name was chosen for better readability. Complexes, if available, were named with trivial names such as NFKB or by combining the names of the subunits with a colon. In addition to complexes, which represent unique entities as physical compounds of molecules, we also use complexes for groups of functionally related molecules, which we refer to as **families**. These families' group nodes are involved in the same reactions, e.g., isoenzymes, to provide more user-accessible diagram designs. We curated these family complexes as ordinary SBML complexes but with a "family" SBML tag. The first version of the AIR in Minerva was released on January 9th, 2020, and included 19 submaps. At the date of this thesis, the number of submaps has increased to 42. All submaps and each submap's number of nodes and edges are summarized in Table 2.

Table 2: Submaps included in the "Atlas of Inflammation Resolution" (AIR) Disease Map at the time of this thesis, with the number of nodes and edges.

Submap Name	Nodes	Unique Nodes	Edges
Angiogenesis	206	122	162
B Cell	177	75	44
Biosynthesis of PIM and SPM from AA	87	57	47
Biosynthesis of PIM and SPM from DHA	122	69	58
Biosynthesis of PIM and SPM from EPA	56	33	30
Chaperone mediated autophagy	97	70	27
Cholesterol Synthesis and Effects	66	60	56
Coordinated Lysosomal Expression and Regulation (CLEAR) network	461	305	341
DC cell	99	33	31
Efferocytosis	188	81	149
Fibroblasts	45	16	55
Leukocyte adhesion and transmigration	235	89	150
Lipoxins signaling	32	14	25
Lysosomal biogenesis	402	213	219
M1 Macrophage	224	32	104
M2 Macrophage	85	12	47
Macroautophagy	150	51	52
Macrophage M1M2 switch	119	24	98
Macrophages phagocytosis	264	109	169
Maresins signaling	36	10	34
Metabolic Pathways	461	399	109
Microautophagy	138	133	15
Monocyte transmigration	153	64	95
Neutrophil extracellular traps (NET)osis	78	31	77
Neutrophil apoptosis	96	45	70
Neutrophil chemotaxis	58	12	39
Natural killer cell	185	140	53
Natural killer (NK) cell chemotaxis	65	17	55

Submap Name	Nodes	Unique Nodes	Edges
PAMP signaling	263	116	110
Protectins signaling	47	15	31
Resolvins signaling	137	37	132
STOP signal	69	27	33
T cell activation	169	76	36
T follicular helper (Tfh) cell	29	11	19
T Helper 1 cell	53	9	31
T Helper 17 cell	94	27	52
T Helper 2 cell	75	13	44
T Helper 22 cell	27	11	17
T Helper 9 cell	57	18	24
Regulatory T (Treg) cell	97	20	50
Vasoconstriction vasodilation and permeability	288	147	208
Wound healing	126	80	87

2.4 Publishing the AIR as an Interactive Disease Map

When AIR was planned to be released in 2019, the DMC had already been established, with a few Disease Map projects already published, many of them utilizing the MINERVA platform [203], [204]. MINERVA was specifically built as a resource to provide functionalities for the interactive presentation of KG diagrams, combined with annotations from drug databases and tools for exploration. Considering that the AIR was intended to be a publicly available, easily accessible, and interactively explorable resource, we also decided to employ MINERVA. The AIR was published on https://air.elixir-luxembourg.org/minerva/ hosted on servers provided by the Luxembourg elixir network and supported by the MINERVA development team. The functionalities provided by MINERVA made it possible to design the AIR interface so that the multi-level layout is accessible to the user [205], [206]. Specifically, linking images to submaps and submaps to other submaps enables interactive exploration of the AIR. The starting point is the overview image (Figure 10), from which the respective submaps were either linked through clickable areas marked on the image either directly or indirectly through another linked image. Figure 11A shows all images in the AIR and how they are linked to the main overview image. This structure allows the user to zoom directly from the top phenotype layer into the process layer and explore the associated submaps.

However, while the submaps curated in the top-down approach can be visualized directly in MINERVA, the MIM at the bottom-up layer contains much more data and is not structured in standardized diagrams. At one point during the development, I developed a Python script to automatically create an SBML diagram from the KG and the CellDesigner

auto layout function from CellDesigner for visualization. To reduce the amount of data, I filtered the MIM for only those edges connected to nodes in the submaps from the top-down approach, utilizing the computational implementation described in Section 2.5. The initial idea was to have an interactive, annotated, and explorable version of the MIM available on MINERVA. However, with the increasing number and size of the submaps, the MIM diagram also grew up to the point where it became unfeasible. It also introduced a significant bottleneck in the MINERVA project upload and, finally, was removed at some point. The latest version of the MIM submap is shown in Figure 11B. Although MINERVA provides various features for exploring Disease Maps, it comes short of some functional information, such as highlighting subgraphs, especially between the submaps, a summary of edges for specific nodes in the complete KG, and linking information from the multi-level structures. The MINERVA developers encounter project-specific functionalities by enabling the development and integration of customized plugins in JavaScript that can interact with MINERVA functions, such as creating overlays or accessing annotations. Multiple plugins have already been created that work on any Disease Map in MINERVA, such as identifying disease variants in the submaps (https://minerva.pages.uni.lu/doc/plugins/disease-variant-associations/) or performing GSEA (https://minerva.pages.uni.lu/doc/plugins/gsea-plugin/). The MINERVA website provides detailed explanations and tutorials on the plugins and their dedicated API (https://minerva.pages.uni.lu/doc/) [205]. For AIR, the development of such a customized plugin was motivated by the need to enable access to the complete KG in AIR, which would make a separate display of the submap unnecessary. However, before a plugin can be developed, the KGs from both the top-down and bottom-up approaches must be processed to be computationally accessible. The next section describes the computational implementation of such conversion in detail.

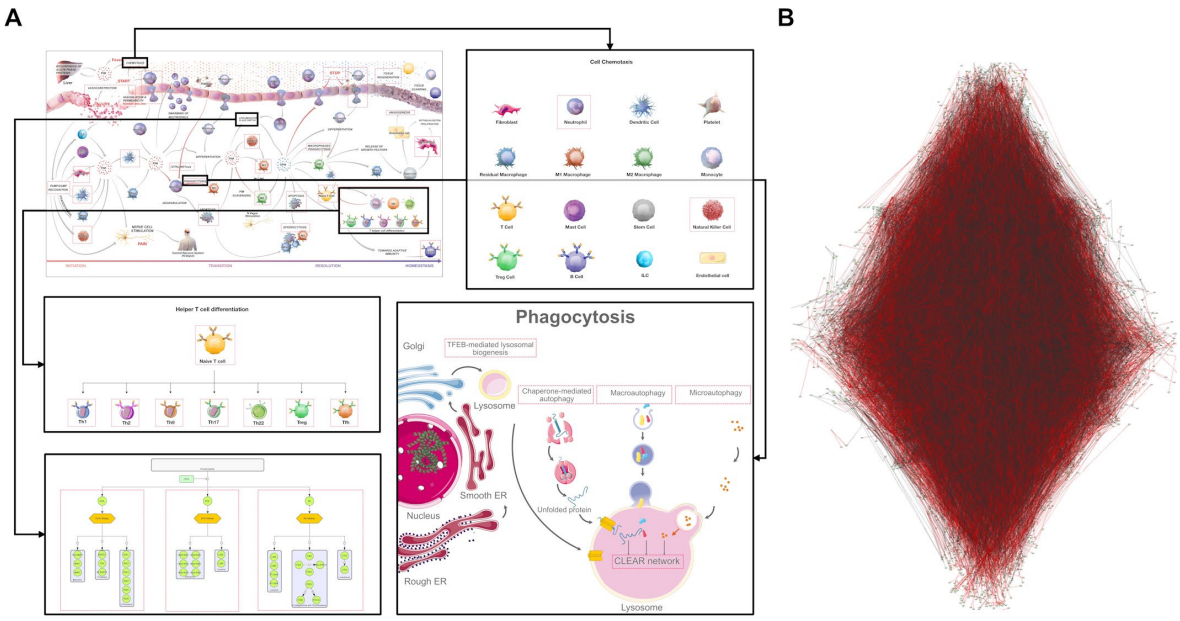

Figure 11: Design of the "Atlas of Inflammation Resolution" (AIR). (A) The overview image visually summarizes the biological processes involved in acute inflammation and its resolution. Red boxes refer to submaps that describe the molecular mechanisms underlying these processes in detail or to other overview images, which themselves contain other links. (B) The complete knowledge graph of the AIR in SBML format was removed in later versions due to its complexity and made accessible through plugins.

2.5 Object-oriented Computation of Knowledge Graph Data

Computational processing of the KG in the AIR requires converting submaps and database files into an object-oriented data structure (Figure 12). I developed a Python framework that creates such structures from supplied KG files and includes functions for their processing. Every project described in this thesis is built on and utilizes this framework. The Python files and Jupyter Notebooks showcasing the loading, processing, and analysis of KGs and data files are available on GitHub at: https://github.com/MattiHoch/KnowledgeGraphAnalysis.

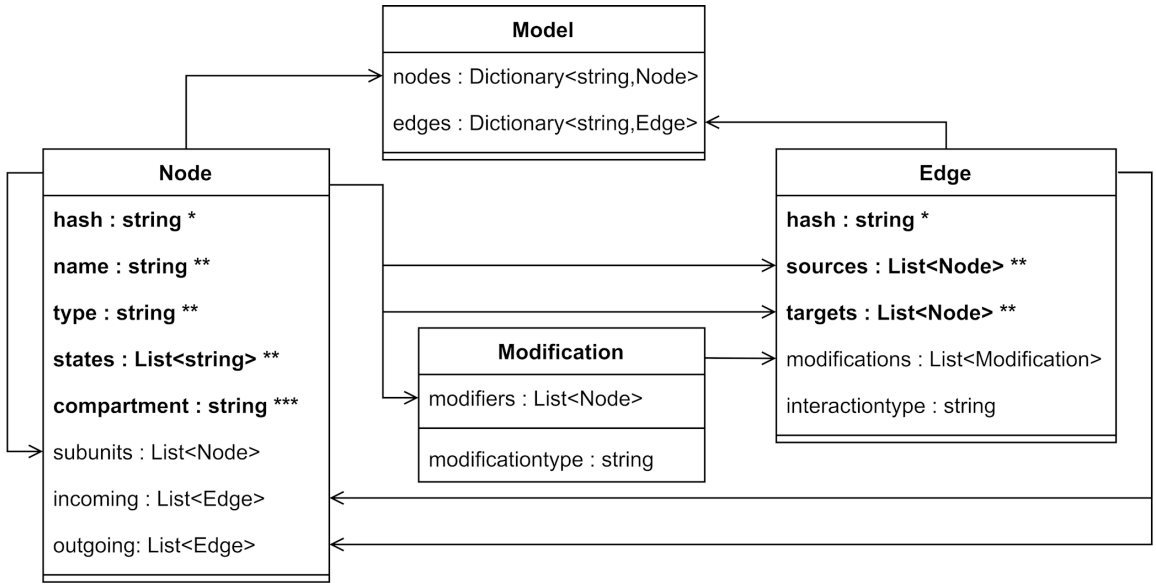

Figure 12: Class diagram of the computational knowledge graph representation in Python. All projects mentioned in this thesis use the presented structure for all Disease Map processing and analyses. * Hash attribute to uniquely identify an object. ** Attributes used to generate the unique hash. *** Unique attributes only in compartmentalized Disease Maps, such as the Sarcopenia Map or the MASLD Map.

Many KG standards, including SMBL and CellDesigner SBML format, are built on the extensible markup language (XML). Submap files can, therefore, be read by XML parsing tools. The **`create_model`** function of the **`read_sbml.py`** file employs the Python **xml.dom** library to extract the properties of all nodes, reactions, and compartments from a given list of submap files. As shown in (Figure 12), a single KG is represented by a **Model** class containing all nodes and edges as dictionaries with objects of **`Node`** and **`Edge`** classes as values and string identifiers as keys. The classes are embedded in the **`model.py`**, **`node.py`**, and **`edge.py`** files, respectively. The keys of the node and edge dictionaries in the model class are **`sha256`** hash strings generated from the attributes of nodes and edges using the **hashlib** Python library. They uniquely identify a **`Node`** or **`Edge`** object. This design allows to determine whether an object exists efficiently and, if so, to return the existing object directly. The attributes used to create the hash depend on the node or edge type and the KG context. Single nodes are identified by their name, while complexes are defined by their set of subunits. Other attributes are the node's states (e.g., phosphorylated or none) and its type (protein, gene, etc.). Some Disease Maps, however, are compartmentalized, such as the Sarcopenia Map (Chapter 6) and the MASLD Map (Chapter 7). In these maps, SBML elements from different compartments are considered distinct nodes, and thus, the hash is further generated from the string of the compartment attribute. Edges are characterized by the type of interaction (positive or negative), and the list of source and target nodes of the corresponding reaction is represented as a list of node objects, respectively, converted into

a string hash using the **hashlib** library. Consequently, reactions with the same source, targets, and type from all submaps are represented as a single **`Edge`** object. In SBML, the type of reactions (τ_e in graph representation) and type of modifications (τ_m) are described as string values using biomedical terminology. When converting the diagrams into graph representations, I used a mapping function to convert the string representation (τ^{str}) of a type τ into an integer ($\tau^{\mathrm{int}} \in \{-1,0,1\}$) value.

$$f_e: \tau_e^{\mathrm{str}} \rightarrow \tau_e^{\mathrm{int}} \\ f_m: \tau_m^{\mathrm{str}} \rightarrow \tau_m^{\mathrm{int}} \tag{2.1}$$

Edges are further extended with a set of objects from the **Modification** class, each defined by a set of modifiers, represented as a node list, and the modification type (e.g., catalysis, inhibition). When reading submap XML files, the **`name`** and **`compartment`** attributes are fetched from **species** and **celldesigner:species** XML elements. From the inner XML of the **celldesigner:complexSpecies** child element, if existing, the parent complex is identified for which it is added as a subunit. If existing, the "family" tag is extracted from the **structuralState** attribute of the **celldesigner:structuralState** XML child element.

The bottom layer information of the AIR on gene regulatory interactions of TFs, miRNAs, and lncRNAs is read from tabular files containing the data downloaded from the respective databases. The data files were formatted to display the information in process description format, as summarized in Table 3. Interaction types of the edges were fetched from the databases and, if necessary, manually assigned to integer values 1 or -1. Finally, after integrating all information, edges were added between corresponding RNA and Protein **`Node`** objects with an edge type of 1. The correspondence was determined by comparing a hash string generated from the parameters of the original hash, excluding the node type.

Table 3: Format and knowledge graph (KG) syntax of database files integrated into the complete KG of the "Atlas of Inflammation Resolution" (AIR)

KG type	Database	Source	Target	Modifier
lncRNA-Gene	lncTarD, lncRNA2Target, EVLncRNAs, lncRNADisease	Gene (DNA)	Gene (RNA)	lncRNA (catalysis)
miRNA-Gene	miRTarBase, miRbase	miRNA	Gene (RNA)	
TF-Gene	TFactS, TRNSFAC, TRRUST, HTRIdb	Gene (DNA)	Gene (RNA)	TF (catalysis)

For the application of graph-theoretical approaches, the PD representation of the AIR KG must be converted into AF format. The conversion in the computational data

structure follows the mathematical descriptions in equations (1.14) and (1.15). The conversion is facilitated by the fact that node objects have direct access to lists of outgoing and incoming edge objects through their **`sources`** and **`targets`** attributes, respectively. Such direct access to neighboring nodes further facilitates the implementation of signal flow algorithms. The edge properties are also passed down when transforming a KG from PD into AF. Modification types from PD edges are mapped to the edges in AF format, e.g., transcription factors that initially catalyzed an edge between the gene and mRNA are now directly connected to the mRNA node through an edge e with $\tau_e^{\mathrm{str}}(e) =$ "gene regulation". Keeping the information allows for functional analysis in graph theoretical approaches, e.g., stopping a pathing or signal flow algorithm at a specific edge type. The **Model** class includes an **as_adjacency_matrix()** function, which converts edges in AF format into a matrix of size $|V(G)| \times |V(G)|$. In this matrix, values of 1 or -1 indicate positive or negative interactions between the source node in the row and the target node in the column. A value of 0 indicates that no edge exists between the two nodes. Representing a KG as an adjacency matrix allows for fast processing and facilitates the application of other tools, such as the **networkx** Python package.

2.6 Knowledge Graph Exploration in MINERVA

Integrating plugins in MINERVA projects can significantly expand the functionalities of the platform. They can access information from Disease Map projects and interact with the UI to create visually appealing-colored overlays on these maps, offering a more interactive user experience. These plugins are loaded as single JavaScript files, seamlessly interacting with a dedicated MINERVA plugin API [205]. As described in Section 2.4, in the initial iteration of the AIR, we integrated the complete KG as a single SBML file, which was soon met with challenges as the file size of the KG grew to proportions that made it overwhelmingly complex and unreadable. Given these limitations and under the anticipation of developing a plugin for data analysis, we decided to use a plugin instead of a separate submap for the KG exploration. We termed this plugin AirXplore and published the associated JavaScript, HTML, and CSS files on GitHub (https://github.com/sbi-rostock/AIR/tree/master/AirPlugins). The plugin was built using the Node.js environment utilizing NPM packages for advanced user interface (UI) implementations. By publishing the files on GitHub, the raw JavaScript file of the plugin (https://raw.githubusercontent.com/sbi-rostock/AIR/master/AirPlugins/AirPlugins.js)

can be accessed and loaded in MINERVA. The plugin aims to extend the information about selected nodes in the map and improve their display through dynamic visualizations. MINERVA does not provide global information about a node, e.g., its neighbors, which edges of other submaps it is involved in, and which higher-level processes it is connected to. The plugin should display such information automatically with minimal additional user input. The plugins developed during the project, including user guides, are summarized at https://air.bio.informatik.uni-rostock.de/plugins. The following section describes the extension of MINERVA plugins for the AIR into dedicated exploration and analysis tools and the associated data infrastructure.

The first step in the development was to make the complete KG available to the plugin. As the information in the AIR is static, the KG data can be pre-processed and stored in an accessible location. I created three JSON files for the nodes, edges, and topological information using the computational AF representation of the KG described in the previous section (Figure 13A-C). The node file contains a dictionary with node ID strings as keys and dictionaries of node attributes as values (Figure 13B). The edge files consist of dictionaries in a list, each describing a single edge and its attributes between two nodes, using the node ID string as a reference to the node dictionary (Figure 13A). Another JSON file was created that contains SPs from every node to its connected phenotypes across all submaps (Figure 13C). The SPs were identified using the Breadth-first search (BFS) algorithm on the adjacency representation described in the previous section. The data files were uploaded to the same GitHub folder as the plugin JavaScript files and fetched when the user loads the plugin. The NPM packages used to develop and compile the plugin file are ChartJS for creating interactive charts, DataTables for tabular display, sorting, and filtering of information, and JSZip for zip file creation during data export. I also employed ProtVista, an interactive JavaScript tool, to visualize protein sequences, structures, and annotations directly from its source URL in the JavaScript file [86].

The current UI of the AirXplore plugin consists of four collapsable panels, each providing a separate functionality to explore or interact with the AIR. Two of these panels were included in the first version of the plugin (Figure 13D and F), while the other two were implemented later as part of the data analysis tools described in Chapter 3, specifically in Section 3.3.1. The plugins are implemented so that they respond to user actions in MINERVA. Each time a user selects a node on the map, its interactions throughout the KG are automatically displayed in multiple tables in the first panel of the plugin (Figure 13D). The user can also enter the name of any node in the KG in a case-insensitive text field. The

tables contain information about nodes of incoming edges (regulators) and outgoing edges (targets), including the node name, node type, edge type, and references describing the edge. Another table lists all phenotypes the selected node is connected to and the length and type (positive or negative) of the shortest path to the phenotype. A clickable icon in the rows of each phenotype highlights the path on the respective submap(s) starting from the currently selected node (Figure 13E). The last two tables contain additional information about the selected node, including a list of HPO terms linked to the gene from the HPO database API and an interactive panel showing the gene and protein sequence generated by the ProtVista module. The second panel contains functions for data export, either as JSON raw files, as phenotype node associations in GMT format or as CSV/TSV tables for the entire KG or specific phenotypes (Figure 13F).

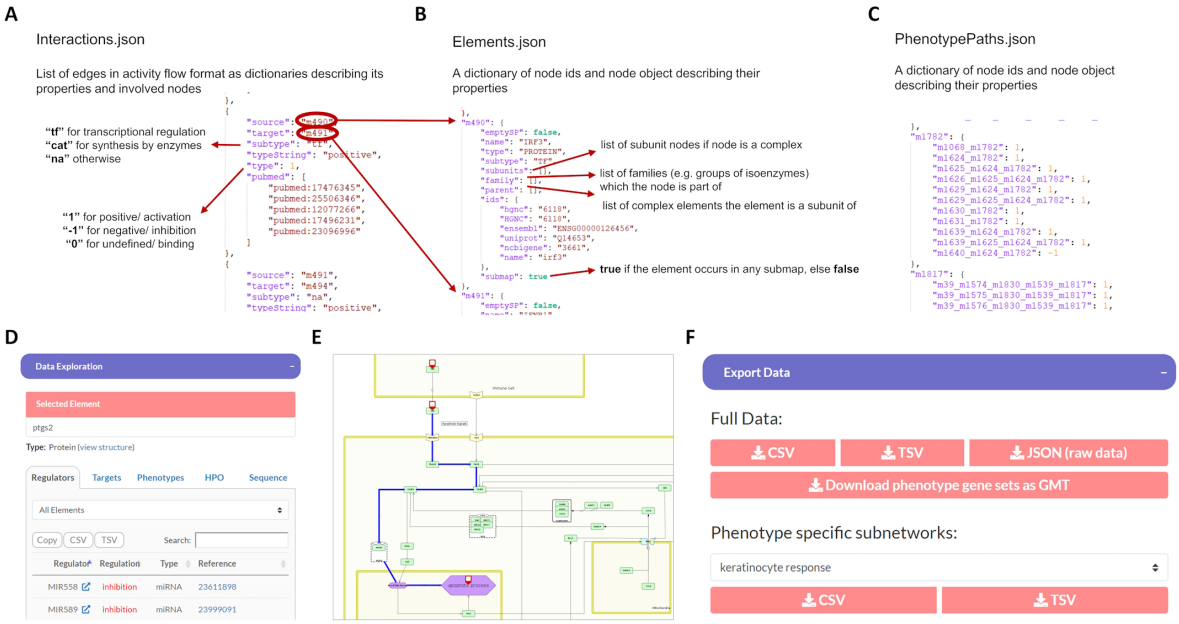

Figure 13: Specifications of JSON files that store data of the knowledge graph from the Atlas of Inflammation Resolution (AIR) (A-C) and screenshots of MINERVA plugins that enable exploration of information in the AIR (D-F).

In addition, I developed other plugins that were not designed explicitly for the AIR but intended to enhance the user's interaction with Disease Maps on the MINERVA platform. One plugin is designed to conduct topological analyses for all submaps on a Disease Map, calculating degree, betweenness, and closeness centrality for the whole Disease Map or individual submaps. The results of these analyses are visually represented, presenting centrality scores for all the nodes in intuitive interactive plots. Another plugin simplifies the user's task of data uploading by allowing multiple samples in either CSV or TSV formats, automatically creating overlays for each sample. This plugin is an

improvement from the default MINERVA upload functionality, which was perceived as somewhat cumbersome, allowing uploads of only one sample at any given time.

2.7 Summary

The foremost application of the AIR is its use as a knowledge exploration and data resource. Non-bioinformaticians, such as wet-lab researchers or clinicians, can utilize Disease Maps as web-based resources to explore the current state of knowledge in their field of research. For bioinformaticians working in related areas, MINERVA's data export functions and API make it possible to employ Disease Maps as KG resources integrated into their own workflows. Similarly, the AIR is fully open-source and can provide its content to other applications. In the years after its publishing, I collaborated on many projects that employed the AIR and presented possible applications of the AIR to interested collaborators. Specifically, in multiple projects, we published applications of the AIR to investigate inflammation-related research questions in different disease and biological contexts.

As described in Section 1.6, facilitating KG analysis requires developing publicly accessible and easy-to-use tools. They should not only ensure reproducibility but also usability for the research community. For this purpose, the plugin functionalities were extensively extended. I developed a novel approach for graph-based enrichment analysis that simultaneously allows user-friendly visualizations and gives users more insights to interpret their results. The methodology and its implementation into the MINERVA tools of the AIR are described in detail in Chapter 3. Chapter 4 describes an example of a functional graph-based analysis of scRNA-seq data using the AIR's KG.

Furthermore, I employed these tools to investigate the effects of drugs in an industrial project by integrating interactome and transcriptome data into the AIR system. The tools I developed for the AIR predicted molecular mechanisms from the individual data but also identified hidden patterns by linking them at multiple levels. My research has strongly influenced the product strategy of our industrial collaborator, providing them with a better understanding of the mode of action of their products. Based on these results, preclinical studies are planned, and it also helped to uncover novel mechanisms from previous preclinical research results. The workflow and results of the data analysis are presented in Chapter 5.

While the AIR provides a comprehensive overview of the functional organization of biological processes, it falls short when considering spatial organization. Due to the great

variety depending on tissue and stimulus, a universal definition of temporal dynamics and cellular composition during inflammation is not possible. However, the AIR could potentially tie spatiotemporal specificity with the functional non-specificity of pathway mechanisms. Given that inflammation plays a central role in many diseases [207], the specifications of other Disease Maps could be utilized as an input for the AIR to infer generalized inflammatory processes. Like a plug-and-play design, the AIR could be made adaptable to different Disease Map projects connecting to disease-specific immune reactions, such as responses of tissue-specific cell types. The AIR could then predict the effects on general immune processes, including immune cell chemotaxis, phagocytosis, and, most foremost, the modulation of inflammation resolution. This design could then be extended modularly, with processes such as chronic inflammation, adaptive immunity, or systemic (side-)effects like sarcopenia.

In general, the connection of disease maps is an important perspective for exploiting their full potential. There is a need for an automated and standardized method, for which the AutoMap workflow from the MINERVA teams provides a potential starting point (https://automap.elixir-luxembourg.org/). Combining the Disease Map effort allows for speed-up curation, identifying interactions between diseases, i.e., comorbidities, and, finally, could potentially enable the development of complete systemic models up to full-fledged virtual twins [208].

Chapter 3

Data Integration and Analysis on Large-Scale Knowledge Graphs

3.1 The 2DEA as a Novel topology-based Enrichment Approach

As discussed in Chapter 1, particularly in Section 1.6, the analysis of biomolecular data is associated with many challenges, and existing tools are limited in their general usability and the interpretability of results. The limitations of existing enrichment methods in not including the full information available in the data and KG (as mentioned in Section 1.4.3) motivated the development of a novel approach for causal reasoning on large-scale KGs and Disease Maps. Analytical tools are needed to support the research community through improved usability and accessibility. Under these considerations, I developed a novel enrichment analysis approach employing information from KG methods. The methodology has been described in detail in the publication "**Network- and enrichment-based inference of phenotypes and targets from large-scale Disease Maps**" [209]. I presented a two-dimensional, topology-based enrichment analysis (**2DEA**) approach that combines topology and data-integration methods to derive information from complex, large-scale KGs such as those from Disease Maps (Figure 14). This approach further enables workflows for data analysis and interpretation that can be navigated even by individuals without a background in bioinformatics. In this way, systems biology approaches can broaden the accessibility and applicability of biomolecular data science.

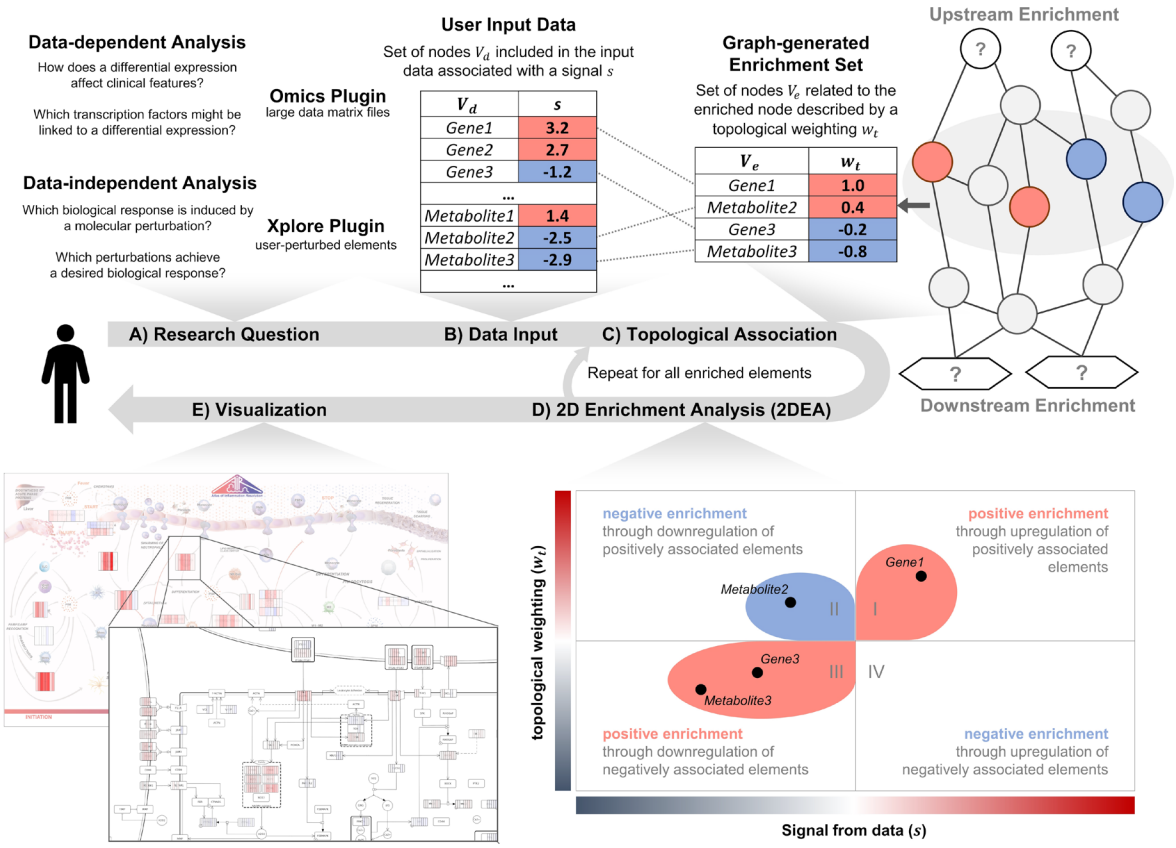

Figure 14: (A) I developed a plugin for the MINERVA platform that allows user interaction and performs *in silico* perturbation analysis on Disease Maps. Depending on the research question, perturbed nodes come either from large experimental data files (Omics plugin) or from nodes on the map individually selected and perturbed by the user (Xplore plugin). (B) In both cases, the inputs can be viewed as a list of nodes ($\boldsymbol{V_D}$) characterized by a signal s, such as an FC value. (C) The nodes of $\boldsymbol{V_D}$ are mapped to nodes $\boldsymbol{V_e}$ the Knowledge Graph (KG) of the Disease Map that is related to (downstream) or from (upstream) the node to be enriched, represented by a topological weighting $\boldsymbol{w_t}$. (D) The 2DEA then statistically evaluates whether the combination of signals and topological weightings is overrepresented towards positive enrichment (same direction) or negative enrichment (opposite direction). (E) Enrichment scores, signals, and topological weightings can be presented intuitively to the user as colored overlays on standardized diagrams and images in MINERVA. From Hoch *et al.*, 2022.

3.1.1 Topological Weighting

As described in Section 1.4.3, enrichment analysis evaluates the overrepresentation of an input list, i.e., nodes V_d mapped from the data, and a set $V_e(v) \subseteq V$ associated with the to be enriched node v. In the 2DEA, V_e is defined by a topology-based weighting factor (w_t) between a node $u \in V$ and v (Figure 14C). The calculation of w_t generally is based an topological parmeters connecting u and v, however, differs for downstream ($w_t(u, v)$) and upstream enrichment ($w_t(v, u)$).

For the AIR, downstream enrichment is solely performed on phenotype nodes, whose modulation is described by the individual submaps. Thus, the phenotype weighting $w_{t,P}$ for nodes that occur in the same submaps as a phenotype node v is calculated differently than for those that do not occur in the submaps. The sets of nodes and edges

originating from the same submaps as v are denoted as $V'(v) \subset V(G)$ and $E'(v) \subset \mathrm{E(G)}$. If $u \in V'$, its weighting is calculated based on the percentage of nodes and paths connected with u. Let's denote $\mathcal{P}_{\to v}$ as the set of all paths to v and of these, $\mathcal{P}_{\to u \to v} \subset \mathcal{P}_{\to v}$ as those that go through u, as well as $V'_{\to v} \subset V'(v)$ as the set of nodes connected to v, and of these, $V'_{u \to v} \subset V'_{\to v}$ as the nodes on paths from u to v. The percentage of paths going through u ($\frac{|\mathcal{P}_{\to u \to v}|}{|\mathcal{P}_{\to v}|}$) is similar to the betweenness centrality $c_B(u)$ from Equation (**1.21**) prioritizing nodes that merge incoming signals and are, thus, integral to the phenotype modulation. The percentage of nodes on paths outgoing from u ($\frac{|V'_{u \to v}|}{|V'_{\to v}|}$) can be comapred to the closeness centrality $c_C(u)$ from Equation (1.22), which prioritizes nodes with initiating or modulating input for the pathway. As both types of nodes might be of high interest to the researcher, they are combined in the topological weighting and signed by the type of shortest path between u and , which is defined as:

$$w_{t,P}(u,v) = \tau(\sigma(u,v)) \cdot \left(\frac{|\mathcal{P}_{\to u \to v}|}{|\mathcal{P}_{\to v}|} + \frac{|V'_{u \to v}|}{|V'_{\to v}|} \right) \tag{3.1}$$

For every node $u \notin V'(v)$ not part of the submaps of v, its weighting is calculated based on the accumulated weightings of nodes $V'(v)$, adjusted by their shortest path.

$$w_{t,P}(u,v) = \sum_{i \in V'(v)} \left(w_t(i,v) \cdot \frac{\tau(\sigma(u,i))}{2^{\ell(\sigma(u,i))}} \right) \tag{3.2}$$
$$with \ |w_t(u,v)| \not> max\{|w_t(i,v)| \mid i \in V'\}$$

Finally, the weightings are normalized by the maximum absolute value, limiting them between -1 and 1. The set of nodes with a non-zero centrality weighting on a phenotype v is, apart from the enrichment set V_e, also referred to as the modulators V_m of v:

$$V_m(v) = V_e(v) = \{u \in V \mid w_t(u,v) \neq 0\} \tag{3.3}$$

Figure 15 exemplary shows the topological weightings calculated from Equation (3.1) for the phenotype "prostaglandin synthesis" visualized on the "Biosynthesis of PIM and SPM from AA."

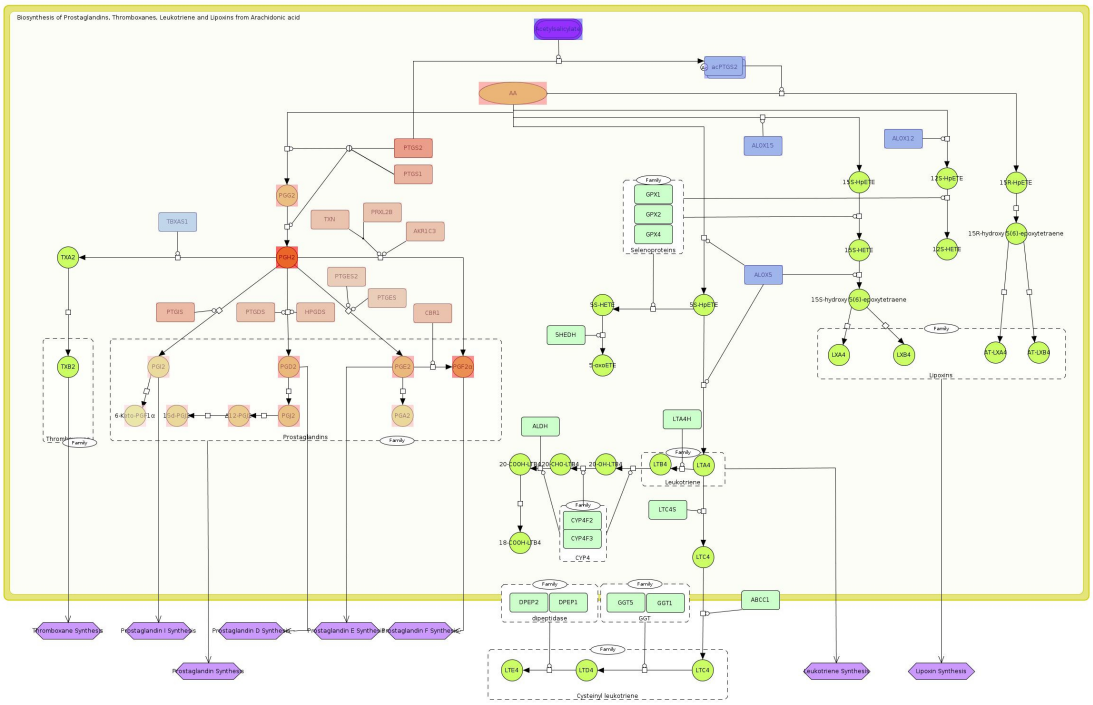

Figure 15: The "Biosynthesis of PIM and SPM from AA" submap with highlighted topological weightings for all nodes modulating the prostaglandin synthesis phenotype (bottom left). Red: positive weighting, Blue: negative weighting. Hub metabolites, e.g., PGH2, and key enzymes, e.g., cyclooxygenase 2 (PTGS2), are the highest weighted nodes. Enzymes that metabolize elements from the pathway have a negative weighting. From Hoch *et al.*, 2022.

While calculating $w_{t,P}(u,v)$ itself is an all-path problem, an adapted Depth-First-Search (DFS) can be leveraged to multiplicatively update stored numbers of nodes and paths when stopping at already visiting nodes. Algorithm 1 shows the Python code to calculate the $\left(\frac{|\mathcal{P}_{\rightarrow u \rightarrow v}|}{|\mathcal{P}_{\rightarrow v}|} + \frac{|V\prime_{u \rightarrow v}|}{|V\prime_{\rightarrow v}|}\right)$ part of $w_{t,P}(u,v)$ for the 2DEA from the data structure described in Section 2.5. The dictionary **nodes_on_paths** records which nodes are on the paths leading to each node, while **paths_with_node** counts the number of paths passing through each node. As the DFS traverses the graph, it keeps track of the current path and updates **nodes_on_paths** and **paths_with_node** for the discovered node and all other nodes it has past on its current path. The KG, however, had to be adapted for enzymatic reactions. Their enzyme nodes can not be included as intermediates between the source and the target during the DFS walk, as shown in Equation (1.16). Otherwise, all substrates would be connected to all products for all reactions the enzyme catalyzes and might connect to paths their real products are not part of. To account for this problem, substrate and enzyme nodes are both connected directly to the targets via separate edges (as described for general modifications in Equation (1.15)) while maintaining the negative feedback of the enzymes to the substrates from Equation (1.16). The **paths_with_node** and **nodes_on_paths** for the enzyme node are updated when the DFS traverses through the

respective substrate and product pair. In Algorithm 1, the enzyme nodes are stored in the adjacency dictionary of the respective source node as a node list **`skipped_nodes`**. The type of shortest path $\tau(\sigma(u,v))$ is calculated seperately through a BFS.

In contrast to downstream enrichment, upstream enrichment is used to identify nodes whose regulated targets are overrepresented in the input list. The input data, e.g., empirical observations, are assumed to be an outcome that the enriched nodes might be likely to trigger or counteract. The selection of upstream nodes depends on the context and nature of the data and is not limited to a specific molecule type. In the special case of enriching nodes upstream to phenotypes, i.e., changes in phenotypes being the input data, then $w_{t,R}(v,u) = w_{t,P}(v,u)$. Otherwise, it could refer to nodes linked to TF in a transcriptomics experiment or enzymes catalyzing metabolic reactions in a metabolomics experiment. The weighting generally describes causalities from the upstream nodes to each $u \in V$, as by their type of interaction if v is a direct regulator of u ($V_R(u)$) or, otherwise, by the shortest distance to all $V_R(u)$. The definition of V_R depends on the context and represents, for example, in gene regulation data, the set of TFs targeting a node u, defined as $V_R(u) = \{v \in V \mid \tau_e^{\mathrm{str}}(v,u) = \text{"}gene\ regulation\text{"}\}$.

$$w_{t,R}(v,u) = \begin{cases} \tau(\sigma(v,u)) & \text{if } v \in V_R(u) \\ \sum_{i \in V_R(u)} \left(w_t(v,u) \cdot \frac{\tau(\sigma(i,u))}{2^{\ell(\sigma(i,u))}} \right) & \text{otherwise} \end{cases} \tag{3.4}$$

Finally, w_t is normalized by its absolute max value, thus being limited to -1 or 1. Since the actual enrichment (see next section) is independent of how w_t is calculated, it can generally be based on any topological information, e.g., $w_t(u,v) = \frac{\sigma(u,i)}{\ell(\sigma(u,i))}$. This way, 2DEA can be easily adapted to other KG specifications and research questions.

3.1.2 Data Integration and Enrichment

As described in Section 1.3, data integrated in KG generates a set of nodes V_d with associated signals s that represent values form the data, such as FC values (Figure 14B). Given the definitions from the previous section, the set of nodes that can be used to enrich another node v is $V_d \cap V_e(v)$. The 2DEA can be thought of as evaluating the distribution of data points in a 2D space, hence the name (Figure 14D). The following section describes the methodology of the approach in detail, visually guided by the concept of a 2D plot (Figure 16). Both the signal from data and the topological weighting are plotted on the x- and y-axis, respectively, defining the set of points in (x,y) format as $\{(s(u), w_t) \mid u \in V_d \cap V_e(v)\}$

with $w_t = w_{t,P}(u,v)$ for a downstream enrichment, and $w_t = w_{t,R}(v,u)$ for upstream enrichment (Figure 16A). Based on the sign of both values, the predicted effect on v is either positive (both values have the same sign) or negative (both values have opposite signs). The enrichment thus needs to evaluate whether the points are prominently oriented towards the first and third (positive) or second and fourth (negative) quadrants. Finally, the approach requires representing this distribution as a numerical score and, through permutation, asses its significance from a standard distribution. Like the GSEA, I refer to this numerical score as Enrichment Score (ES). I calculated ES as the slope of a regression line that passes through the origin (zero) of the plot. This way, ES > 0 for distributions in the first and second, and negative in the second and fourth quadrants. However, the slope increases with higher y-values and lower x-values, giving much more significance to the centrality weighting and can sometimes cause high signal values on the x-axis to reduce the ES. To ensure that ES is bound between -1 and 1 and also is equally influenced by both values, I mapped the points onto both diagonals, as shown in Figure 16B, by multiplying the x and y values with the absolute value of their counterparts, respectively, defining the mapped points as:

$$\{(s(u) \cdot |w_t|, w_t \cdot |\mathrm{s}(u)|) \mid u \in V_d \cap V_m(v)\} \tag{3.5}$$

The remaining issue is that the slope is independent of the number and value of the points if all are on the same diagonal, constantly being 1 or -1 and failing to assess the statistical significance. As a solution, I added two points, (-1,0) and (1,0), that act as a counterweight onto the x-axis, forcing ES towards 0. This way, a score closer to -1 or 1 indicates a strong alignment with either of the 45° diagonals, suggesting a distribution across the respective quadrants. The general form of a regression line is $y = mx + b$, but when the line is forced through the origin, the intercept b is zero, simplifying the equation to $y = mx$. For a set of points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the slope m of the regression line can be calculated as:

$$m = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{3.6}$$

Integrating the mapped points from equation (3.5) and the two counterweight points defines the ES to enrich a node v as:

$$ES(v) = \frac{\sum_{u \in V_d \cap V_m(v)} \left(|w_t \cdot s(u)| \cdot w_t \cdot \mathrm{s}(u)\right)}{2 + \sum_{u \in V_d \cap V_m(v)} \left({w_t}^2 \cdot \mathrm{s}(u)^2\right)} \tag{3.7}$$

In addition to the ES, the absolute **level** for each enriched node v can be calculated by aggregating the weighted signals of all nodes denoted as the phenotype signal s_p (Equation 4.6).

$$s_p(v) = \sum_{u \in V_d} w_t \cdot s(u) \tag{3.8}$$

Its purpose is to have a quantifiable score that allows for comparing the strength of enrichment, e.g., representing the predicted activity of a higher-level process across samples. It is important to note that because the level is arbitrary, comparing it with the level of other enriched nodes is unreasonable. Additionally, the appraoch should provide information on the saturation of the enriched node in the sample, calculated as the weighted percentage of modulators that are DCEs (Equation 5).

$$\text{Saturation}_v[\%] = \frac{\sum_{u \in V_d \cap V_m(v)} (\,|w_t(u,v)|)}{\sum_{u \in V_m(v)} (\,|w_t(u,v)|)} * 100\% \tag{3.9}$$

In upstream enrichment, the enriched nodes can either be positive, affecting nodes in the data according to the sign of their values, or negative, having the opposite effect. Both may be of interest to the user, as suppression of positive or activation of negative upstream nodes (or vice versa) serves the same purpose. Calculating an aggregated level for the upstream enrichment, like in the downstream enrichment, may biologically not be very meaningful, as targets are regulated in parallel. Thus, in addition to the ES, instead of the level, the nodes are ranked according to their sensitivity (= true positive rate, i.e., ability to affect V_d) and specificity (= true negative rate, i.e., ability not to affect $V \setminus V_d$).

$$\text{Se}(v) = \frac{\sum_{u \in (V_m(v) \cap V_d)} (\,w_t(v,u) \cdot s(u))}{\sum_{u \in (V_m(v) \cap V_d)} (\,|\text{FC}_u|)} \tag{3.10}$$

$$\text{Sp}(v) = \frac{\sum_{u \in (V_m(v) \setminus V_d)} (\,1 - |w_t(v,u)|)}{\sum_{u \in (V_m(v) \setminus V_d)} (\,1)} \tag{3.11}$$

Sensitivity is greater than zero for positive targets and less than zero for negative targets. Sensitivity (= true positive rate, Equation 8) will be 1 (= positively enriched node) if $w_t(v,u) = 1 \; \forall \; u \in V_d$. For example, a predicted target with a sensitivity of 1 in a transcriptomics experiment refers to a transcription factor that directly induces the expression of all DEGs with a positive FC value and represses the expression of all DEGs with a negative FC value. Conversely, the sensitivity will be -1 (= negatively enriched node) if $w_t(v,u) = -1 \; \forall \; u \in V_d$. Specificity (= true negative rate, Equation 9) will be 1 if $w_t(v,u) = 0 \; \forall \; u \in V \setminus V_d$.

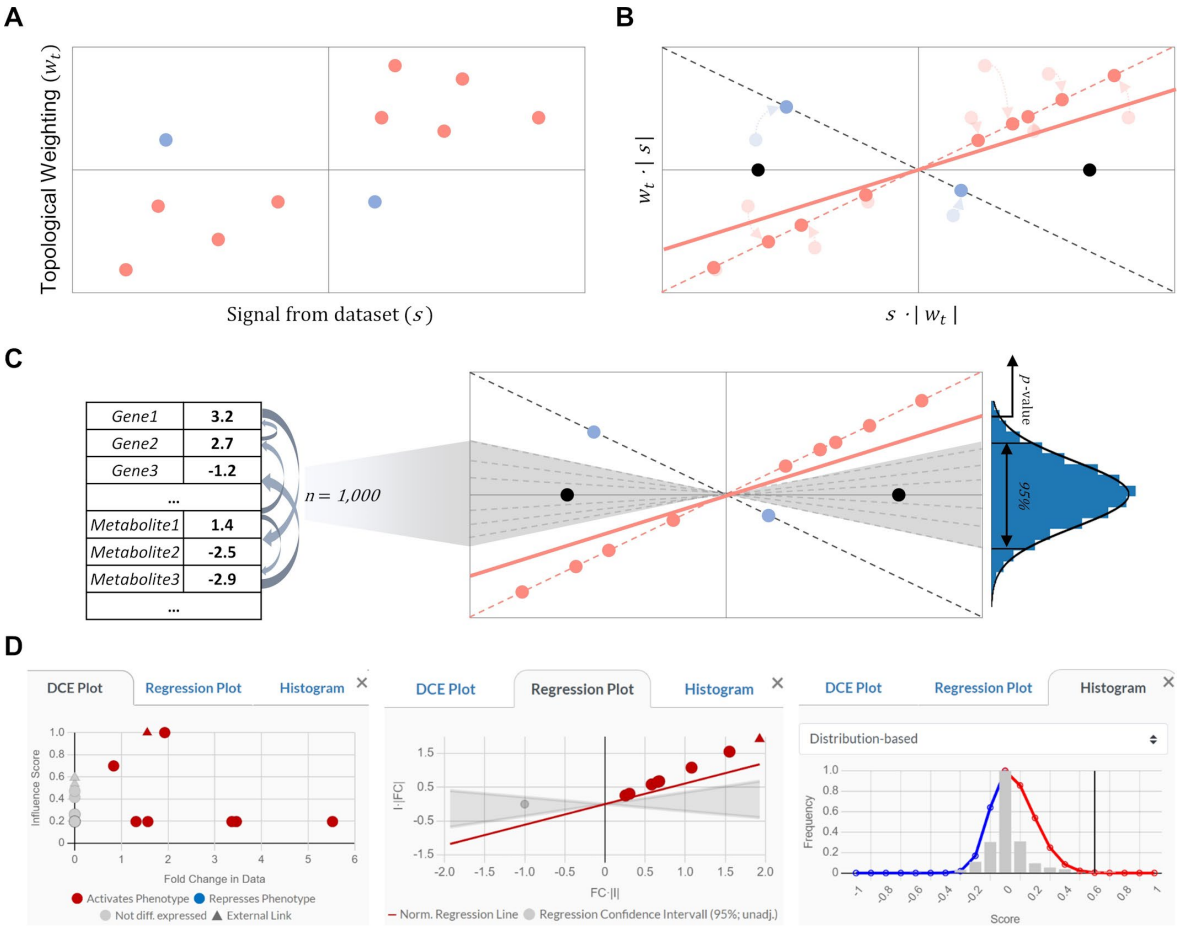

Figure 16: Visual representation of the enrichment score (ES) calculation in the 2DEA. (A) *Log2* fold (FC) change values of entries in the input data and their topological weightings ($\boldsymbol{w_t}$) generated from the knowledge graph (KG). (B) To normalize their distribution, all points are shifted on the diagonals with slopes of 1 and -1 (dotted lines). The ES is defined as the regression line's slope through the origin (red line). Two baseline points (black) are added as a counterweight, forcing the regression towards the x-axis, making the ES dependent on the number of nodes, and ensuring normal distribution. (C) Recalculating ES for randomized input lists (dotted lines) identifies its statistical significance, thus creating a reference null distribution around the x-axis. (D) User interface screenshots of the AIR plugins show how statistical features are interactively presented for each result. From Hoch *et al.*, 2022.

3.1.3 Statistical Evaluation

Statistical analysis of enrichment analysis that integrates continuous parameters from the data is challenging. Biomolecular data, such as gene expression levels or corresponding FC values, often do not follow a normal distribution. They require non-parametric tests, calculating p-values through permutations of the data. Therefore, to identify the statistical significance in the 2DEA, I generated n randomized permutated sets by either permutating V_d generating $V_{d_r} = \{V_{d_1}, V_{d_2}, \ldots, V_{d_n}\}$ or permutating $V_e(v)$ generating $V_{e_r}(v) = \{V_{e_1}(v), V_{e_2}(v), \ldots, V_{e_n}(v)\}$ with $n = 1000$ by default (Figure 16C). Some enrichment studies suggest using term label perturbation instead of gene list permutation to avoid skewing co-expression relationships in the data and thus produce more biologically accurate null

distributions [210]. I opted for a permutation of V_d because 2DEA as a topology-based approach employs co-regulations in the KG that would introduce a similar skewing to those from the data. Secondly, in most cases, the number of samples will be less than the number of enriched nodes (phenotypes), making the enrichment less computationally expensive. While this would normally not make a major difference in calculation times, it becomes relevant when implementing the method in JavaScript for the plugins. The permutated sets are generated for each sample in the supplied data, maintaining the same size as V_d or $V_m(v)$ through stratified permutation, i.e., randomizing $s(u)$ or $w_t(u, v)$, respectively, among all $u \in V(G)$ of the same type (e.g., genes or metabolites). Either way, enrichment scores are calculated for each permutated sample $ES_R = \{\mathrm{ES}_1, \mathrm{ES}_2, \ldots, \mathrm{ES}_n\}$ and levels $Level_R = \{Level_1, Level_2, \ldots, Level_n\}$. The normal distribution may differ for positive and negative ES values because the topological weightings and FCs are not evenly distributed between positive and negative values. Therefore, Gaussian curve fitting determines a separate half-distribution for each direction (positive and negative values). Then, from the standard deviation σ and mean μ of the identified distribution, the z-score for the original ES or Level is calculated as:

$$z - score = \frac{ES - \mu(ES_R)}{\sigma(ES_R)} \quad or \quad z - score = \frac{s - \mu(Level_R)}{\sigma(Level_R)} \tag{3.12}$$

The p-value represents the probability of achieving the same or higher absolute ES or Level than in the original V_d by random. Finally, adjustment for multiple testing among all enriched nodes in each sample was performed through false discovery rate (FDR)-correction by *Benjamini–Hochberg* to generate adjusted *p*-values [211].

3.2 Comparison with Established Approaches

2DEA distinguishes between up- or downregulation of positively- or negatively associated elements by combining information from data with the topology-weighted relationship to the element that will be enriched. Thereby, 2DEA can statistically evaluate whether an enriched up- or downstream element is positively or negatively enriched in the input data. Other enrichment approaches usually do not or only partially include this information, as shown in Table 4.

Table 4: Comparison of 2DEA with other established enrichment approaches.
* The approach has not been applied for up- or downstream analysis, but it is theoretically possible.

			Information from the input list		Relationship between the input list and the enriched element	
	Down-stream	Up-stream	Fold Change direction	Fold Change value	Regulatory Direction	Regulatory weighting
ORA[212]	✓	X*	X	X	X	X
GSEA[210]	✓	X*	✓	✓	X	X
RCRA[120]	X*	✓	✓	X	✓	X
IPA[127]	✓	✓	✓	X	✓	X
ROMA[213]	✓	X*	✓ (PCA analysis of expression matrices)		X	X
PADOG[214]	✓	X*	X	X	X	✓
Weighted GSEA[119]	✓	X*	✓	X	X	✓
BD-Func[118]	✓	✓	✓	X	✓	X
2DEA	✓	✓	✓	✓	✓	✓

To show how differences in integrated information affect the results of enrichment approaches and their interpretability, I compared 2DEA with GSEA [210] in a case study analyzing a bulk tissue RNA-seq dataset from a murine colitis model [215]. As an input list for both enrichment approaches, I identified significant DEGs in all eight samples. For every sample, I applied 2DEA and GSEA to enrich all 42 phenotypes in the AIR. The gene sets associated with each phenotype were the same for both approaches, i.e., all nodes within the AIR KG that have a weighting on the enriched phenotype that is non-zero. I then selected three enrichment results: one significant only in GSEA, one in both approaches, and one only in 2DEA. Figure 17 shows the output graphs of both approaches for each selected result. In Figure 17A, the enrichment by GSEA, but not by 2DEA, is significant. Although upregulated DEGs are overrepresented (the left side of the GSEA panel and the right side of the 2DEA panel), these DEGs have ambiguous effects (similarly distributed positive and negative w_t). GSEA does not include information on the relationship between DEGs and enrichment phenotype, thus identifying a significant overrepresentation of upregulated DEGs. In Figure 17B, DEGs are upregulated, but all have positive weightings, so 2DEA and GSEA identify significant enrichment. In Figure 17C, GSEA predicts non-significant enrichment when upregulated and downregulated nodes are equally represented. However, the upregulation of DEGs with positive w_t and the downregulation of DEGs with negative w_t can be considered the same result and vice versa. The 2DEA

shows its strength by accounting for these correlations and allows such cases to be predicted as significant.

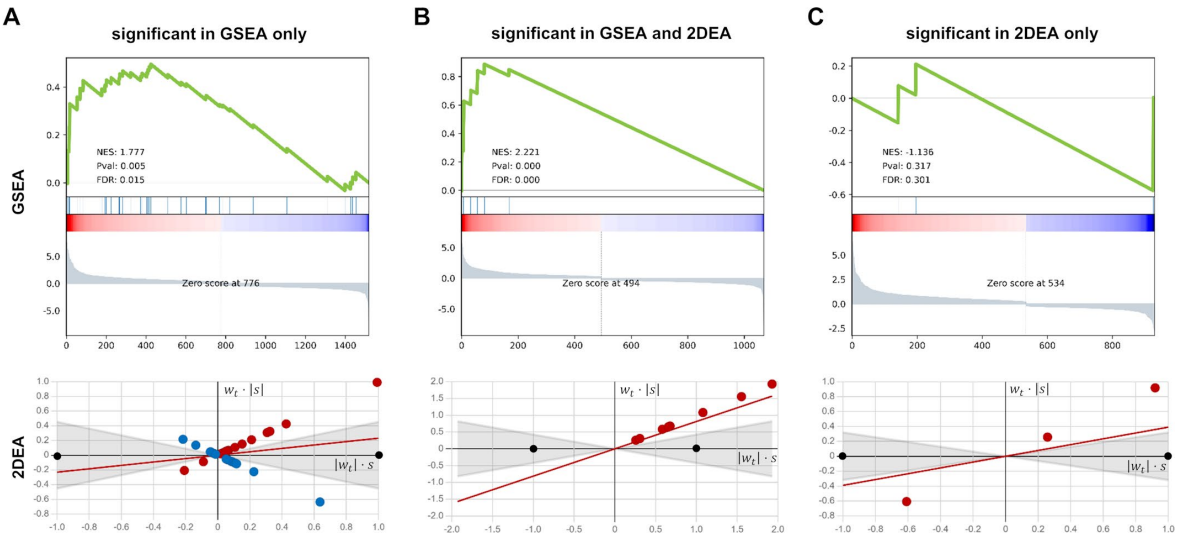

Figure 17: I employed GSEA and 2DEA to identify enriched phenotypes in AIR from an RNA-seq dataset of differentially expressed genes (DEGs; adj. p-value < 0.05) generated by DESeq2 as an input gene list. Three results were selected that were significantly enriched in GSEA only (A), in both approaches (B), or 2DEA only (C). In the panels of GSEA, a running sum of enriched scores is generated over the list of DEGs, ordered by their log2 fold change (FC) value from upregulated (red, left) to downregulated (blue, right). In 2DEA, normalized signals ($\boldsymbol{s}$, x-axis) from the data, such as FC values, are linked to topology-based weightings ($\boldsymbol{w_t}$, y-axis) to identify the distribution of DEGs in either the direction of positive (red) or negative (blue).

3.3 Integrating the 2DEA into Disease Map Tools

Finally, I integrated the 2DEA approach into the MINERVA plugins for the AIR. The plugin design needed to distinguish between simulations of minor disturbances by users (hypothetical data) with more exploratory intentions and those that analyze empirical data, which differ in complexity and level of detail. I decided to integrate the hypothetical data analysis into the AirXplore plugin as it fits the idea of Disease Map exploration, where users can perturb selected nodes with a few clicks and observe the impact on the KG. For analysis of empirical data, however, I developed a new plugin called AirOmics, which allows the uploading of large data files and offers detailed customization of the analysis and a detailed presentation of the results. Consequently, I redesigned the single plugin into a tool suite that includes AirXplore and AirOmics as separate tools that can be selected via a navigation tab bar. This solution allows the extension of the plugin with independent JavaScript files for each tool that can share data by exposing its objects globally. The following sections describe the integration into the AirXplore tool and then explain the functionalities for large-scale data analysis using the AirOmics tools in more detail.

3.3.1 AirXplore - Enabling User Perturbations

I implemented the 2DEA approach in the AirXplore tool to facilitate *in silico* simulations with non-empirical input data. Users may wish to observe the effects of selected molecular perturbations or identify specific regulators of certain phenotypes. Preparing and uploading individual data files takes unnecessary time and interferes with streamlined data analysis. To solve these problems, I integrated intuitive UI elements into the tool to define perturbations and enable immediate presentation of results. Two additional panels have been added to the UI, one for downstream and one for upstream enrichment (Figure 18A-B). In the former, the user enters a list of node names separated by a comma or inserts the currently selected node into the map via a button (Figure 18A). The nodes then appear in a table with a slider that can be used to select a perturbation value as a pseudo-FC value between -1 and 1. Any change to the perturbation automatically performs the 2DEA and displays the predicted values and adj. p-values in a table. The second panel enables upstream enrichment and works similarly, but the user selects the levels of any phenotype in the map, already listed in a table, using sliders (Figure 18B). Similar to upstream enrichment, any change automatically performs the enrichment, but in this case, the results for all upstream mediators are visualized in a graph, with specificity on the x-axis and sensitivity on the y-axis.

3.3.2 AirOmics - Analyzing Empirical Data

When developing the plugin, I ensured that it offered much flexibility to prevent the user from extensive and unnecessary formatting of the files. The AirOmics tool accepts a preprocessed data file containing signal values and/or p-values from multiple samples (columns) for any biomolecular probes (rows). The file is expected to be in a tabular format, with or without headers. One column contains the row identifiers, and the others contain the values for each sample, with samples iterating over the columns. The plugin automatically recognizes the file specifications, such as headers and separators, but the user can also set them manually. The first column is assumed to contain the identifiers; otherwise, the user can manually select the identifying column. If p-values are included, they must be in a separate column after the FC value column of the same sample. The plugin attempts to detect whether p-value columns are contained in the data by searching for an occurrence of the string "**`pvalue`**" and similar variations in the column headers and checking whether the number of columns minus one for the identifier column is even. After the initial

processes, the file is mapped to nodes in the KG. At the start of the plugin, node IDs from the knowledge graph are associated with MINERVA objects based on the same rules as the KG generated from the submap. This association also provides nodes with the dataset IDs from MINERVA's automated annotation function. Requiring performing this association on every plugin load is a significant limitation of the pure front-end design and is further discussed in Section 3.3.3. In case multiple IDs in the dataset map to the same node, the user can select how the information is merged, either by mean value, highest/lowest signal, or highest/lowest p-value. Any values in the data that are not readable as a number are assigned 0.0 for the FC and 1.0 for the p-value.

The downstream enrichment is the default active window after initializing the data. Usually, the analyses can be started directly by clicking the button, with settings optimized for large datasets with many probes. In the advanced settings drop-down menu, different settings can be specified to fit the parameters of the analysis (Figure 18D). All results are directly displayed in a table (Figure 18C) of columns for each sample containing the estimated value (between -1 and 1) and the p-value in parentheses for each u in the rows. Clicking on the column header, i.e., the sample name, will instantly color the phenotypes with their estimated levels (blue for negative, red for positive values) on the submaps if the p-value is lower than the supplied value. Clicking on a single value itself will pop up a new graph plotting the as shown in Figure 16C. Clicking on a value will show a scatter plot in a popup window with detailed information on the phenotype's modulators, their signal in the data, their topological weighting $w_{t,P}$, and statistical information. Additional information in the table includes (i) the number of nodes in $|V_d \cap V_e|$ as the mean [+ std. dev.] among all samples, (ii) the weighted percentage $\frac{\sum_{u \in V_d \cap V_e(v)} |w_{t,P}(u,v)|}{\sum_{u \in V_e(v)} |w_{t,P}(u,v)|}$ in all samples, and, (iii) the top five nodes ranked by their combined signal and weighting among all samples and phenotypes. The next panel (Figure 18E) allows users to customize the visualization of the estimated phenotype levels as overlays in MINERVA, with the possibility of visualizing the signal values from the data. By setting a phenotype threshold, the user can decide which phenotypes to include in the overlays. The name of the overlay is equal to the sample name in the data, with a user-specified suffix to distinguish multiple analyses (e.g., with different p-value thresholds) from one another. There are buttons to hide, show, or delete all overlays automatically. The node ranking panel (Figure 18F) shows how much nodes in the KG contributed to the results based on their topological weightings and signal values calculated as $\sum_{u \in V_d \cap V_e(v)} \left| w_{t,P}(u,v) \cdot s(v) \right|$. The overview is for one sample only, which the user can

select. The ranking is presented as a horizontal bar plot, with one bar for each node, showing their relative importance in percentage. Additionally, it can be specified whether only significant phenotypes (based on the p-value threshold defined in the table panel) will be considered. Its content can be downloaded in JSON format.

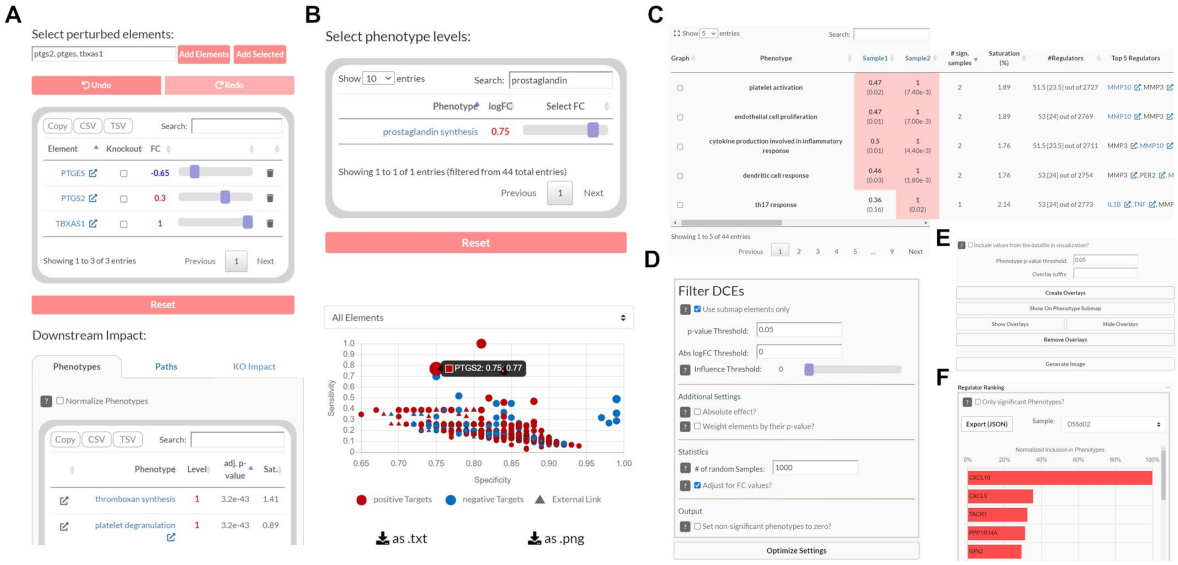

Figure 18: User interface of the Atlas of Inflammation Resolution (AIR) tools for downstream enrichment analysis from large-scale data files. (A) While the standard analysis can be carried out directly, users can make additional settings, e.g., filtering the data. (B) The results are presented in a table showing levels, p-values, and other details on predicted phenotypes. (C) Overlays highlighting the phenotypes and values predicted from the user data can be created by coloring the corresponding nodes on the maps, allowing for intuitive visualization of the results. (D) In each sample, probes from the data, i.e., genes or metabolites, are ranked by their impact on the analyses, providing an overview of potentially highly relevant data patterns.

The upstream enrichment tool predicts nodes in our dataset that may be the most probable modulators for the data sample's observations (i.e., FC values). Highly potential enriched nodes interact with nodes from the dataset (V_d) in the same way as their signal (positively enriched node) or the opposite way (negatively enriched node). To perform the analysis, the user selects the sample and, if desired, filters the enriched nodes by their molecule type (e.g., transcription factor, miRNA) (Figure 19A). In the user interface, enriched nodes are referred to as “targets,” as “enriched” might be an unfamiliar term to many users, and identifying drug targets may be the tool's foremost application. The supplied input data can be filtered by their signals and p-values. Additionally, enriched nodes with different signs of signals in the data compared to their sensitivity can be filtered out. Because the upstream enrichment tool requires fetching data for the analysis, results are stored in memory for the time of plugin use, which the user can turn off. The enriched nodes are then displayed in a scatter plot (Figure 19B) by their sensitivity and specificity score. The colors of the regulators are based on their type of regulation: red for positive enrichment and blue for negative enrichment. Nodes marked as 'external' are not included

on the maps and will link to their entry on public databases. Upstream enrichment can enrich not only a single node but also combinations of up to four different nodes. In these cases, the enrichment is performed similarly but with V_e combined and $w_{t,R}(v,u)$ aggregated for each enriched node v. However, the large number of possible combinations strongly impacts the calculation time, and thus, the number of nodes used to iterate through the combinations must be capped. The user can supply a value representing the number of significant targets with the highest sensitivity value from the single target enrichment parsed to the combinatory analysis.

Figure 19: User interface of the Atlas of Inflammation Resolution (AIR) plugins for upstream enrichment analysis from large-scale data files. (A) Similar to the downstream enrichment, the plugin is designed to be performed with minimal user input, and the analysis can be performed directly after data upload with default settings. Users can filter the data further or perform the analysis for a combination of upstream nodes. (B) The results are presented in a scatter plot showing the specificity or sensitivity of all considered upstream nodes or their combinations.

3.3.3 Ongoing Efforts

The tools described in the previous sections are constantly being updated, issues are being fixed, and the MINERVA plugin is adapted for more and more Disease Map projects, listed in Table 5. However, adapting the plugin to individual projects requires a major effort to prepare and publish the data files on GitHub manually. Additionally, given the increasing complexity of the tools to meet the requirements of a comprehensive, multi-level data analysis, their initial implementation became insufficient. Running the analyses purely in the front end bottlenecked the speed due to the user's hardware and bandwidth to load

all required data at every start of the plugin. Eventually, efforts were initiated to perform a large-scale update of the computational infrastructure, which, at the time of this thesis, is still underway. A dedicated web server is being set up to improve the user experience and, most importantly, allow a dynamic function of the plugin to be run on any Disease Map without being required to upload Disease Map-specific data files to GitHub (Figure 20). A demo version of a server-based plugin was set up using the PythonAnywhere platform (https://www.Pythonanywhere.com/) utilizing the Flask Python package as its web framework [216].

Table 5: List of Disease Map projects to which the MINERVA plugin originally developed for the "Atlas of Inflammation Resolution" (AIR) has been adopted.

Disease Map	Disease / Biological Process	Reference
IBD Map	Inflammatory Bowel Disease MeSH (D015212)	Hornschuh M *et al.*, 2021[217] https://www.sbi.uni-rostock.de/minerva/index.xhtml?id=IBD_V2
Melanoma Map	Melanoma MeSH (D008545)	https://www.sbi.uni-rostock.de/minerva/index.xhtml?id=Melanoma_V4
NOSE-OE	Smell MeSH (D012903)	Genovese F *et al.*, 2022 [218] https://www.sbi.uni-rostock.de/minerva/index.xhtml?id=NOSE-OE_13102023
NaviCenta	Pre-Eclampsia MeSH (D011225)	Scheel *et al.*, 2023 [148] https://www.sbi.uni-rostock.de/minerva/index.xhtml?id=NaviCenta
COVID-19 Disease Map	COVID-19 MeSH (D000073640)	Ostaszewski *et al.*, 2021 [149] https://covid.pages.uni.lu/
LSD Map	Gaucher Disease MeSH (D005776)	Confidential industrial project Accessible within the consortium
NPC Map	Niemann-Pick Disease, Type C MeSH (D052556)	Confidential industrial project Accessible within the consortium

The development of such a server for plugins is a step-by-step process in which the data retrieval is first outsourced, i.e., the processing of the data objects previously stored as JSON files on GitHub can be applied automatically for every Disease Map. When the plugin is started in MINERVA, a POST request is sent to the server containing the URL of the MINERVA instance, the project ID, and the creation date of the project, which together uniquely identify a Disease Map. As the content of the Disease Map can no longer be changed once uploaded to MINEVA, the creation date indicates whether a new version of the project exists. Server-side, this information is converted into a project hash string using the **hashlib** library. **`Model`** objects are stored in a dictionary with the project hash as a key and can thus be easily accessed. If the key does not exist, the server tries to create a new **`Model`** object from the disease map files stored in a folder with the project hash as a name. If the folder does not exist, the project files (submaps and images) are downloaded via the MINERVA API and saved in a new folder. The **`Model`** object is further enriched with regulatory interaction files from a global database folder. The project hash is returned to

the plugin frontend and sent with subsequent requests. Finally, the **`Elements.json`**, **`Interactions.json`**, and **`PhenotypePaths.json`** data can be retrieved from the server for the respective project by individual requests that return the data, as was previously implemented via GitHub. In the future, more and more functions of the plugin tools will be moved from frontend to backend processes.

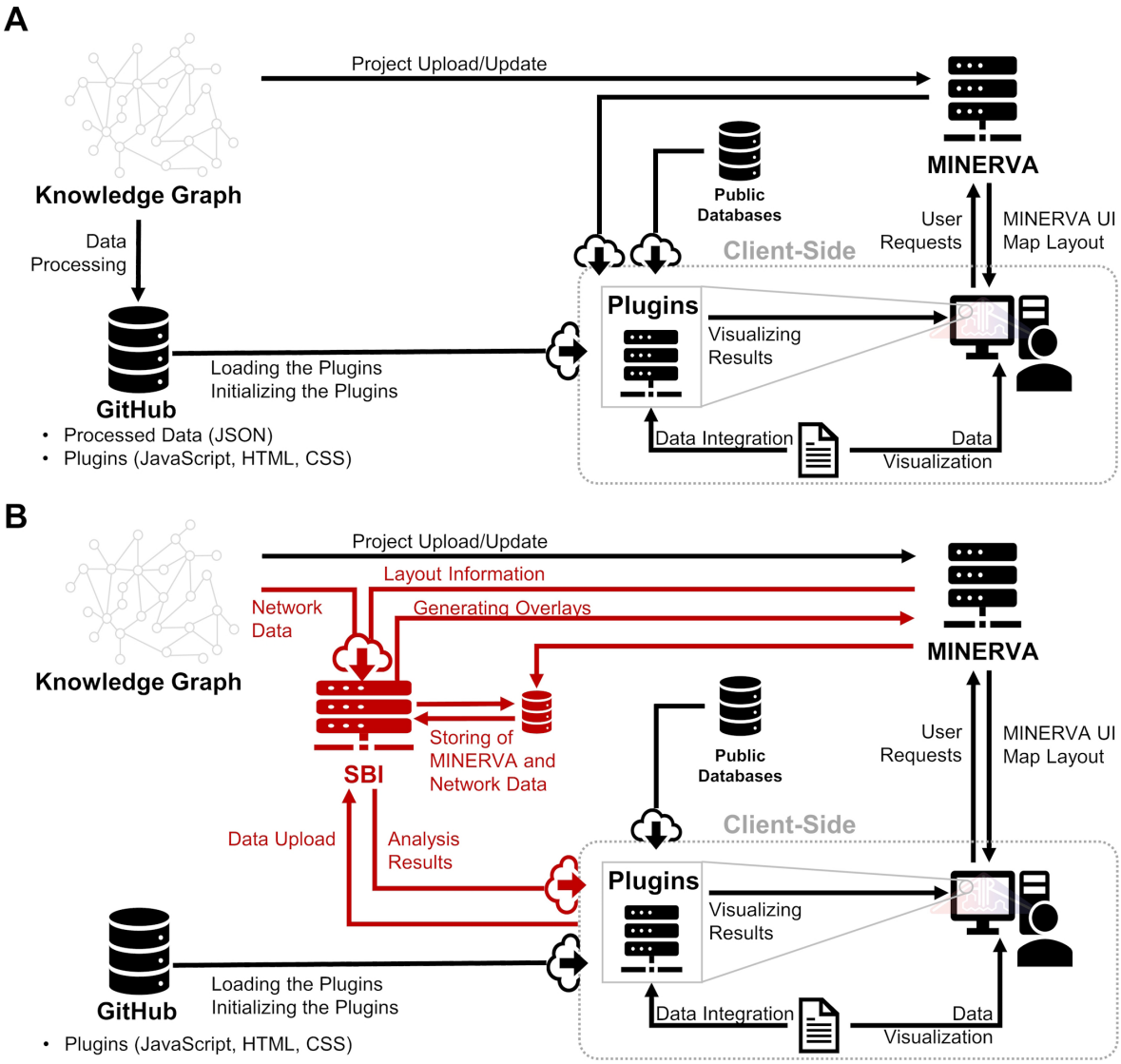

Figure 20: Network diagram showing the infrastructure of the MINERVA plugins developed for the Atlas of Inflammation Resolution (AIR). (A) The initial structure in which the KG was stored as data files on GitHub accessed at each start of the plugins. (B) A dedicated Python server was set up in the updated workflow, which automatically processes the KG and interacts with the MINERVA API.

3.4 Multi-Omics Data Integration

The development of two new tools extends the capabilities of the existing AirXplore and AirOmics plugins by enabling the integration of genetic and mass spectrometry data. These tools convert information from the respective formats into fold-change values. This

conversion enables the effective merging of this data with the omics tool. By combining multiple data types, these tools facilitate the identification of links between different biological levels. This integration is critical to uncovering causal relationships in experimental data and promotes the application of systems biology in personalized medicine. This approach improves the understanding of complex biological systems and supports the development of personalized medical solutions.

3.4.1 Genetic Variant Data

Genomic data is of high interest in clinical research to investigate the links between mutations and diseases, evaluate patient-specific responses, and ultimately enable personalized medicine [219]. A common format to represent genetic variants is the Variant Calling Format (VCF) [220]. VCF files show the frequency of base changes at specific genomic positions compared to a reference genome in a tabular format. Disease Map provides an excellent platform to visualize the role of individual genes in disease processes, and MINERVA already provides an interface to clinicians by highlighting drug targets and disease-related gene variants [205]. However, it misses a direct integration and analysis of large-scale genomics data. To this end, I have developed a plugin integrated into the plugin suite that allows users to upload genomic data files, visualize mutations on the disease map, and assess their biological relevance. Results obtained on gene variant data can then be linked to the Omics plugin to assess their involvement and potential effects on higher-level processes.

In the plugin, the user selects one or multiple local VCF files by clicking on the 'Choose Files' button (Figure 21A). The plugin includes the human genome releases hg19 and hg38, which can be selected in a dropdown list. Furthermore, the user can also select a checkbox to consider negative-strand transcripts in the analysis. After clicking the button 'Read VCF,' all uploaded variants for each sample are mapped to transcripts to genes in the KG. I implemented an interval tree-like indexing algorithm to map the genome positions from the VCF files to the intervals on the chromosomes from the transcripts. When the plugin is loaded, dictionaries are created for each chromosome, with tree-like dictionaries for every order of 10 from the starting positions of the transcripts. The queried genome position is then compared step-by-step for each order of magnitude, and, finally, direct comparisons with transcript intervals only at the lowest order of magnitude, significantly reducing the search space.

Given that no chromosome in the human genome contains more than $3x10^8$ base pairs and only very few transcripts are shorter than 100 base pairs, the tree starts with a multiplier of 10^8 and ends at 10^3. This "range partitioning" principle significantly reduces the number of comparisons required to find the appropriate transcript, reducing the complexity from $O(n * m)$ to approximately $O(logn + m)$. When we switched the plugin from frontend JavaScript to a server-based approach using Python, I employed the **intervaltree** Python package for a simple but even faster solution to map the VCF files. I once evaluated its performance against our previous plugin implementation and a brute-force search (results not shown). The brute force method was up to 1000 times slower. The performance difference between the **intervaltree** and the implementation using range partitioning in the JavaScript plugin was minimal (< 10%), with the former being slightly faster, as expected for an established Python package. Overall, this result shows the efficiency of the previous plugin implementation.

After mapping the VCF data, the transcript and all mapped variants are interactively presented in a table (Figure 21A). Each row in the table contains a checkbox to select the specific gene for further predictions, the gene name, chromosome, number of transcripts (that can be filtered for), and the number of unique variants in each sample. The plugin sends requests to the Ensembl Variant Effect Predictor (VEP) API to analyze the consequences of variants (http://www.ensembl.org/info/docs/tools/vep/index.html). The genes selected in the table are displayed to the user in a text field, which serves as a filter before the VEP, as retrieving the results takes some time (approx. five variants analyzed per second) (Figure 21B). Alternatively, the user can enter the official gene names separated by a comma. For ease of use, the "Select all map elements" buttons can simultaneously add all genes visible in the MINERVA submaps or the "Reset" button to reset all genes already selected. The text field "gnomAD" is an input parameter for the VEP and filters common variants according to their frequency in the populations as provided by the Genome Aggregation Database (https://gnomad.broadinstitute.org/about). Clicking on the 'Predict Variant Consequences' button automatically submits all variants to Ensembl for prediction and displays results as mutation impacts for each sample in the table below. By storing fetched results in memory, variants in multiple samples will only be parsed once, and changing the frequency threshold will not result in a new request. Results can either be downloaded as a .txt or displayed as overlays on the specific genes on the map itself. Mutation impacts are color-coded with high as red, medium as yellow, modifier as grey,

and low as green. Otherwise, the number of transcripts or the number of variants for each gene can be visualized (color-coded as a gradient from white to red).

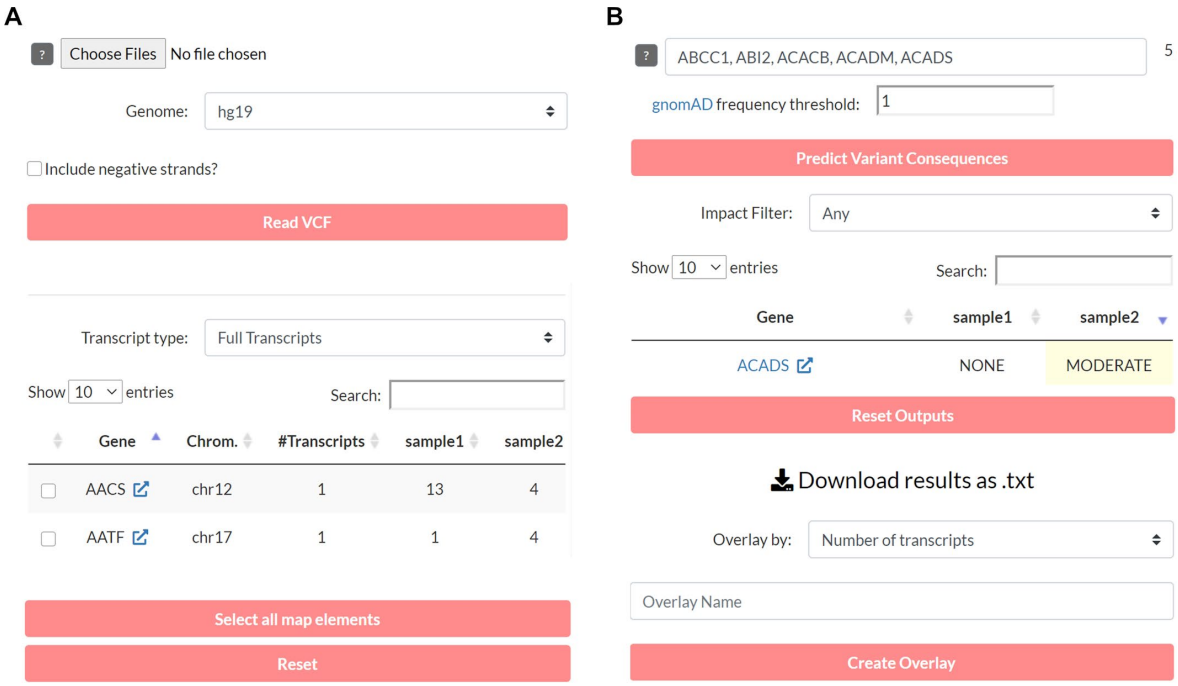

Figure 21: User interface of the Atlas of Inflammation Resolution (AIR) plugins to integrate and analyze genetic variant data. (A) The plugin accepts one or multiple VCF files. The genomic positions in the files are mapped to transcripts of genes included in the AIR, and the information on mapped transcription is shown in a table. (B) Users can select genes for which the effect of variants is then predicted using the ensemble's VEP tool and presented in a table. Genes for which an impact on protein function was predicted can be highlighted on the map or exported to the omics plugin for up- or downstream enrichment analysis.

3.4.2 Mass Spectrometry Data

The last tool added to the plugin enables the integration of targeted metabolomics data. In the first step, the reference peaks are mapped to metabolites or their adducts in the AIR by name or database identifiers, such as ChebiID. The tool enables the upload of two files, one with data of mass spectrometry peaks with their intensity, mass-to-charge ratio (m/z), and other parameters such as retention time (Rt), if combined with chromatography or collision cross section (CCS) in ion mobility spectrometry, and another file describing values for reference peaks of metabolites (Figure 22A). File types and parameter columns are fetched automatically. Additionally, users can select the polarity of the input file. The columns for control and case samples to be compared are selected in two checkbox lists.

The peaks can be filtered by a user-defined delta threshold for all three values, either absolute or relative (Figure 22A bottom). The adducts are mapped to a reference peak by a prioritization based on all three parameters (CCS value, m/z ratio, and retention time). The user can define the type of prioritization method, such as Euclidean distance,

highest/lowest FC, and highest/lowest peak intensity in either sample. The reference and mapped peaks are visualized in a 2d scatter plot with two of the three parameters, showing all other plots fitting the filtering (Figure 22B). For a selected compound, each of its adducts' reference and mapped peaks are visualized in an interactive 2D-scatter plot colored by the FC of the intensity between the case and control samples [red for up-and blue for downregulation]. Users can select which parameter to show on each axis and whether the values in the graph should be filtered by the z-axis, i.e., the third parameter. If Euclidean distance was selected as the mapping method, the distance can be further weighted towards parameters with lower thresholds. The overview tab lists the FC of all mapped peaks and resulting metabolites. The "Adduct" pane shows a table with all adducts and their mapped peaks, including information on the metabolite, the peak's FC, and p-value, calculated through a t-Test, and three parameters with the reference peak's values in brackets (Figure 22C). In the "Metabolite" pane, the adduct's peaks are summarized for each metabolite with an overall FC value associated with the metabolite (Figure 22D). The user can select how the overall FC is inferred when mapping adducts to compounds (selecting FC values of adducts with the highest or lowest intensity value) or when mapping compounds to metabolites (averaging or aggregating FC values). The final result table shows the metabolite (with a link to the map position), FC value, and p-value. These results can be visualized by creating overlays. The user will select a FC value range in which the color will be mapped (blue for negative values, white for zero, and red for positive values). This way, the color range will stay the same when performing a new analysis with a different sample comparison to make them comparable in the KG visualization. As with the Variant plugin, the metabolite results can be imported to the Omics plugin to perform further analysis, such as the upstream enrichment, to assess the influence of the metabolic changes on biological processes.

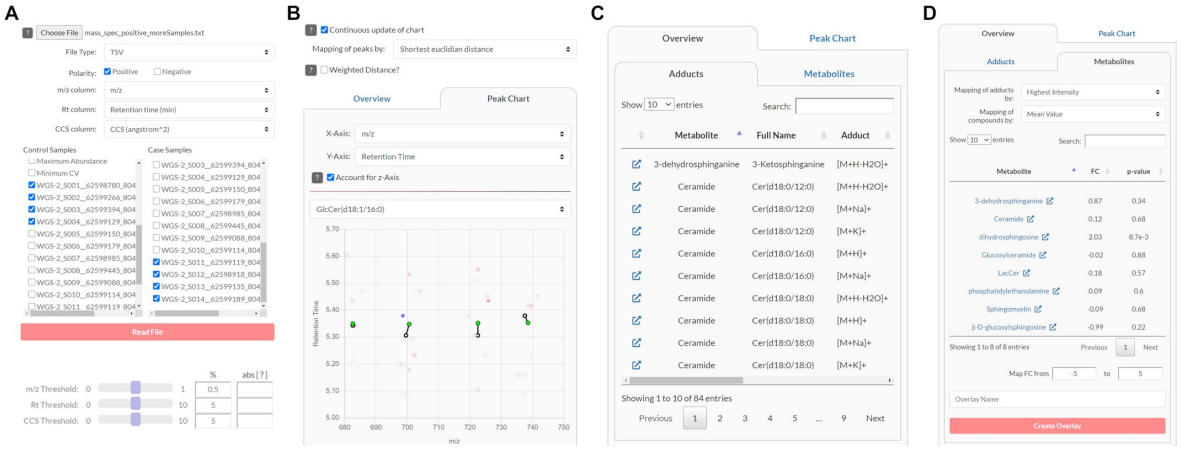

Figure 22: User interface of the Atlas of Inflammation Resolution (AIR) plugins to integrate and analyze metabolomics data. (A) After uploading a tabular data file, users must specify columns containing the rt, CCS, and m/z values and one or multiple columns for a comparative analysis. (B) Peak values from the data file are mapped to reference peaks using user-defined filters. (C) For the mapped peaks of each adduct of a metabolite, log-2 fold change and p-values are calculated between the intensity between the two sample groups. (D) Results for adducts are then mapped to metabolites based on user-defined thresholds.

3.5 Summary

I developed a two-dimensional enrichment analysis (2DEA) that combines relationships from KG topology between the data inputs and the enriched node, with FC values of input data as weighting factors. Including both scores allowed for more detailed evaluations by assessing the direction and strength of the responses. By integrating the topological weightings, I improve the informational value of data enrichment by giving higher weights to topologically more relevant nodes and considering the direction and sign of these relations. The topological weightings I employed for the enrichment can further be used to generate (gene) node sets for other enrichment approaches by filtering them for a defined threshold. Even on their own, topological weightings are a valuable tool for exploring relationships between nodes across a large-scale KG. Their visualization on Disease Maps, as I implanted as an automated function for the MINERVA plugins, provides an intuitive overview of causalities in molecular processes.

The 2DEA can detect enrichments with a small number of associated inputs if they have a higher biological relevance, possibly allowing more insightful predictions. Additionally, by giving the values from the data and topological weightings a high impact on the results, the enrichment becomes less biased towards the size of the enrichment sets. By converting large-scale KGs of Disease Maps into enrichment sets of context-specific associations between nodes, I developed a KG-size-independent solution for data analysis on Disease Maps. I managed to keep computation times to a minimum so that the analyses

could be performed on the client side of the plugins, avoiding the need to upload or store data and preventing data security issues. The approach is highly customizable in that the calculation of the topological weightings can be adapted for various Disease Map types. This customizability improves enrichment capabilities for different types of data, e.g., catalytic scores for metabolomics data and transcriptional scores for transcriptomics data.

Despite great advances in the amount of information considered in enrichment-based approaches such as I have presented here, they remain limited. They consider the elements on which enrichment is based as independent and ignore their mechanistic interactions. Therefore, they do not consider the actual signal events but try to combine data into representative measurements by making assumptions about their aggregated behavior. In addition, the issue mentioned earlier of generating null distributions from data and topology randomizations may be biased by biological phenomena such as gene co-expression [221]. Nevertheless, I believe they will remain relevant in biological data science because they are extremely efficient in assessing data in functional and spatiotemporal contexts. Especially in environments where large amounts of data are generated and analyzed in an exploratory manner, they can significantly facilitate data processing and interpretation.

With the presented methodology, I provide an intuitive solution that enables web-based perturbation experiments and data analysis directly on Disease Maps. I addressed many challenges in developing Disease Map analytic tools, intending to make the method intuitively usable for any researcher. The weightings can be precalculated and stored on the server, enabling fast analyses with large datasets. Plugins require no data upload and can even be performed offline because they are executed as JavaScript on the web browser, and computation times are minimized. Systems biology approaches should help scientists understand their data and point them to potentially important aspects rather than simply displaying computational results or rankings. The plugins provide as much information as possible by incorporating graphical visualization of nodes and their weightings in the enrichment sets, helping users interpret the results and making computations transparent.

Chapter 4

Investigating Cell-Specific Gene Regulation in the Lipid Mediator Switch

4.1 Cell Type Specificity in the Innate Immune Response

Although acute inflammation involves a large number of cells and molecules, many pathways, especially during the initiation of inflammation, trigger relatively straightforward and ubiquitous cascades depending on the type, amount, and timing of stimulus (e.g., production of PIMs, vasodilation, chemotaxis of various immune cells) that ensure a rapid response in any tissue [222], [223]. In contrast, the resolution of inflammation mechanisms (e.g., type and levels of SPM production and their downstream signaling cascades) strongly depends on the tissue microenvironment [224], [225]. The SPMs biosynthetic pathways, including regulatory enzymes, are now largely identified and are the very same involved in PIM production, but the actual regulatory processes underlying cell type-specific mediator profiles remain elusive. In 2018, Norris and Serhan performed a metabo-lipidomics analysis of human whole blood and identified functional and cell type-specific LM profiles [226]. Their results showed that haematopoietically and functionally distant cell types have similar LM profiles and, vice versa, closely related cells can synthesize substantially different LMs, indicating individual cell type-specific regulations. LMs are secreted to neighboring cells in an auto- and paracrine fashion [227], [228]. Such a highly localized response would require cell type-specific transcriptional programs and, thus, a cell type-specific expression of GRNs.

Usually, cell types are defined by cell type-specific markers, morphological features, and functional properties or by their distinct (multi-)omics profiles [229]. With the advancement of single-cell RNA sequencing (scRNA-Seq), new subsets of existing cell types are constantly being defined, and the established boundaries between cell types seem to disappear [230]. Thus, modern experiments focus on single-cell data rather than bulk samples of apparently related cells. However, the idea of subsets of a defined cell type also

adds new complexity to understanding cell type-specific signal transduction that distinguishes them from others. Unsupervised machine-learning approaches proved extremely useful for identifying patterns in single-cell expression profiles to address the challenge of analyzing single-cell physiological or functional relationships [231], [232]. In addition to clustering cells based on their omics profiles, generating topological features from cell type-specific molecular interaction KGs enables the study of functional relationships between molecules and genes [233].

While scRNA-Seq data stores lots of information, it is challenging to discern the impact on cell-specific processes, especially in the context of EDA. With the causal reasoning of KG approaches, differences in expression levels can be targeted to specific outcomes, such as the synthesis of LMs. Second, insights from KG can filter the large amounts of data and reduce it to important information targeted to the processes of interest. The AIR provides a detailed description of the biosynthetic pathways of PIMs and SPMs from their precursors AA, DHA, and EPA, together with large-scale GRNs. The availability of such knowledge in standardized KG diagrams motivates the KG-based analysis of their specific functions in the immune response. In this project, I investigated LM synthesis at the transcriptional level by in silico analyses of cell type-specific GRNs from scRNA-Seq data. In the first part (Section 4.2), I integrated scRNA-Seq data into the GRNs extracted from the AIR, assessed the expression of LM enzymes, and identified topological differences in LM synthesis pathways. In the second part (Section 4.3), I applied a signaling simulation approach to quantify the importance of TFs in each cell type. In both parts, clusters of related cell types are subsequently identified by unsupervised machine-learning methods. The contents of this chapter are published in Hoch *et al*., 2023 [4] and are revisited in the following sections.

4.2 Analyzing Cell Type Specific Knowledge Graphs

4.2.1 Integrating Single-Cell Data

The complete KG of the AIR was filtered by all extracting edges from the “lipid mediator biosynthesis from arachidonic acid,” “lipid mediator biosynthesis from DHA,” and “lipid mediator biosynthesis from EPA” submaps of the AIR. The edges from the maps were converted into AF format by integrating enzymes as intermediate nodes between substrates and products, as described in Equation (1.16). The resulting KG was then extended with all

TF and gene target interactions from the AIR to create a new KG from the original AIR KG G, which I refer to in the following as $G' = (V', E')$ with $V'(G') \subseteq V(G)$ and $E'(G') \subseteq E(G)$. Two murine single-cell RNA-seq profiles (GSE122108 and GSE109125) with preprocessed and library-size normalized read counts (q) by the Immunological Genome (ImmGen) Project were downloaded from their website (http://rstats.immgen.org/DataPage/). These datasets include various immune cell types from different tissues with extensive descriptions of the origins of the samples and cells and their sorting markers. Both datasets have been described in detail in their respective studies [234], [235]. I mapped the murine genes from the data with genes in the AIR using human-mouse gene identifier associations from the Ensemble database (https://www.ensembl.org/), associating the read count $q_i(u)$ for each $u \in V_d$ in every cell type i for the set of nodes V_d that were mapped to the data (see Section 1.3). I defined a read count of 10 as a threshold to mark a gene as expressed or unexpressed, slightly higher than the threshold of 5 used by the ImmGen project to exclude more genes with non-functional expression levels [234], [235]. Genes with read count values below the threshold in a cell type were removed, resulting in cell-specific subgraphs $G'_i = (V'_i, E'_i)$ with:

$$V'_i = \{v \in V' | v \notin V_d \; or \; q_i(v) > 10\} \tag{4.1}$$

$$E'_i = \{(u, v) \in E' | u, v \in V'_i\} \tag{4.2}$$

For each cell type i, the read counts for every node v were normalized to:

$$\hat{q}_i(v) = \frac{q_i(v)}{max(\{q_i(u) | u \in V_d \cap V'_i\})} \tag{4.3}$$

In each subgraph G_c, the shortest paths between precursors (AA, EPA, DHA) and the final products (LM phenotypes) in the LM biosynthesis were identified using the Breadth-First-Search.

4.2.2 Immune Cell Type Clustering

We clustered the cells based on the expression profile of genes included in the G', i.e., being directly related to immunological processes (Figure 23A). The dimensionality reduction largely restored the cell type clusters as they are defined in the metadata of both datasets. I investigated the expression of LM enzymes in the cells and whether clusters of enzyme expression correspond to UMAP clustering (Figure 23C). UMAP reduces the high

dimensionality of the input data to a two-dimensional graphical representation where each point corresponds to a cell in the data. In this way, cells with similar values are positioned close to each other, while separated cells indicate greater differences. Cell clusters were identified using manually adjusted k-means clustering on the generated embeddings. To visualize distributions across all LM classes, the embeddings of each class were combined into a single dataset, on which a new UMAP was applied. Additionally, I analyzed whether an SP exists for each cell from the substrates of LM biosynthesis, AA, DHA, or EPA, to the LM class phenotype nodes (Figure 23B).

The GSE122108 dataset consists of mononuclear phagocytes, mainly macrophages, of different tissues, with various pro- and anti-inflammatory stimuli. The smaller groups of cell types, such as monocytes, dendritic cells, and microglia cells, were mostly restored (Figure 23A). In contrast, macrophage cells are widely scattered and partially mixed with the clusters of the other cell types, most likely because they originate from various tissues. One macrophage cluster separates from all other cells and consists mainly of peritoneal cells. These peritoneal macrophages also show a distinct LM enzyme profile, with an expression of many genes and the only cells with consistently high expression of ALOX15 and PTGIS and, thus, are the only cell types expressing the required enzymes for all LM classes (Figure 23C). The analysis showed that almost all cell types can synthesize prostaglandins, leukotrienes, and thromboxanes, while only very few cell types can synthesize SPMs, except E-resolvins, which show the same pattern as Leukotrienes (Figure 23B). In contrast, lipoxins, protectins, and D-resolvins are only produced by a subgroup of macrophages. The cell types in which three classes occur are incapable of synthesizing maresins. A subgroup of dendritic cells expresses only the enzymes required to synthesize E-resolvins and leukotrienes.

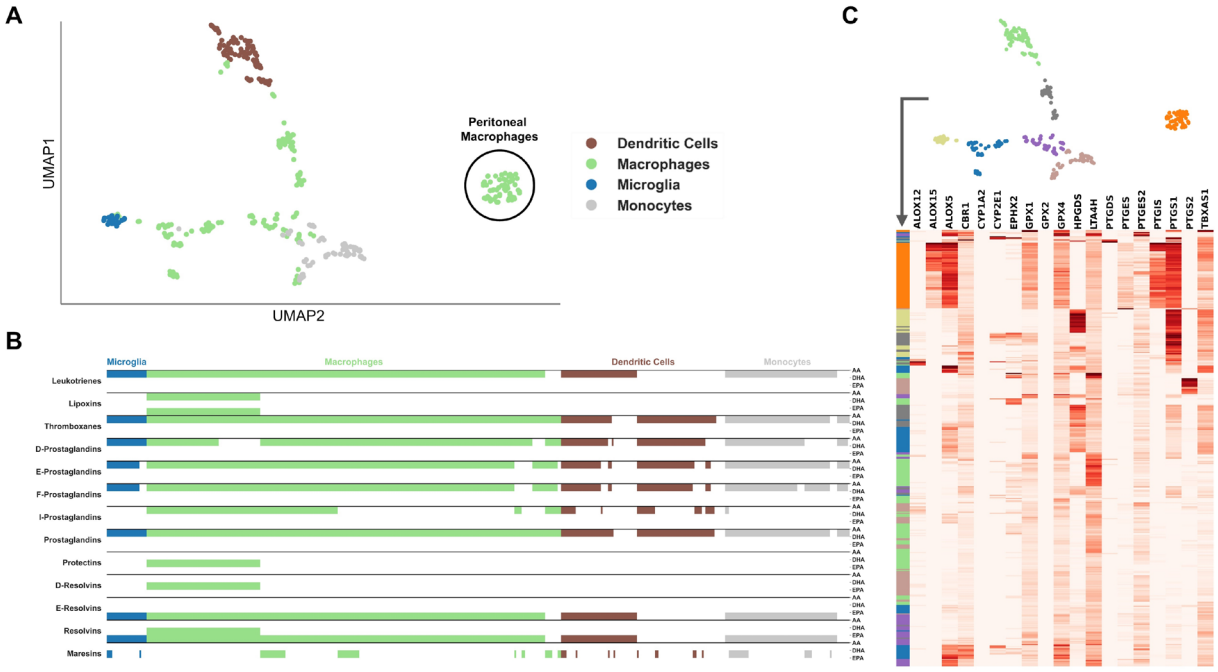

Figure 23: Clustering of immune cell types in the GSE122108 dataset. (A) UMAP plot of immune cell scRNA-Seq data with highlighted clusters based on scRNA-Seq cell sorting. Genes in the dataset were filtered for those included in the "Atlas of Inflammation Resolution." (B) Cell type-specific *de novo* biosynthetic pathways of each lipid mediator class from the precursor molecules arachidonic acid (AA), docosahexaenoic acid (DHA), or eicosapentaenoic acid (EPA), based on the expression of catalyzing enzymes. (C) Clustered heatmap of lipid mediator enzyme expression color-coded by clusters defined from the UMAP in (A). From Hoch *et al.*, 2023.

The GSE109125 dataset consists of many different cell types spanning the hematopoietic lineage. It includes stem cells, epithelial cells, and both compartments of innate and adaptive immune cell populations, with monocytes being the only missing cell subsets. A UMAP analysis on the gene expression $q_i(u)$ for all nodes in $u \in V'$ restored the cell type groups to a high degree (Figure 24A). The two-dimensional projection of the UMAP graph shows the cell branching in two directions, starting from the hematopoietic cell group. Except for B cells, which are placed closer to the myeloid cells, these two groups coincide with the lymphoid and myeloid lineages. The topological analysis revealed that the ability to synthesize LMs, based on the expression of required enzymes, is much lower in lymphoid than in myeloid cells (Figure 24B). In particular, cells belonging to the myeloid lineage and hematopoietic stem cells are the ones most capable of biosynthesizing both PIMs and SPMs, with macrophages, mast cells, and granulocytes (neutrophils, basophils, and eosinophils) being the most efficient due to the high expression of LM enzymes (Figure 24C). NK cells and NKT cells share a similar ability to synthesize the same class of LMs, limited only to prostaglandins (except for I-prostaglandins) and thromboxanes; however, only NKT cells can produce maresins. Interestingly, epithelial cells display a biosynthetic pathway identical to NKT cells.

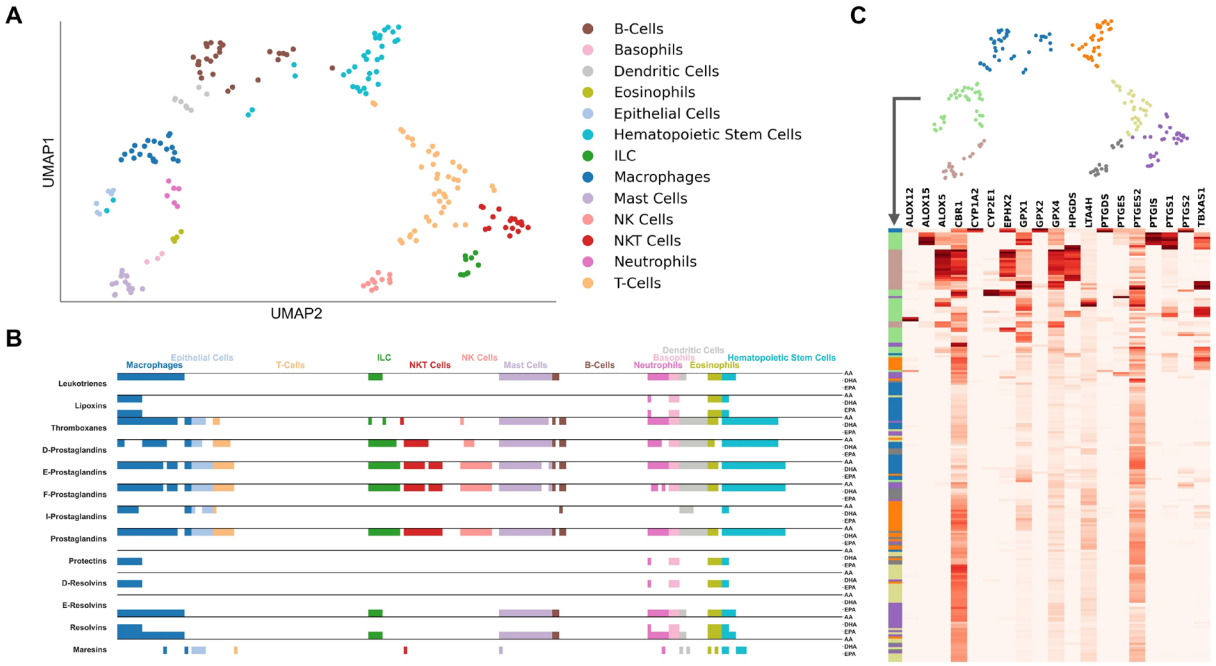

Figure 24: Clustering of immune cell types in the GSE109125 dataset. (A) UMAP plot of immune cell scRNA-Seq data with highlighted clusters based on scRNA-Seq cell sorting. Genes in the dataset were filtered for those included in the "Atlas of Inflammation Resolution". (B) Cell type-specific *de novo* biosynthetic pathways of each lipid mediator class from the precursor molecules arachidonic acid (AA), docosahexaenoic acid (DHA), or eicosapentaenoic acid (EPA), based on the expression of catalyzing enzymes. (C) Clustered heatmap of lipid mediator enzyme expression color-coded by clusters defined from the UMAP in (A). From Hoch *et al.*, 2023.

4.3 Simulating Cell Type Specific Gene Regulation

4.3.1 Signal Transduction in Single-Cell GRNs

We performed a signal simulation based on the approach presented by Lee and Cho [236], starting from the enzymes in each LM synthesis and continuing in the reverse direction through each G'_i (Figure 25A). The algorithm is based on a distinct propagation of continuous signal values throughout the KG. At each step, the signal for each node in the KG is updated based on weighted signals of nodes from incoming edges. The weightings are based on the target node's in-degree and the source node's out-degree as normalization factors to avoid overestimating highly connected nodes. The simulation was initiated separately for each LM class by setting the starting scores to $s_i(v,0) = w_{t,P}(v,p)$ for each enzyme v in the LM class p. Node signals were updated at each step based on degree centralities (= number of interactions) in the G', their targets' scores in the previous step, and their normalized read count $\hat{q}_i(v)$ (Equation (4.4) and Figure 25B).

$$s_i(v,t) = \sum_{u \in N_{out}(v)} \frac{\hat{q}_i(v) \cdot s_i(u,t-1)}{\sqrt{c_{d_{in}}(u)} \cdot \sqrt{c_{d_{out}}(v)}} \tag{4.4}$$

In contrast to Lee and Cho, I excluded the parameter α, which defines the proportions in which the incoming signal and the initial signal of the node are aggregated. The subgraphs for each LM synthesis extend hierarchically from the enzymes, which are the only nodes with an initial signal and are not passed again. Normalizing signals at each step to the starting signal would thus not affect the results. Also, instead of identifying the signals in a steady state, the simulation should estimate how the gene regulation is distributed across the GRN. The initial signal for the LM enzymes is thus kept for a defined number of steps (20) and set to 0 for the remaining steps. The simulation can, therefore, be interpreted as if a batch of weighted signals were sent from the enzymes backward through the graph and measured at TFs. Finally, the regulatory score $\bar{s}_i$ for each node v is then defined as the aggregated score over 100 signaling steps (Figure 25C).

$$\bar{s}_i(v) = \sum_{t=0}^{100} s_i(v,t) \tag{4.5}$$

For each LM class, a UMAP analysis is then performed on all regulatory scores $\bar{s}_i$ from all graphs G'_i. Cell clusters were identified using manually adjusted k-means clustering on the generated embeddings. Then, the embeddings were combined into a single dataset to visualize distributions across all LM classes, and a new UMAP was applied to the combined embeddings. Clustering in the enzyme expression heatmaps was performed using the Euclid-based hierarchical clustering method of the Python package **seaborn** [237]. Furthermore, I extracted core regulatory networks (CRNs) from each G'_i by combining paths from selected TFs to the LM enzymes with maximized $\bar{s}_i$ for all nodes on the paths. This maximization is a widest path problem and was solved using an adaptation of Dijkstra's algorithm described in Algorithm 2. The edge weights are based on the edge's target node u and were set to either $\bar{s}_u$ for CRNs of a single cell or $|\Delta\bar{s}_u|$ when comparing two sets of cells.

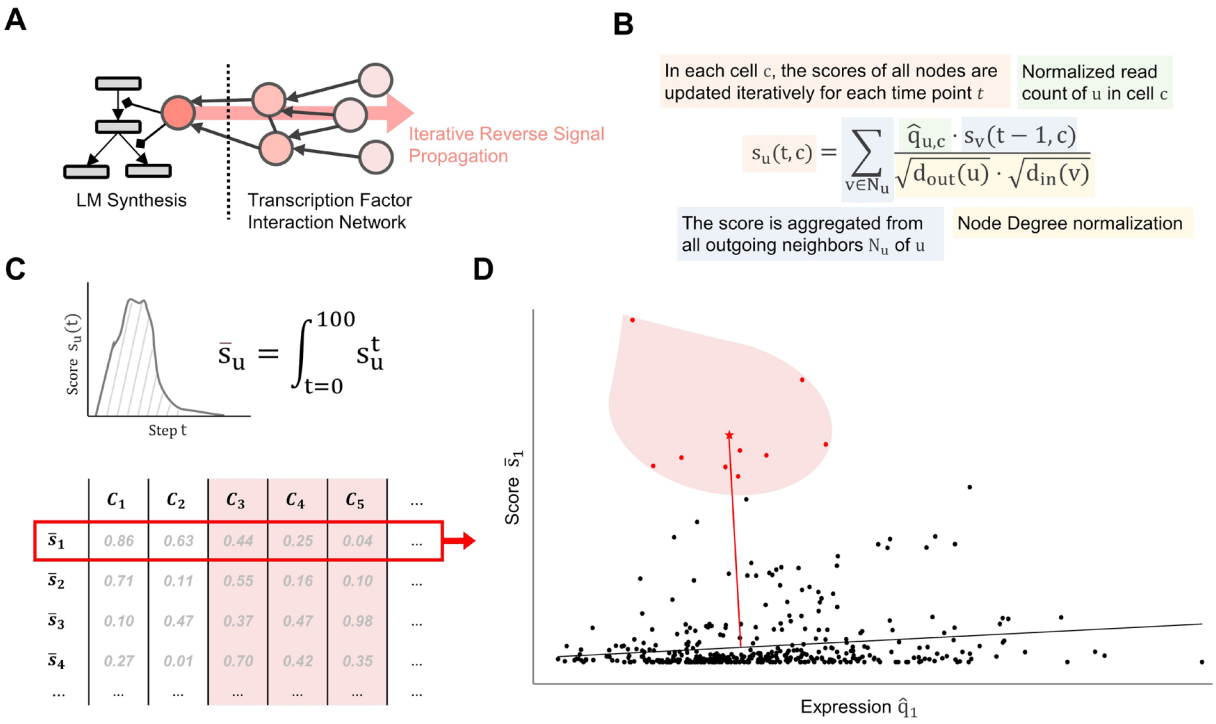

Figure 25: Feature extraction from the cell type-specific gene regulatory networks (GRNs). (A) The starting signal is traversed in reverse throughout the GRN, starting from lipid mediator enzymes. (B) For a distinct number of steps, a score is updated for each transcription factor (TF) based on its gene expression, target score in the previous step, and the node degrees of both the TF and its target. (C) The final score is defined as the AUC of the scores over 100 steps. (D) The statistical significance of a score is calculated based on its distance to a regression line representing the correlation between the score and the expression of the TF across cells.

Because the calculation of regulatory scores is based on the expression of the feature in the cell, the final scores are biased toward $\hat{q}$. Therefore, instead of calculating the highest scores, the features were statistically analyzed in relation to $\hat{q}$. In an LM class, the $\hat{q}$ and $\bar{s}$ values of all cells not in the cluster were fitted to linear regression, and a half-normal distribution was created from the absolute distances of each cell from the line (Figure 25D). The p-value of the feature in the cluster is then calculated from the z-score of the average distance of the cluster's cells in the distribution. Finally, adjustment for multiple testing among all enriched nodes in each cell was performed through false discovery rate (FDR)-correction by *Benjamini–Hochberg* to generate adj. *p*-values [211]. For each cluster, I identified the genes with the most significant differences compared with all other cells (adj. *p*-value < 0.05).

4.3.2 Gene Regulation of Lipid Mediator Synthesis in Immune Cells

In the GSE122108 dataset, I observed many separate clusters and good restoration of the main cell types, i.e., dendritic cells, macrophages, microglia, and monocytes (Figure 26A, Supplementary File 1). Of note, macrophages appeared as smaller clusters that are partially composed of tissue-specific cells, e.g., from the aorta, heart, or liver. I identified the

significant (adj. *p*-value < 0.05 for any LM class) genes of the microglia cells, which build the most defined cluster in the UMAP plot (Figure 26C). For the two highest-ranked genes, MEF2A and XRCC5, I additionally showed their regulatory score in relation to their expression in all cells. The plots show how the score is significantly increased in the microglia cells and, especially for MEF2A, is independent of its expression. Information on tissue-specific transcriptional regulation of LM biosynthesis is very sparse in the literature. Hence, to compare the results with experimental data, I searched the literature for evidence of the immune modulatory function of the genes related to microglia. Of the 13 genes, I found clear evidence in the literature for eight genes on their relevance in microglial function and neuronal inflammation (MEF2A [238], HDAC11 [239], [240], SMAD3 [241], MEF2C [242], ARID1A [243], [244], ZFHX3 [245], [246], ETS1 [247], and JUN [248]). Four genes were mentioned in experiments on microglial inflammation (XRCC5 [249], [250], ZFP191 [251], PRDM1 [252], and USF2 [253]), whereas no information was found in the literature for only two genes (Znf383, and Nfrkb). The mode of action of the predicted genes in modulating microglia function has been attributed to their impact on cytokine expression. These results suggest that they modulate the immune response also by regulating the expression of enzymes involved in the biosynthesis of LMs. SMAD3, JUN, USF2, and XRCC5 have already been described in their regulation of prostaglandins, while little to no research is available on other LM classes [254], [255], [256], [257]. MEF2A and MEF2C have been identified as downstream effectors of PGE2, which could indicate a feedback loop on the prostaglandin e synthesis [258], [259].

In contrast, in the GSE109125 dataset, the original cell types are more heterogeneously distributed between clusters (Figure 26B, Supplementary File 2). The differences in the expression of immune-related genes between the major immune cell types are not reflected in the TFs associated with the LMs. However, two clusters consisting of hematopoietic stem cells and mast cells, respectively, are strongly separated. While no significant TFs were identified for the latter, the former shows a division into three subclusters, from each of which several significant TFs were identified. Interestingly, based on cell metadata, the three subclusters appear to represent stages of lymphoid hematopoiesis, namely (i) bone marrow-derived stem cells (BMSCs), followed by (ii) early (DN1, DN2a lymphocytes) and (iii) late lymphoid progenitor cells. While BMSCs express many LM enzymes, they are downregulated in lymphoid progenitors. When comparing the regulatory scores of stem cells and early lymphoid progenitor cells, HLF had the greatest difference in its score for all LM classes (not shown). HLF is an important regulator

of lymphoid development in the hematopoietic Lineage [260]. The results suggest that modulation of LM synthesis by gene regulation of LM enzymes may play a role in shaping the fate of lymphoid cells by HLF.

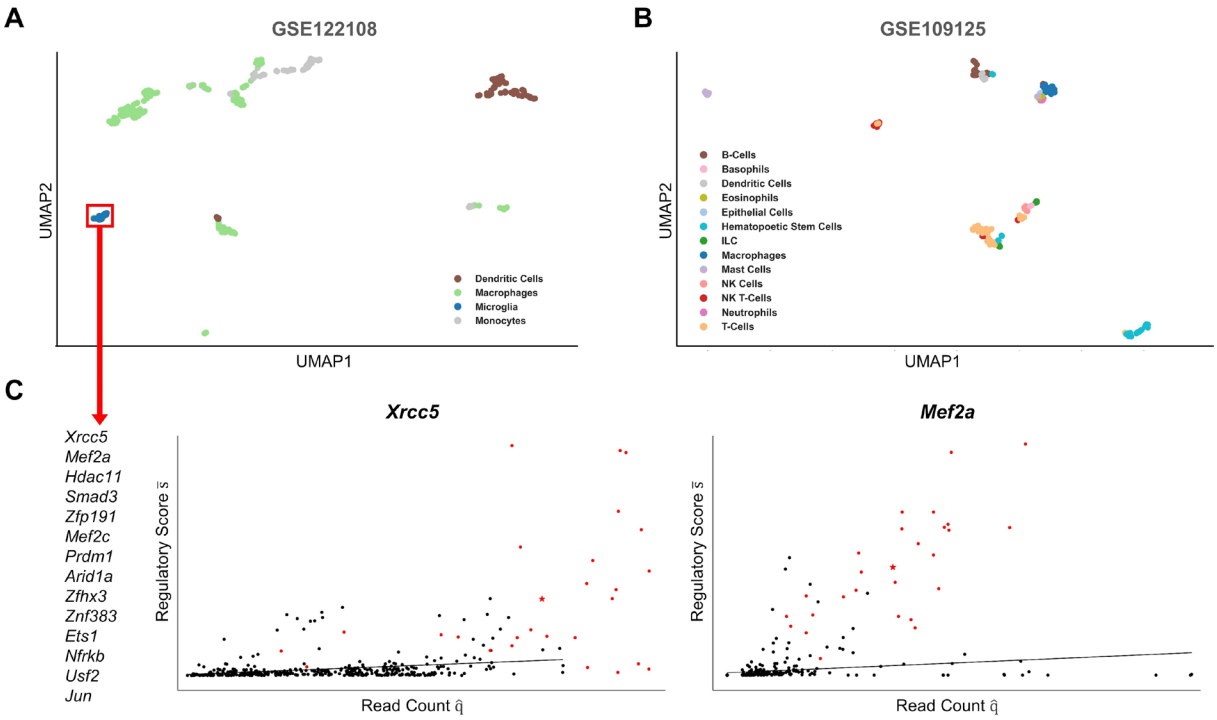

Figure 26: UMAP clustering of individual cells based on their topological association and expression of transcription factors (TFs) related to lipid mediator biosynthesis. scRNA-Seq profiles of two data sets, GSE122108 (A) and GSE109125 (B) were mapped to a gene regulatory network (GRN), and topological features were extracted for the UMAP. (C) For the microglial cell cluster, TFs with significantly higher scores than other clusters are shown. For the two highest-scoring TFs, XRCC5 and MEF2A, their score and their expression in the cluster (red) compared with all other cells (black) are shown in a scatter plot.

Several samples in the GSE122108 data were treated with pro- or anti-inflammatory stimuli at several time points, including lipopolysaccharide stimulation (LPS), *C. albicans* infection, induction of injury, paracetamol, and thioglycolate. I compared the cells in successive time points for each stimulus and identified the TFs with the strongest changes in their gene regulatory activity for each LM class (Figure 27A). For selected genes, I additionally show violin plots comparing their expression values (read counts) and topology scores, showing that the estimated change in connectivity is independent of their expression (Figure 27B). In general, the predicted TFs show a strong variability between cells and the different stimuli, suggesting that gene regulation of LMs in the immune response is highly cell type and environment-specific. Additionally, especially at early time points, the identified TFs differ substantially between PIMs (e.g., the prostaglandin classes) and SPMs (e.g., the resolvin classes) due to the distinct enzyme profile, arguing for fine-

tuned gene regulation. At later time points, the difference between PIM and SPM classes becomes smaller, and the number of overlapping TFs increases.

Many predicted genes are well-known regulators in the immune response to the respective stimuli. For example, in liver macrophages stimulated with APAP, HES1 appears to be a key regulatory TF of most SPM classes. *In vivo* experiments showed that blocking the Notch signaling pathway in mice reduced HES1 levels and increased susceptibility to APAP-induced liver injury [261]. In thioglycolate-stimulated monocytes/ macrophages, our model predicted several genes related to both PIMs and SPMs synthesis, which have also been described in the literature, such as EPAS1 (prostaglandins), EGR2 (prostaglandins), CEBPB (all LM classes), and SREBP1 (SPMs). EPAS1, coding for HIF-2α, is an important mediator of cellular processes and macrophage recruitment in response to hypoxia [262]. In an experimental thioglycolate periodontitis model, EGR2 and CEBPB were required for macrophage activation [263]. In SREBP1 knockdown mice, thioglycolate-elicited macrophages showed increased levels of pro-inflammatory cytokines and reduced levels of DHA and EPA during the resolution phase after TLR4 activation [264]. Although they are related cell types, the five subtypes of LPS-stimulated lung macrophages also differ in the predicted TFs. Two subtypes of lung macrophages originate from broncho-alveolar lavage (BAL) and show a similar gene regulation of prostaglandins through KLF10 and VHL. Both genes have already been associated with inflammatory responses in BAL macrophages [265], [266]. For other LMs, both BAL subtypes do not overlap in the predicted TFs. The remaining lung macrophage subtypes are defined by cell sorting markers. Their samples for which data are available at days zero and three after LPS stimulation overlap at STAT1, STAT2, and PIAS1. The results become more diverse at later time points (day six vs. day three). The three MHC-II$^-$ macrophage and monocyte subtypes partially overlap in FOXK2, RORA, and ING4. These genes are associated with cytokine production in response to LPS [267], [268], while for the MHC-II$^+$ subtype, autophagy-related genes RB1CC1, RB1, and HDAC2 were predicted [269], [270], [271]. Whether or not this difference is caused by MHC-II is yet to be determined, as only a few pieces of evidence connect MHC-II with the predicted genes.

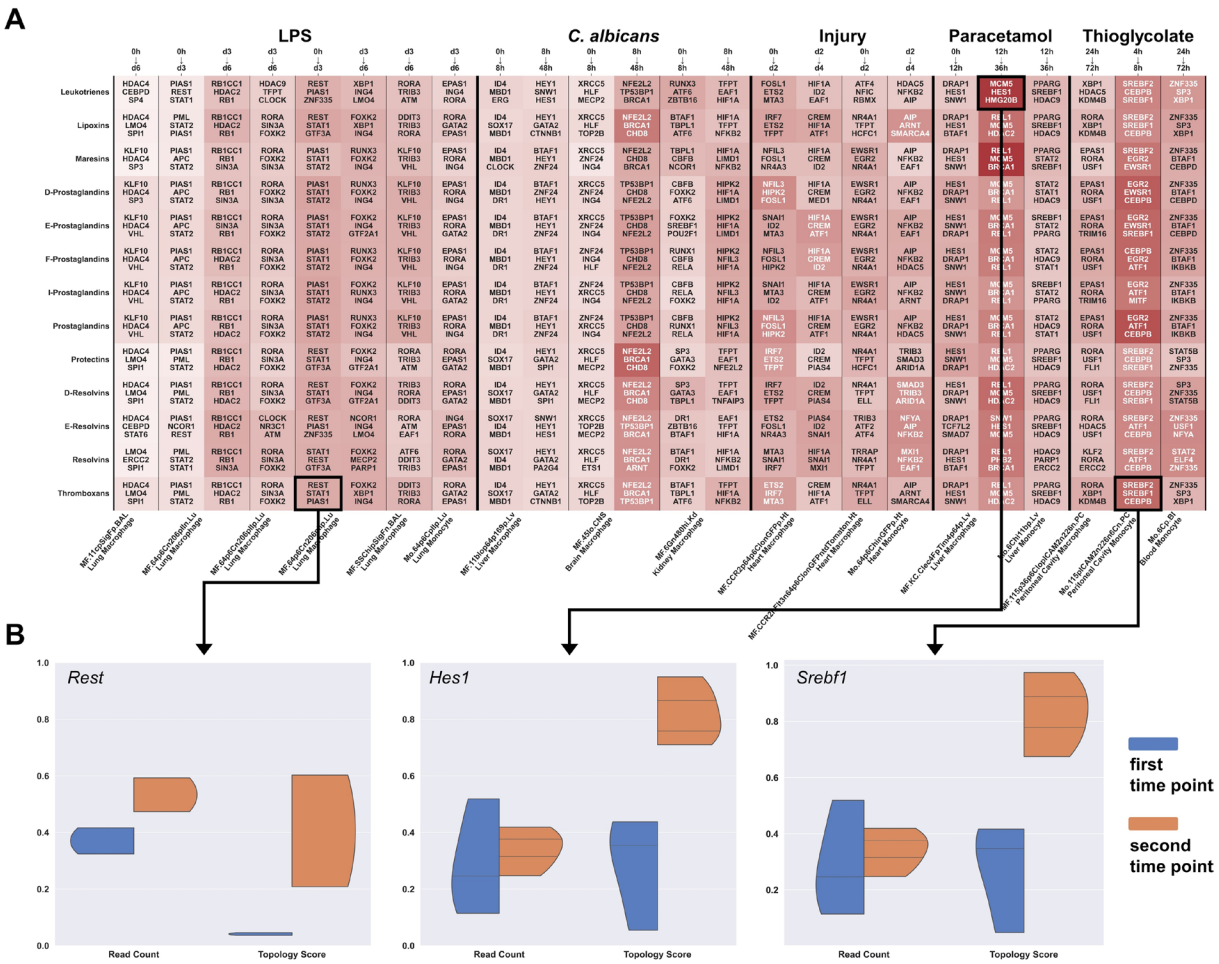

Figure 27: Transcription factors (TFs) associated with stimulation of immune cell types. (A) The GSE122108 dataset includes gene expression data of immune cell types stimulated with inflammatory agents for different time points. I identified the three major TFs with increasing topological association to each lipid mediator class between time points of each cell type. For three selected genes, REST, HES1, and SREBF1, the normalized expression levels and topology scores for all samples are shown in a violin plot.

Since the transcriptional regulation of LMs appears to be tightly regulated and cell type-specific, I then investigated the extent to which closely related cell types may differ in the GRNs of PIM and SPM synthesis. I identified the cell pairs with the smallest distance in expression-based UMAP but the largest distance in transcriptional topology-based UMAP. The top-ranked sample pair consists of aorta macrophages and lung macrophages stimulated with LPS (Figure 28A). Both tissue-specific subtypes of macrophages appear to have a nearly identical transcriptomic profile but substantially differ in the LM gene regulation. Thus, I extracted the CRNs to gain further insight into the genes contributing to the observed differences (Figure 28B). I additionally generated a CRN of an unstimulated sample of the same lung macrophage subtype but without LPS stimulation to ensure that the difference is not caused by the response to LPS. Interestingly, the CRN shows that the expression of most LM enzymes is similar except for PTGS2, which is not expressed in aorta macrophages. In contrast, PTGS2 is highly expressed in aorta macrophages with high

expression levels of the TFs JUN, EGR1, and FOS. All these three genes are highly associated with atherosclerotic inflammation [272], [273], [274]. Egr1 is involved in the response to mechanical or oxidative stress and, thus, the development of atherosclerosis from plaques and hypertonia [272], [275], [276].

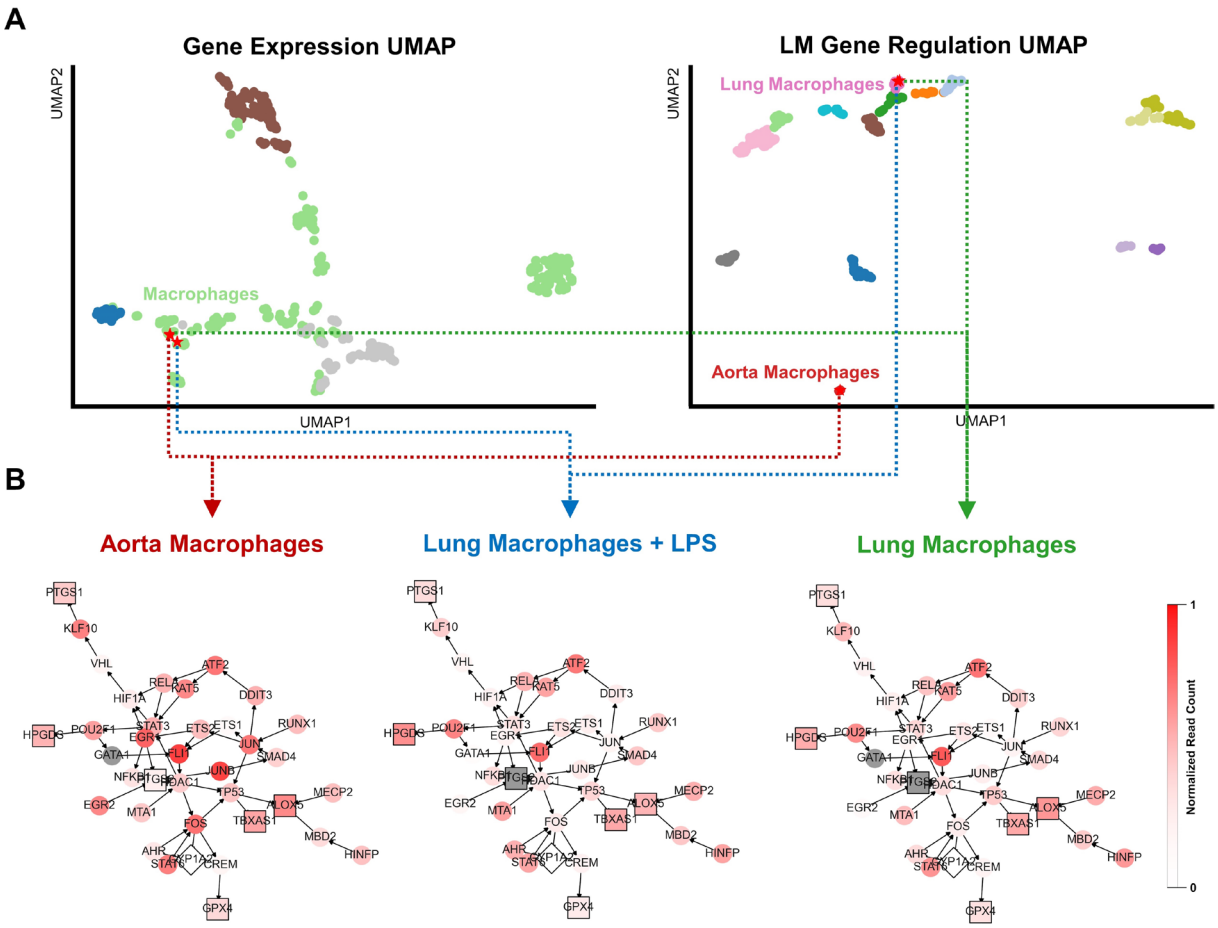

Figure 28: Identification of cells with similar expression profiles but different transcriptional regulation of lipid mediators (LMs). (A) Cell samples with minimized distance within the expression-based and large distance in the LM-regulation-based UMAP. The highest-ranked cell pair consists of a sample from aortic macrophages and one from lung macrophages stimulated with LPS. I included unstimulated lung macrophage to show that the difference is not caused by the reaction to LPS. (B) Gene regulatory networks connected to LM enzymes for all cells colored by their normalized read count values. The shape of the nodes distinguishes between TFs (round), LM enzymes (square), and nodes not included in the transcriptomics data (diamond). Nodes with read counts below the absolute threshold of 10 are highlighted in gray to distinguish them from lowly expressed ones.

4.4 Summary

The project provided further insights into the gene regulation of lipid mediators in the immune response and the contribution of selected cell types to their biosynthesis. In the study, I employed KG-based approaches to analyze functional molecular pathways and their gene regulation using scRNA-Seq data. I investigated the expression of LMs that

synthesize enzymes in different immune cell types and identified key TFs that control the expression of their genes.

The signaling-based approach I used overcomes some of the limitations of the enrichment-based methods mentioned in Section 3.5. The approach considers the interdependencies of TFs and thus identifies cell-specific differences hidden in the cell-specific KG topology. In this way, I was able to identify relevant TFs that do not show a change in their gene expression but disturbed signaling due to the expression of connected TFs. However, the KG was constructed exclusively from GRNs, limiting mechanistic information, such as the effects on and from protein signaling and feedback regulation of metabolic processes. Consequently, this method can analyze a snapshot of the cells during the RNA-Seqeuencing but cannot predict the temporal dynamics in the LM synthesis. Although integrating other KG types is theoretically possible, the gene expression data would have to be linked to the protein and metabolic level, and mechanistic interactions would have to be included, requiring sophisticated multi-scale approaches.

With the approach I presented, I could integrate single-cell data into KGs and perform their analysis at the transcriptomic level. I have shown that gene regulation highly depends on cell type and stimuli. Although the cell types have similar expression profiles, they express different regulations of LM synthesis and thus respond with different LM production to experimental conditions. These results suggest a finely tuned transcriptional modulation of immune cell types and emphasize the need for systems biology approaches to understand the underlying mechanisms.

Chapter 5

Multi-Target Drug Mechanisms in Knowledge Graphs

5.1 Multi-Target Approaches in Pharmacology

Traditionally, drug research has primarily focused on a single-molecule, single-target approach (Figure 29). Drug treatment often has an inhibitory effect of alleviating symptoms by downregulating pathological processes. However, the clinical benefit of single-component therapy can be limited [277], [278]. While they may positively affect the desired symptom, they may also be associated with various side effects. One explanation could be that the dysregulated pathways are involved in many more processes than are relevant to the disease pathology. Therefore, a strong inhibition could often go beyond the desired effects, spreading to physiological pathways or affecting pathways essential for regeneration. Similarly, the traditional treatment of inflammation is carried out using the non-steroidal anti-inflammatory drugs (NSAIDs) class, which includes aspirin, diclofenac, and ibuprofen. NSAIDs have a potent anti-inflammatory, pain-reducing, and partially anti-phlogistic effect. However, they also show many side effects, such as stomach ulcers and reduced tissue healing, and are usually not suitable for long-term use or open wounds. Their potent inhibitory effects on a key innate immune response pathway may interfere with pro-resolving and tissue-healing processes.

Conversely, multi-target approaches of multi-component drugs aim to overcome these issues by simultaneously utilizing the interferences in the effects from multiple sites of action. The effects of individual drugs should positively overlap with the targeted disease processes while not overlapping with other undesired processes. Thus, the same, or even greater, effect can be achieved through lower doses, reducing side effects through a drug-sparing effect, and potentially overcoming adaptive resistances [279], [280], [281], [282], [283]. The multi-target is understood as a “fine-tuning” of disease processes. As outlined in Chapter 2, current scientific evidence emphasizes that the resolution of acute inflammation is not a passive event but an active, orchestrated process [284], [285], [286]. Resolution pharmacology may be more suitable for long-term treatment as it

simultaneously regulates inflammation and promotes rather than inhibits natural physiological processes, particularly resolution pathways. Therefore, multi-target approaches aimed at resolution processes could be more effective than the traditional single-target inhibition of inflammatory diseases.

However, without a deeper understanding of the underlying molecular mechanisms, pharmacology is a black box from which the effects of drugs can only be assessed based on higher-level responses. This severely limits the potential for more sophisticated and targeted approaches, especially for drug repurposing. The possibilities of multi-target combinations are endless, making assessing their effects through in silico simulations essential before translating them into *in vitro* and *in vivo* experiments. Aside from these multi-target drug combinations, there are multi-component products, such as plant extracts, whose combined effects might be known. However, there is a lack of understanding of the individual component's mechanisms [287], [288]. Investigating the multi-component actions in detail would allow us to adjust the formulation to improve efficiency or repurpose the drug. KG modeling approaches have become invaluable in these applications, shaping the field of systems pharmacology [289], [290], [291], [292], [293], [294].

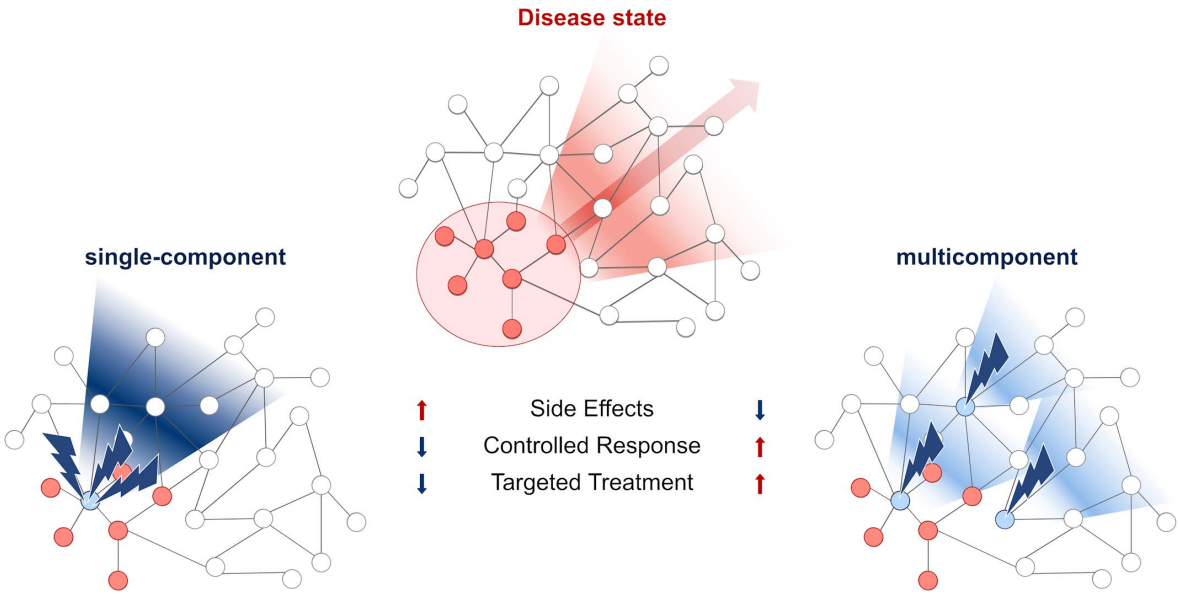

Figure 29: Principle of multi-target drug application vs. the conventional single-target drug approach. Diseases are caused by dysregulations in molecular processes that propagate through time and space. The conventional single-target approach aims to identify and inhibit a key regulator that promotes the disease phenotype and thus prevents or even reverses disease progression. A sufficient response from a single target requires strong perturbations, which can cause uncontrolled side effects in undesired processes. The multi-target approach aims to overcome these limitations by affecting multiple targets with interferences on disease processes, thus achieving more targeted effects with fewer doses.

In collaboration with *Heel GmbH*, we investigated the molecular mechanisms and effects of a multi-component formulation of natural products, referred to as Tr14 in the following. Evaluating the mode of action of a multi-component drug through KGs is a multifaceted process that involves multiple steps of analysis. These could include (i) identifying the direct targets of the individual components, (ii) evaluating the combined drug's effects on higher-level processes detected by experimental or clinical means, and (iii) identifying the mechanisms that link (i) and (ii). Points (i) and (ii) can be assessed with some degree of confidence because of the empirical evidence of the experimental measures and because they can be evaluated from direct molecular interactions or process-specific KGs. However, the underlying link signaling mechanisms present a more complicated picture. They require large-scale KGs to connect the drug targets to the observed effects. For example, if the latter were RNA-Seq data, the KG analysis would require a PPI part to simulate downstream signaling from the drug targets and a GRN part that connects the final signal on TFs to their regulated genes.

The AIR provides such a multi-level KG specifically curated in the context of the acute immune response. For Tr14, we had access to time-series RNA-seq data from a previous study on a murine wound healing model treated with Tr14 or the single-component drug diclofenac [295], [296]. Both datasets consist of 7 samples, at 12h, 24h, 36h, 72h, 96h, 120H, and 192h, compared to the respective control, either drug-free topical ointment or saline injection. Diclofenac is an NSAID known to inhibit the synthesis of prostanoids such as prostaglandin-E2 (PGE2), prostacyclins, and thromboxanes by blocking both cyclooxygenase 1 (COX-1) and cyclooxygenase 2 (COX-2) enzymes [297], [298]. Previous studies have shown that Tr14 regulates several pathways associated with the resolution of acute inflammation, including apoptosis, leukocyte migration, and angiogenesis [295], [296], [299], [300]. Tr14 positively impacts the synthesis of specialized pro-resolving lipid mediators (SPMs) in human monocyte-derived macrophages. In addition, it enhanced efferocytosis and SPM production in a zymosan-induced mouse model [301]. While RNA-Seq shows indirect effects on the transcriptional level at specific points in time, KG-based analyses allow us to use this information to understand higher-level processes and identify connections between samples that might suggest cause and effect. A separate project was also undertaken to create a Tr14 drug interactome using publicly available databases and predictions of compound-target binding by artificial intelligence methods. Mapping signals of the drug targets throughout the KG in interferences, giving insights into the potential effects of Tr14 on immune processes.

Comparing the results from the RNA-seq data and the drug interactome analysis generates more confidence in the overlapping results (Figure 30). This also allows to distinguish between direct and indirect drug effects, providing a deeper understanding of the drug's mode of action.

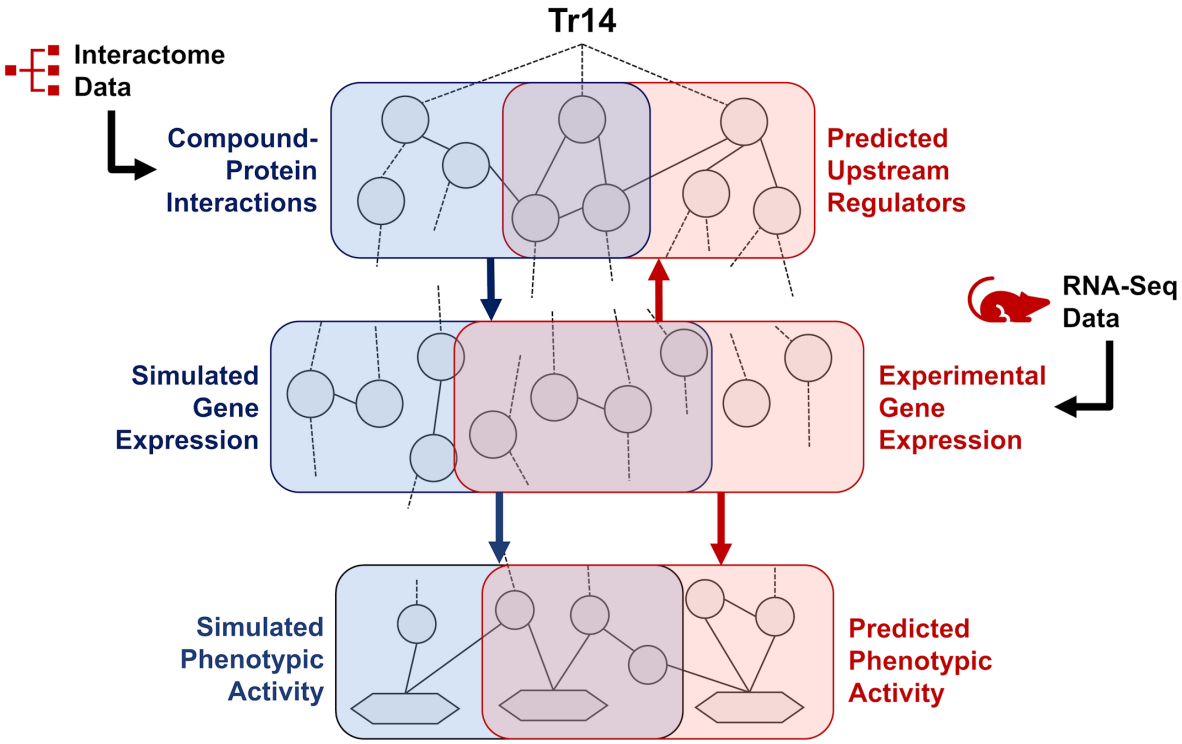

Figure 30: Concept of knowledge graph-based analysis to investigate the molecular mechanisms of a multi-component drug Tr14. The effects on gene expression and higher-level biological processes can be predicted from possible protein targets in drug interactome data by downstream enrichment analyses using knowledge graphs. At the same time, gene expression data from *in vivo* drug response experiments provide insights into higher-level processes through downstream analyses and into potential gene regulators through upstream analyses. The information from both approaches can be combined to increase confidence in the predicted mechanisms.

The project was divided into three core components, explained in more detail in the following sections.

(i) RNA-Seq data analysis utilizing the 2DEA KG-based enrichment approach, as outlined in Chapter 4.

(ii) Creating a Tr14 interactome and performing 2DEA on the compounds' drug targets. This way, one can compare the direct effects with the indirect ones observed in the RNA-seq data.

(iii) Simulating signal transduction from Tr14 targets to gene regulation to generate a simulated gene expression and compare the results to the empirical data.

5.2 Analyzing a Multi-Component Drug Transcriptome

5.2.1 Integration of RNA-Seq Data

While previous studies on the data identified the differential expression of selected genes or performed undirected enrichment analyses, the effects on inflammatory processes and the underlying molecular mechanisms were still unknown. The goal is to infer knowledge of clinically relevant drug effects and mechanisms from local bulk-tissue samples in the mice model. In the first part, I re-analyzed the data using the AIR tools employing the 2DEA. I combined the FC values and adjusted p-values from the diclofenac and Tr14 RNA-Seq data. In the plugins, the node sets $V_{d,T} \subseteq V(G)$ and $V_{d,D} \subseteq V(G)$ represent the set of nodes in G mapped to the probes in the data. The mapping was performed using the official gene names with node names. For each sample i, the FC values of DEGs (adj. p-value < 0.05) were integrated as node signals in the KG and are denoted as $s_{D,i}(u)$ for every $u \in V_{d,D,i}$ or $s_{T,i}(u)$ for every $u \in V_{d,T,i}$. I then performed the 2DEA for every phenotype and every sample i in both the diclofenac and TR14 data, generating a set of phenotype signals $\{s_{p,i}(v) | v \in V_p\}$. The analysis results are summarized in Figure 31A-D for both diclofenac and Tr14, using a pie chart to visualize the predicted levels of inflammatory processes (slices) at each time point (inner circles).

5.2.2 Phenotype-Specific Subgraphs

While the AIR tools statistically evaluate the aggregate effect on an upstream or downstream node, they do not provide insights into the actual regulatory mechanisms of nodes not included in the data. Therefore, I added a feature to the tools to generate subnets for each predicted phenotype in each sample. I adapted an approach from Khan *et al*. in 2017 based on ranking motifs that are gene triplet feedback loops by scoring and weighting normalized topology and expression features [302]. A Pareto set of motifs with the highest feature scores is generated by iteratively changing the weights and selecting the highest-scoring motifs at each iteration. These motifs are then merged into a single CRN. For a selected phenotype v, the score for a k-mer motif of nodes $S \subset V(G)$ in a sample i weighting scenario j is calculated as:

$$s_M(S, v, i, j) = w_{1j} \cdot \sum_{u \in S} w_{t,P}(u, v) + w_{2j} \cdot \sum_{u \in S} \frac{1}{\ell(\sigma(u, v))} + w_{3j} \cdot \sum_{u \in S} s_i(u) \tag{5.1}$$

with $\{w_1; w_2; w_3\} \subseteq \{0.33; 0.66; 1.0\}$. In addition, I integrated the functionality for creating interactive CRNs into the UI of the AirOmics tool of the MINERVA AIR plugin. After performing a downstream enrichment analysis on phenotypes, the user can select a phenotype and sample for which a CRN should be generated. Additionally, the tool includes options to filter the motifs by a maximum $\ell(\sigma(u, v))$ and define the number of motifs in the CRN.

5.2.3 2DEA Predicts Treatment Effects

Figure 31 shows selected upregulated (red) or downregulated (blue) processes/phenotypes at each time point during the four phases of acute inflammation described in the AIR as predicted by the 2DEA, using either all DEGs or only the unique DEGs in each treatment condition. By comparison, at time point 120h, many inflammation resolution processes were downregulated in the diclofenac treatment while being upregulated in the Tr14 treatment group; most of them were related to immune cell type activation. Treatment with Tr14 resulted in limited gene expression changes at 12h and 24h, but at 120 h, the effect peaked, especially on processes/phenotypes associated with acute inflammation resolution. Among diclofenac-treated mice, most of the selected acute inflammatory processes and phenotypes were affected at early time points compared with placebo-treated animals. In the diclofenac group, the highest number of significantly differentially regulated phenotypes occurred at 36h. Interestingly, there were only small differences between the predicted phenotypes for both DEG sets indicating that phenotype enrichment was driven mainly by the unique DEGs. These findings further argue for a fundamental difference in the mode of action of both treatments.

The 2DEA revealed that Tr14 treatment potentially influences neutrophil- and macrophage-related pathways. Specifically, Tr14 downregulated NETosis by 96h, while the 120h mark was characterized by the induction of apoptosis- and phagocytosis-related cytokines and receptors. Drawing on these findings, we hypothesize that Tr14's downregulation of proinflammatory NETosis genes leads to an extension in neutrophil survival, an increase in neutrophil marker genes, and the initiation of neutrophil apoptosis. This process might set the stage for the subsequent phagocytosis of apoptotic neutrophils by macrophages – a process termed efferocytosis – recognized as a critical step towards the resolution of inflammation and is known to stimulate tissue cleansing and repair.

When examining the CRNs for each phenotype, four processes, in particular, showed a substantial difference between the two treatments: "M2 phenotype and behavior" (Figure 31E and F), "apoptotic process", "apoptotic cell clearance" (efferocytosis), and "NETosis" (shown in the supplementary data of Hoch et al., 2023 [5]). Whereas we observed downregulation of NETosis-inducing genes, such as PADI4, by Tr14 after 96h, Tr14 also upregulated apoptosis-related genes (CASP1, CASP3, CASP7, and CASP8) and apoptosis-inducing receptors (Fpr2) after 120h. At the same time point, Tr14 treatment resulted in a general upregulation of neutrophil marker genes (ITGAM, NCF1, and NCF2). The expression of efferocytosis and M2 macrophage cytokine markers IL4, IL10, and IL13 were too low to be detected in any of the treatments and time points. However, we see significant upregulation of many related receptor genes at 120h by Tr14, including the IL2RG subunit of the IL4 receptor, the IL10RA subunit of the IL10 receptor, and the IL13RA1 subunit of the IL13 receptor. An activation of efferocytosis by Tr14 would go hand in hand with its cytokine profile and the strong upregulation of phagocytotic markers. On the other site, diclofenac downregulated IL2RG and IL10RA at 120h and the IL4RA subunit of the IL4 and IL13 receptor at 36h. At 120h diclofenac additionally downregulated FPR2 and upregulated *PADI4*. These results indicate that the neutrophil-macrophage axis is a central part in the different modes of action between Tr14 and diclofenac.

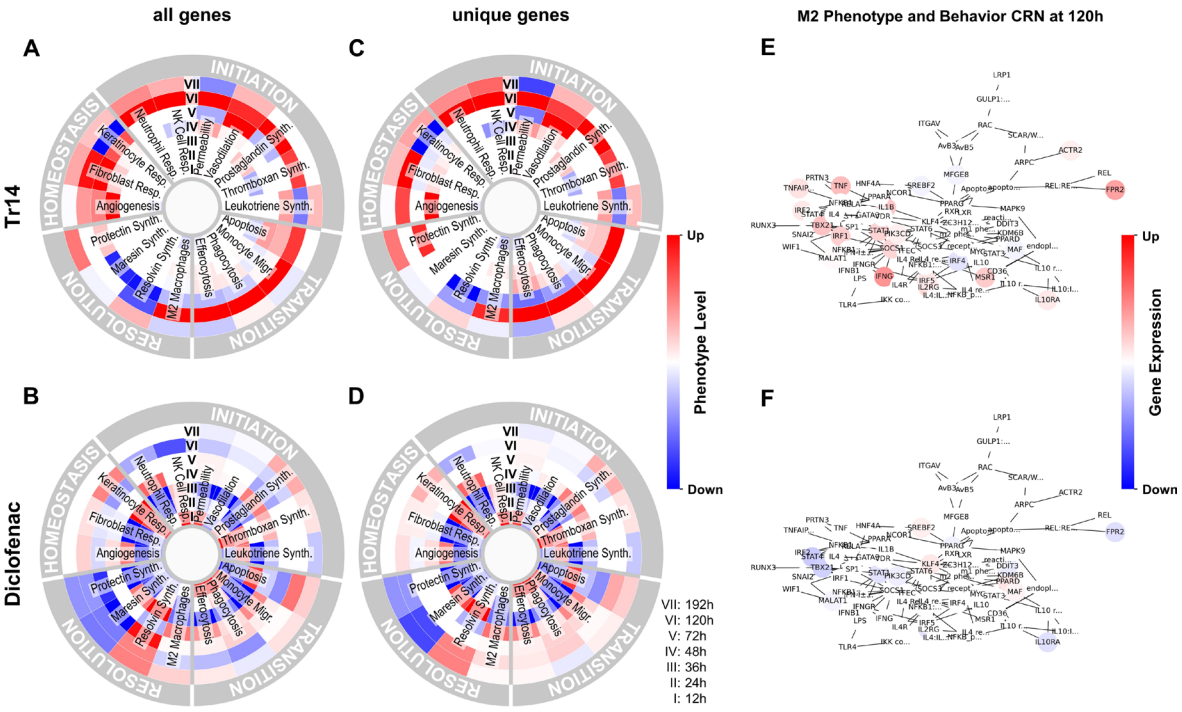

Figure 31: Impact on selected acute inflammatory processes and phenotypes in Tr14 vs. saline control (A and C) and diclofenac vs. placebo control (B and D). (A-D) The processes and phenotype levels were normalized between +1 (upregulation; red color) and -1 (downregulation; blue color). Acute inflammatory processes and phenotypes were grouped into 4 phases (inflammation initiation, transition, resolution, and homeostasis).

Circles from inner to outer regions represent treatment time points 12h, 24h, 36h, 72h, 96h, 120h, and 192h. (E-F) Knowledge Graph (KG)- and expression-based motif ranking create a central regulatory network (CRN) representing the molecular interaction associated with the selected phenotype node (e.g., M2 Phenotype and Behavior) for each process at a given time point and treatment. The CRN highlights the up-regulated (red) or down-regulated (blue) differentially expressed genes (adj. *p*-value < 0.05) in the sample. From Hoch *et al.*, 2023.

5.2.4 Summary

Upon comparing the two treatments, we found opposing responses and temporal differences, pointing towards notably different pharmacodynamics between single-target and multi-target drugs in the context of inflammation resolution. Unlike diclofenac, Tr14 did not suppress the expression of proinflammatory genes early in the acute inflammation timeline but instead supported the expression of these genes at a later stage. Tr14 induced opposite transcriptional changes compared to diclofenac, especially at 120h. Conversely, some processes induced by Tr14 at 120h are also induced by diclofenac, however, already at 36h. One explanation may be that the early inhibitory effect of diclofenac on inflammation causes some processes to shift in their timely activation while others remain blocked. Our observations suggest that Tr14 strengthens the late physiological immune response otherwise downregulated at an earlier stage by the anti-inflammatory drug diclofenac. The difference in the phenotypic effects of the two treatments may have been caused by their fundamentally different pharmacodynamic nature. Diclofenac, as an NSAID, has a direct, potent inhibitory effect on cyclooxygenase enzymes (PTGS1 and PTGS2), leading to noticeable changes in downstream signaling and metabolic cascades associated with SPM biosynthesis [296], [303], [304]. Following administration, an initial effect on early gene transcription continued over time. By comparison, the multi-component drug Tr14 appears initially to have a lesser effect. Tr14, as a multi-component natural product, presumably modulates the SPM biosynthesis or its effects through multitarget mechanisms. Consequently, the early effect of Tr14 on the lipid mediator pathway on individual targets might not be directly detectable at the transcriptional level, especially in bulk tissue samples.

Accumulating evidence suggests that a pro-inflammatory phenotype at the early stages of acute inflammation is essential to promote inflammation resolution and restore tissue homeostasis [174]. Early and immediate suppression of pro-inflammatory signals has been shown to cause various long-term chronic complications, suggesting that events occurring during the early acute inflammatory phase are needed for tissue healing [305], [306]. During acute inflammation, certain inflammatory cells, including neutrophils and

macrophages, undergo functional repolarization to acquire phenotypes contributing to the onset of inflammation resolution. Additionally, some mediators that initially promote the proinflammatory phase, including PGE-2, can switch roles to initiate a program for active resolution [307]. Whether different mediators act in a proinflammatory, anti-inflammatory, or pro-resolution manner is determined in part by their spatiotemporal relationships with other cells and the surrounding microenvironment during the entire time course of acute inflammation. Assuming that Tr14 acts simultaneously and slowly on multiple molecular targets, we hypothesized that small changes in regulatory components accumulate over time and lead to significant late modulation of the inflammatory response without disrupting important initial processes [308], [309], [310]. We suggested that using multitarget drugs with smaller but longer-lasting influences on different cellular processes could be of greater clinical value in reducing inflammation and improving inflammation resolution over time than drugs with a strong, early inhibitory effect.

Using the AIR, I examined altered gene expression associated with inflammatory processes and cellular profiles. Comparing the two treatments, I found opposing responses and temporal differences, suggesting markedly different pharmacodynamics of multitarget and single-target drugs in resolving inflammation. The KG-based enrichment analysis I developed facilitated the identification of genes with high relevance to each process.

5.3 Knowledge Graph Investigation of Drug Interactomes

In the next part of the project, we shifted our focus to analyze a Tr14 interactome of biologically active molecular compounds and their protein targets in the immune response.

5.3.1 Phenotypic Effects from Drug Interactomes

An interactome was compiled from public scientific literature and databases in a separate project. This interactome is integrated as a second dataset and mapped to corresponding nodes $V_{d,I} \subset V(G)$ with a signal $s_I(u)$ for every $u \in V_{d,I}$ representing the impact of Tr14 on u, either positive for activation or negative for inhibition, comparable to an FC value. I predicted the phenotype levels $\{s_{p,I}(v) | v \in V_p\}$ from identified Tr14 targets data using the 2DEA. The resulting predictions showed similarities to the RNA-Seq analyses, especially in the upregulation of pro-resolving processes (M2 macrophages, neutrophil apoptosis, efferocytosis, synthesis of maresin and protectins) and the downregulation of pro-

inflammatory processes (cytokine release, M1 macrophages, and PIM synthesis). Apoptosis is the most (positively) affected phenotype, with 31 related genes targeted by Tr14, of which 28 are modulated in the direction that supports the induction apoptosis, i.e., a positive impact on positive modulators and vice versa. Of the negatively affected phenotypes, "cytokine production involved in inflammatory response" and "M1 phenotype and behavior" are predicted to have the strongest change in activity, with 34 and 33 targeted genes, respectively. The strongest overlaps are visible at time points from 72h onwards. Most overlaps occur for the "apoptotic process" phenotype, which is predicted to be upregulated in both datasets at the 72h, 96h, and 120h time points. Tr14 affects proteins involved in the cellular processes of neutrophils and macrophages while simultaneously inhibiting the production and release of proinflammatory mediators. The findings indicate that, based on the Tr14 interactome, its main mode of action might be the modulation of the cellular immune response by modulating apoptotic and efferocytotic processes.

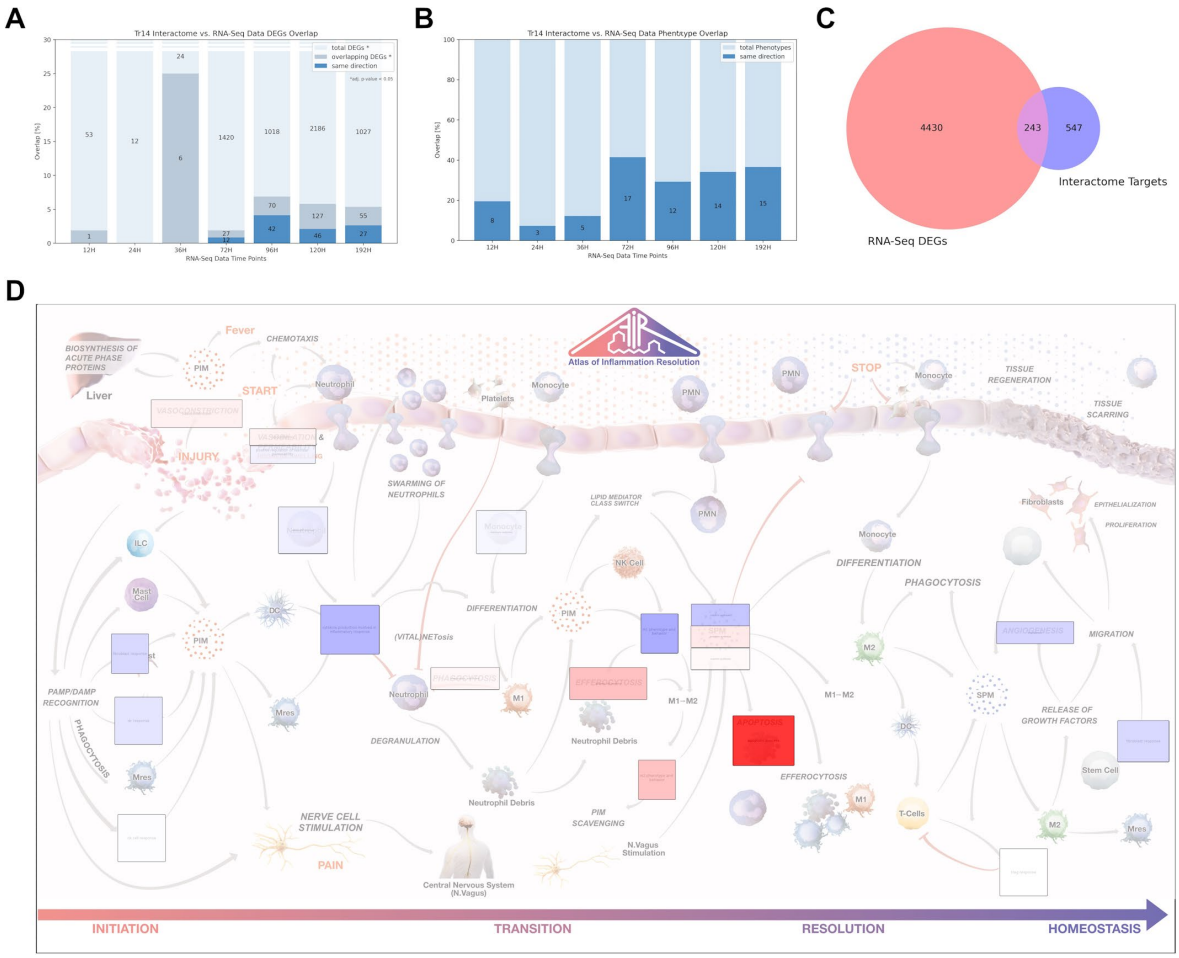

Figure 32: Overlap between Tr14 interactome and RNA-Seq Data. (A) Overlap of differentially expressed genes (adj. p-value < 0.05) at each time point with interactome targets. (B) Overlap of phenotype levels predicted from the RNA-Seq data at each time point with those predicted from the interactome. (C) Total overlap of unique genes and targets across all time points. (D) Predicted phenotype levels of the Tr14 interactome data using the 2DEA on the AIR.

5.3.2 Modulation of Drug Responses through Pharmacological Priming

In the next analysis step, I wanted to investigate whether overlaps exist between genes regulated by diclofenac and gene products directly targeted by diclofenac. Combining both pieces of information could evaluate whether a potential combined application, simultaneously or delayed, improves treatment outcomes. The idea behind the analysis is to identify if the differential gene expression induced by diclofenac on Tr14 targets can, and if so, to what extent, alter its biological effects.

First, I overlapped the predicted drug targets from the interactome with the DEGs from the RNA-Seq analysis in all samples. The absolute and relative overlaps at each time point i were identified using three approaches with varying weightings:

(i) counting the number of DEGs (Figure 33 A and D),

$$\left|V_{d,I} \cap V_{d,T,i}\right| \tag{5.2}$$

(ii) aggregating their absolute FC value (Figure 33 B and E)

$$\sum_{u \in V_I \cap V_{T,i}} \left|s_{T,i}(u)\right| \tag{5.3}$$

(iii) aggregating the FC value weighted by aggregated topological weightings of the gene on all phenotypes in the AIR (Figure 33 C and F).

$$\sum_{u \in V_{d,I} \cap V_{d,T,i}} \left(\left|s_{T,i}(u)\right| \cdot \sum_{v \in V_p} |w_t(u,v)| \right) \tag{5.4}$$

Additionally, the overlaps are calculated again only for nodes with the same sign of s_I and $s_{T,i}$, i.e., $\{u \in V_I \cap V_{T,i} | s_I(u) \cdot s_{T,i}(u) = 1\}$, highlighted in red in Figure 33. It shows that most genes overlap at 36h, which is to be expected as this time point contains the most DEGs. However, overlapping genes at the 24h time point appear to play a more important role in inflammatory processes. Interestingly, relative to the total number of DEGs, all time points show a similar overlap with the Tr14 interactome, which leaned more towards later time points when weighting by FC values and again towards the 24h when weighting by the gene's impact on inflammatory processes.

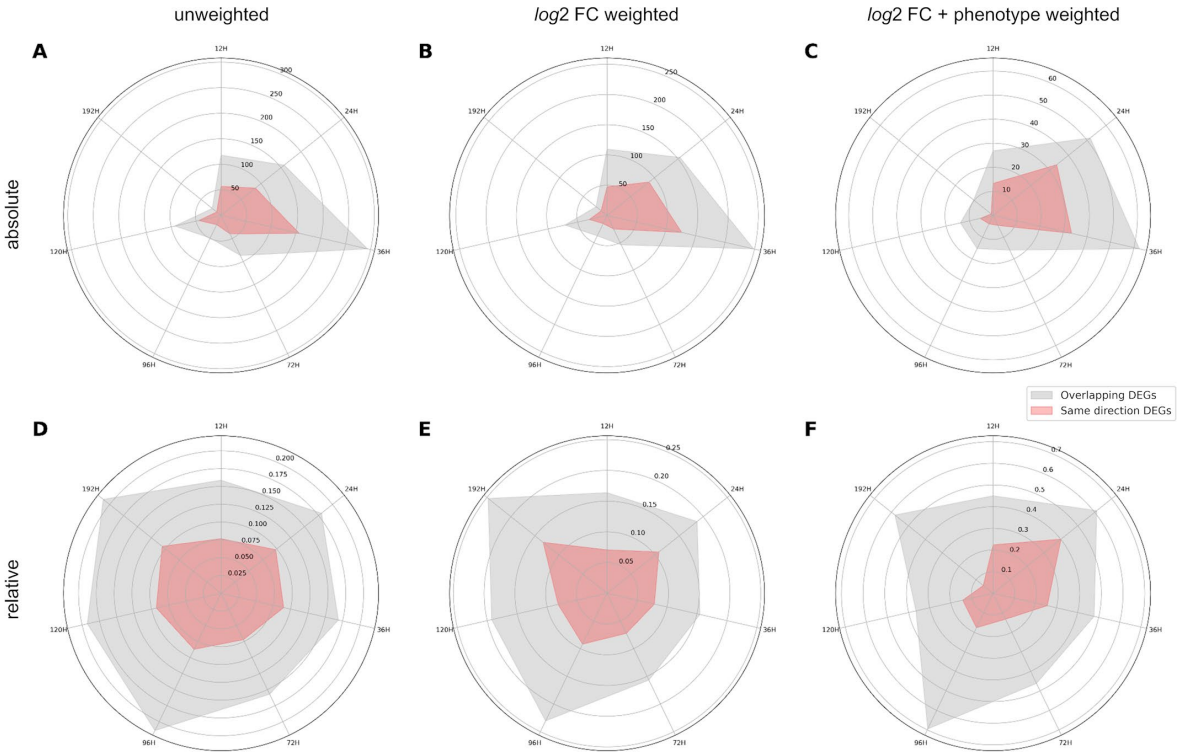

Figure 33: Overlap of protein targets from the Tr14 interactome with their respective coding differentially expressed genes (adj. p-value < 0.05, DEGs) in the RNA-Seq Data after diclofenac treatment at seven time points. Values are presented as the sum of absolute signal values of overlapping genes (A-C) or relative to the sum of all DEGs (D-F). DEGs are either counted (A, D), summed by their absolute signal value (B, E), or additionally weighted by their topological score in process-specific knowledge graphs from the "Atlas of Inflammation Resolution" (C, F).

Considering the transcriptional effects of diclofenac, I evaluated the potential amplified efficacy when diclofenac is applied before Tr14. To assess the impact of differential gene expression of diclofenac, I weighted the regulatory score s of the interactome targets based on its direction and the direction of the corresponding signal value from the data. Targets with the same direction of regulation are enhanced, while those with the opposite direction are reduced.

$$s_I{}'(u) = s_I(u) \cdot \begin{cases} 2^{|s_{D,i}(u)|}, & s_I(u) \cdot s_{D,i}(u) > 0 \\ 2^{-|s_{D,i}(u)|}, & otherwise \end{cases} \tag{5.5}$$

and

$$\Delta s_I(u) = s_I{}'(u) - s_I(u) \tag{5.6}$$

It shows that most overlapping genes are downregulated by diclofenac as a consequence of its primarily inhibitory effect. Consequently, many targets are downregulated and inhibited by Tr14 compounds, mostly pro-inflammatory cytokines and TFs, including IL1B, IL6, and STAT1. Of the targets with positive scores that are upregulated by Tr14, mainly those of the 12h are involved in signaling pathways of

biological functions, namely glucose metabolism (PGK1), SPM synthesis (GPX1), and apoptosis (CASP8, CASP9, APAF1). Next, I re-analyzed the effect of overlapping targets of inflammatory phenotypes using the AirOmics tool of the AIR plugins. I compared the impact of weighting the interactome with the diclofenac RNA-Seq data (Figure 34). Similar to the changes on the molecular level, I observed a primarily enhanced negative impact on pro-inflammatory processes, such as prostaglandin and thromboxane synthesis, and fibroblast response, with the most substantial change at the 24h time point. Positively affected processes included leukotriene synthesis, apoptotic process, resolvin synthesis, and keratinocyte response, but are less profound than the downregulated ones. Other time points only show marginal changes compared to earlier time points due to their low FC values in the diclofenac data and/or low impact on inflammatory processes. The effects of Tr14 on the later time points 120H and 192h especially seem to barely overlap with diclofenac treatment.

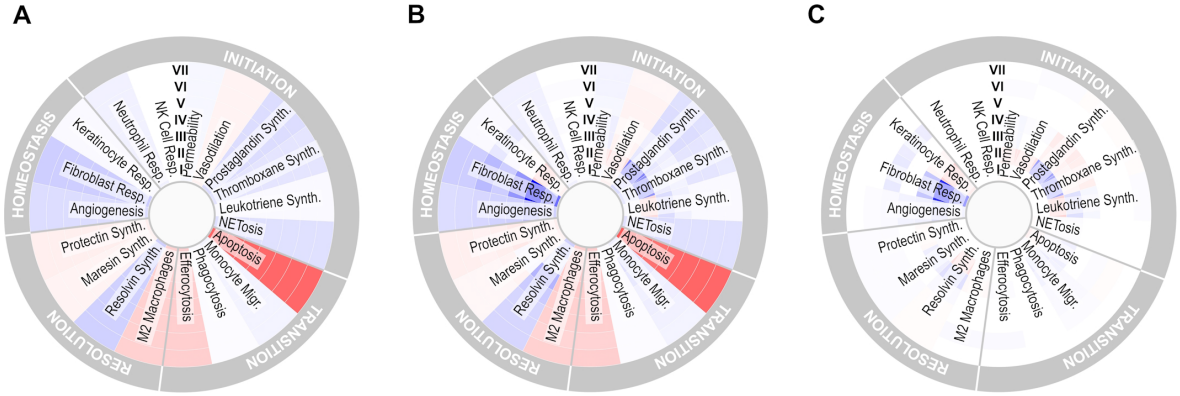

Figure 34: Estimated effect of the Tr14 targets overlapping with differentially expressed genes from diclofenac RNA-Seq data on inflammatory processes using the "Atlas of Inflammation Resolution." Shown are the phenotype activities before (A) and after (B), weighting the regulatory scores of the Tr14 interactome with *log*2 fold change values from the DEGs and the difference between both (C).

5.3.3 Summary

Several challenges arise when comparing drug perturbation data with RNA-sequencing (RNA-seq) data. The first and foremost challenge is that the changes in protein activity due to drug perturbation do not necessarily correlate with changes in gene expression. The physiological response to drug perturbation can often be counterintuitive. For example, a drug that inhibits a specific protein might disrupt a negative feedback loop controlling the transcription of that protein's gene, leading to an increase rather than a decrease in gene expression. Consequently, we only observed a small overlap between the Tr14 interactome and the differential expression in the RNA-Seq data. A second and arguably more significant challenge involves the temporal aspects of biological responses

to drug effects. The changes observed in mRNA expression are not direct consequences of the drug perturbations but rather the outcome of complex signaling processes that operate at multiple biological levels. For instance, the effect of a drug like Tr14 is primarily caused by its binding to its target proteins. However, the observed phenotypic effects depend on the system's biological state at the time of drug administration.

Factors such as the composition of local cell types and their activity state at a given moment can significantly influence their reaction to drug effects. Such spatiotemporal state-dependent pharmacodynamics means that the tissue might need to reach a certain physiological or pathological state susceptible to the drug's mechanism of action before it can exert its therapeutic effect. The following factors could favor such a state: **(i) Cellular composition** at the injection site: Molecular compounds of the drug might bind to or are taken up by specific types of cells only. Because the immune response involves a highly time-dependent acquisition of immune cells, the drug effect might be more favorable at one phase. **(ii) drug accumulation**: It is possible that the drug only achieves its therapeutic concentration after multiple doses, leading to a delayed effect. **(iii) pharmacokinetics**: The activation and transport processes of the drug compounds might be slow, causing a delay in the drug's effects, and **(iv) biological response time**: Some biological processes, such as the initiation of apoptosis, can take up to 48 hours.

In summary, these issues underscore the complexity of drug responses and the challenges inherent in correlating drug effects with changes in gene expression. They also highlight the importance of considering the biological state and temporal dynamics when interpreting drug perturbation and RNA-seq data.

5.4 Strategies to Simulate Drug-Induced Gene Regulation

To overcome the concerns from the previous section about the limited dynamic processes in the analysis of drug perturbation data, in the last part of the Tr14 analysis, I aim to consider signaling mechanisms when comparing the two data. As described in Section 5.1, large-scale KGs, such as the AIR, could be employed to simulate the signal transduction from the direct targets downstream through the KG to TFs and, finally, to the regulated genes. The following sections describe ongoing ideas and strategies on how such an approach could be developed for AIR KG to simulate the Tr14 drug interactome and be computationally implemented in the framework from Section 2.5.

5.4.1 Simulating Signal Transduction

I employed the approach by Lee and Cho in 2018, which I also used for the analysis in Chapter 4, and adopted it to the specification of the AIR and the Tr14 data. Conversely to the methodology described in Section 4.3.1, I performed the simulation in the downstream direction, updating the signal $s(v,t)$ of a node v at step t based on the signal of incoming neighboring nodes $N_{in}(v)$ in the previous step. However, in contrast to the LM study, we were now faced with a highly connected large-scale KG, and there can be a high number of input signals from the Tr14 interactome distributed across the whole KG. Thus, I included the full equation from Lee and Cho, including its second part, which adjusts the node's signal towards its initial state.

$$s(v,t+1) = \alpha \cdot \sum_{u \in N_{in}(v)} \frac{\tau(u,v) \cdot s(u,t)}{\sqrt{c_{d_{in}}(u)} \cdot \sqrt{c_{d_{out}}(v)}} + (1-\alpha) \cdot s(v,0) \tag{5.7}$$

As an input for the simulation, I set $s(v,0) = s_I(v)$ for every $v \in V_I$ in the interactome dataset.

The simulation is then run until a steady state is reached, defined as the step t_s at which the accumulated difference to the previous step falls below a selected threshold.

$$t_s = min\left\{t \in \mathbb{N}: \sum_{v \in V} (|s(v,t) - s(v,t-1)|) < 10^{-6}\right\} \tag{5.8}$$

The final signal of a node $\bar{s}$ is then defined as the accumulated signals over all steps.

$$\bar{s} = \sum_{t=0}^{t_s} s(v,t) \tag{5.9}$$

Additionally, the end signal can also be defined as the signal at the steady state $\bar{s}_{ss}(v) = s(v,t_s)$.

5.4.2 Computational Implementation

In the LM analysis from Chapter 4, the KG was sparse as it only included LM synthesis pathways and TF interactions, and the number of steps was relatively short as the approach was designed to prioritize direct TFs. In contrast, in this study, I used the full large-scale KG from the AIR and simulated changes in gene expression regardless of the distance of the signal. Furthermore, the research question of this project required a large number of simulations (hereafter referred to as samples) to explore the many combinations

in the Tr14 formula and the permutations for statistical analysis, which is further described in Section 5.4.3 However, a simplification was also possible here as there is no gene expression data and the weightings are therefore sample-independent. I implemented the approach into the computational framework described in Section 2.5 so that several simulations can be performed simultaneously.

The computational design is schematically summarized in Figure 35A. The signals of nodes are stored in a 3D np array of size ($|V| \cdot T \cdot n$) with n being the number of simulations. The 2D array signal states at step t are aggregated with the topological weights, which, since their values are static, are generated as a constant 2D array of size ($|V| \cdot |V|$) at the beginning of the simulation. Both arrays are aggregated at each step by calculating their dot product. The new signal is then calculated by combining the aggregated array and the values of the initial state in proportion to the parameter α.

Similar to inferring how phenotypic effects by the Tr14 targets could be affected by prior diclofenac treatment in Section 5.3.2, a similar weighting could also be applied here. This way, the signal transduction through a system prestimulated by diclofenac can be simulated. The idea is that the absolute signal, and subsequently transduction to the next nodes, of a node with upregulated expression is increased if the signal is positive or decreased at a negative signal, and vice versa for reduced expression (Figure 35B). The signal of all nodes at each step is weighted by the FC values from the diclofenac data at each sample i is thus changed to:

$$s'(v,t) = s(v,t) \cdot 2^{s_{D,i}(u) \cdot sign(s(v,t))}$$

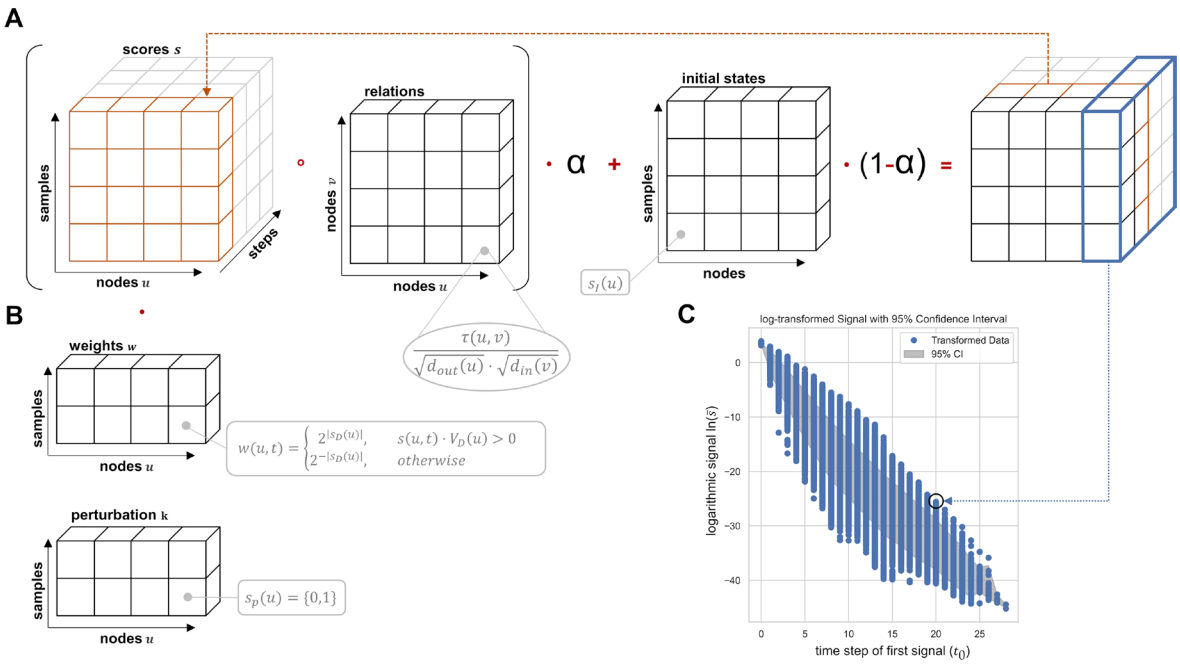

Figure 35: Computational implementation of the signal flow estimation algorithm from Lee & Cho, 2017, adapted for multiple runs (samples) with different conditions. d_{out}: out-degree; d_{in}: in-degree; r: relation of edge {-1,0,1}; α: hyperparameter set to 0.5.

5.4.3 Statistical Evaluation

The statistics in this study require a permutation solution to identify significant response values. Especially in highly interconnected KGs, nodes with a high degree may exhibit a strong signal regardless of the type and origin of the input stimulus. Secondly, at every step, the signal is normalized by a factor larger than one 1, causing signals that are further away from the inputs to have lower values. Consequently, a statistical evaluation would need to be based on a permutation of the input signals and consider the distance of the signal. Given that throughout its transmission, the signal is incrementally reduced by a defined value, the final signal can be described as a decay function, dependent on the step t_0 where a signal is first detected (= distance)

$$\bar{s}(v) \propto \left(\frac{\alpha}{w}\right)^{t_0(v)} \tag{5.10}$$

Indeed, if the signal $\bar{s}$ is plotted against t_0 on a logarithmic scale, as shown in Figure 35C, a negative linear correlation becomes visible. Thus, statistical analysis can be performed by permutation of the initial signals across the KG and generating a normal distribution of the signals at every distance. The p-values for all RNA nodes in the KG are then calculated by assessing the z-scores of the signal from the fitted Gaussian curve and adjusted for multiple testing using Benjamini-Hochberg FDR correction.

5.4.4 Limitations and Outlook

Logic-based models of large KGs are characterized by uncertainties due to the lack of mechanistic information [311]. Consequently, the simulation of signal transduction within these models will always be accompanied by some inaccuracies. Lee & Cho reported an accuracy of 60-80% when comparing their approach with experimental data [312]. Therefore, only highly significant signals likely to receive strong and potentially superimposed signals from the perturbations should be considered. On the other hand, this leads to many false negatives and prevents the prediction of expression patterns unaffected by the perturbations.

Secondly, the weighting of the simulation by additional RNA-Seq data, e.g., diclofenac, is limited by the biological relevance of the changes in expression in relation to the type and spatiotemporal context of the perturbation. Theoretically, if time-scale data is available, the weightings could be adjusted differently at different steps during the simulation, representing the points in time from the dataset. However, the biological significance is severely limited, as the interference from the input signal in reality would drastically affect gene expression. The empirical differential expression that was used for the weighting would, thus, no longer be representative. A possible solution could be to continuously adjust the empirical expression data by the simulated measurements, for example, by lowering the FC values of upregulated genes if they receive a negative signal in the simulation. However, validation would be a major challenge, as highly specific experimental data would be required. Nevertheless, such an approach could provide many details on the interactions between two conditions by effectively combining information from empirical and simulated data. The general possibility that such methods could be used in the future, provided that the most accurately curated KGs and validation data are available, motivates research into large-scale KGs in systems pharmacology.

When investigating potential strategies for influencing drugs, one should not only consider the effects predicted for a static KG but also try to understand the entire KG as a dynamic system. The group of Barzel *et al.* investigated the recoverability of KGs that are perturbated to the point of unresponsiveness due to changes in the topological structure [313]. One could apply their ideas to pharmacology by considering the transcriptional changes in diseases as a collapse of crucial functions. In this way, processes such as inflammation resolution in inflammatory diseases can be described as a "failed network," as it is referred to by Barzel *et al*. Their methods could provide insight into how the failed

processes can be reconstructed at the topological level and identify perturbations that "revive" these beneficial programs. Such an approach could complement the one described in the previous sections by identifying drug targets that can restore specific processes instead of predicting the effects of known perturbations.

Chapter 6

Logic-based Modeling of Systemic Diseases

6.1 Clinical Associations Between Sarcopenia and Malnutrition

Malnutrition (MN) is a common and characteristic feature of gastrointestinal diseases, such as liver cirrhosis (LC) and intestinal dysfunctions (ID), e.g., short bowel syndrome (SBS), and is associated with high mortality rates [314]. For LC patients, the prevalence of MN is indicated at up to 90% [315]; for patients suffering from SBS, it is around 10 to 40% [316]. Disease-related MN is closely related to mild, chronic inflammation [317]. Both MN and inflammation contribute to muscle wasting, which, combined with the loss of muscle function, can eventually result in **sarcopenia**. This vicious cycle of MN, inflammation, sarcopenia, and the underlying disease itself leads to an unfavorable prognosis for the patient [318]. A sufficient supply of energy and nutrients is needed for the homeostasis of muscle anabolism and catabolism. Conversely, an inadequate nutrient uptake by intestinal malabsorption, a deficient metabolism of nutrients, and a deficient breakdown of muscle waste products in the liver can impair muscle growth [319], [320]. Additionally, microbial invasion caused by a disrupted epithelial barrier in ID and LC leads to systemic inflammation that stimulates catabolic processes in the muscle [318], [321], [322]. The liver, as a primary producer of cytokines and hormones, also releases many pro-inflammatory mediators during injury that favor muscle atrophy [320], [323], [324]. The control of muscle physiology is consequently highly dependent on intestinal and liver function, making sarcopenia a common secondary phenomenon in ID and LC [318].

Given the physiological and pathophysiological association of intestine, liver, and muscle function, it is not surprising that they are linked by complex molecular communication processes [325], [326], [327]. Although the role of many molecules has been elucidated by extensive *in vitro* and *in vivo* experiments, understanding the system as a whole is challenging. Therefore, *in silico* approaches, i.e., converting available knowledge into KG formats, can help unravel this complexity. In the context of nutrition and

sarcopenia, models have already been developed to investigate various systems, such as nutrient absorption [328], muscle fiber physiology [329], [330], pathologic liver metabolism [331], and diabetes [332]. However, an approach that links gastrointestinal diseases, nutrition, and muscle (patho-)physiology on a larger scale and enables simulations across tissues has been lacking. In collaboration with the Department of Gastroenterology, we developed the "Sarcopenia Map," a Disease Map that describes the gastrointestinal and muscular processes through KGs and provides tools for exploring underlying molecular mechanisms through Boolean Modeling.

6.2 Designing and curating the "Sarcopenia Map"

Like inflammation, gastrointestinal processes also show non-linear behavior through complex mechanisms, such as intertwined hormonal regulation and nutrient cycling between organs. However, compared to the processes described in the AIR, a graph representing nutrition and sarcopenia has defined starting (food ingestion) and endpoints (muscle) from a spatial as well as temporal point of view. Starting with food ingesting, nutrients are digested, absorbed into the bloodstream, metabolized in the liver, and ultimately affect muscle growth and function. These considerations allow for a hierarchical design of the Disease Map. Also, from a methodology perspective, the Sarcopenia Map differs largely from the AIR. The latter was designed to perform EDAs on modulated processes and molecular patterns in the observed samples. It selects possibly affected processes, which are then evaluated in the data through KG-based enrichment approaches like the 2DEA.

In contrast, the analysis becomes more targeted in the Sarcopenia Map, assuming that all the curated processes in the KG are mechanistically relevant. Instead, the map should provide insights into how the intrinsic mechanisms are affected in response to changing conditions, such as an altered diet or disease states. Experimental data is thus not the primary focus, mainly because the dataset would need to combine omics profiles from multiple tissues and organs for patients in various disease states. Data provided by the user are instead the specifications of the underlying (mainly clinical) conditions, such as a nutrient profile, genomic variations, e.g., in digestive enzymes, and dysregulations through diseases.

The Sarcopenia Map was developed entirely using a top-down approach, manually curating the KG from the higher-level intestine, liver, and muscle processes, specifically for

LC, ID, and sarcopenia. We screened the PubMed database for published literature focusing on recent reviews describing the intestinal uptake of nutrients and their metabolism in the liver, hormonal communication between the liver and muscle, and regulation of muscle growth and function. Simultaneously, we sought information on the effects of ID and LC on these processes. The information was then further examined to ensure that the interactions identified were direct, such as protein-receptor interactions. We collected the information in three SBML-standardized submaps in CellDesigner to improve clarity and ease curation efforts. Intracellular molecules were enclosed in compartments reflecting the organ. In contrast, extracellular molecules were placed outside the compartments, representing molecules in the bloodstream (e.g., nutrients or cytokines) or systemic conditions, such as acidosis or hyperammonemia. This separation distinguishes tissue-specific processes and their communication through secreted molecules. Figure 36A provides a schematic overview of the map organization and the hierarchical flow of information through the submaps. Similar to the AIR, the Sarcopenia Map is curated in combined AF and PD formats. More extensive metabolic pathways, e.g., glycolysis, have been combined into a single catalytic reaction leading from the initial reactant (glucose) to the final product (pyruvate), omitting all intermediates. The reaction is catalyzed by a phenotype (glycolysis) representing the metabolic pathway per se. All regulations, e.g., product-feedback inhibitors or hormonal ones, were added as reactions to the phenotype (Figure 36B).

Table 6: Overview of the submaps currently included in the Sarcopenia Map

Submap Name	Edges	Nodes	Unique Nodes
Intestine	264	251	218
Muscle	224	152	119
Liver	165	129	87

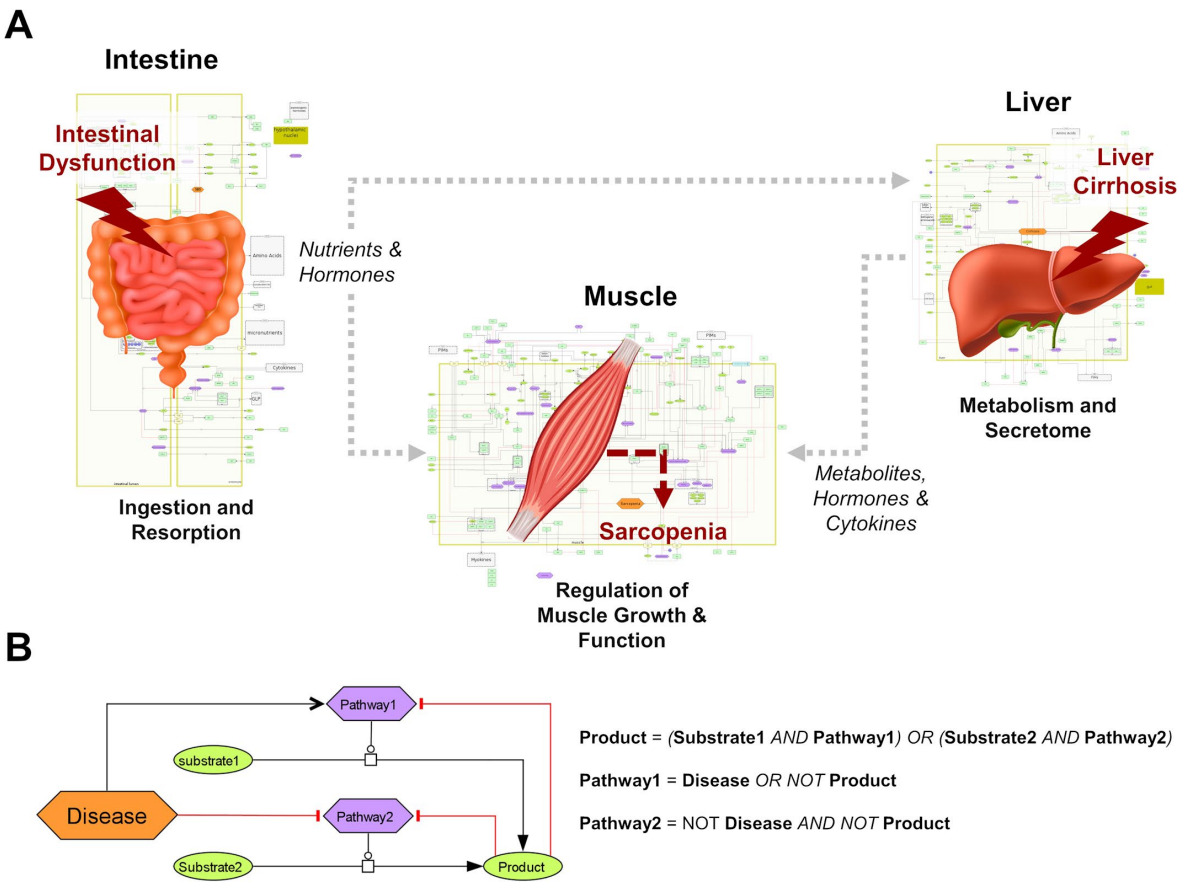

Figure 36: Overview of the hierarchical organization of the Sarcopenia Map. (A) We summarized information on molecular interactions related to sarcopenia from the literature into three tissue-specific submaps. In addition, we integrated the effects of liver cirrhosis (LC) and intestinal dysfunction (ID) on these molecular processes. (B) Combined activity flow and process description formats for reduced representation of molecular pathways and disease interactions. Boolean rules define a node's state by converting SBML reactions into logical gates.

6.3 A Systemic Boolean Model

During the design of the Sarcopenia Map, I faced challenges in defining logical rules that enable a comprehensive model of nutrient resorption, metabolism, storage, and hormonal regulation. Because many metabolic processes are regulated depending on the available quantities of metabolites, I developed the model using multivalued logic rules. A node $v \in V(G)$ is defined by a multivalued state $s(v,t) \in \{0,\dots,s_{max}(v)\}$ where is $s_{max}(v)$ the maximum value explicitly defined for v with $s_{max}(v) = 1$ as the default value.

Additionally, the node v can be perturbed by a perturbation state $s_p(v,t) \in \{-1,0,1\}$. Thus, when the state of v is accessed by others, its actual state $s(v,t)$ is expressed as a Boolean pseudo-state $s'(u,t)$ defined as:

$$s'(v,t) = \begin{cases} true & \text{if } s_p(v,t) = 1 \\ false & \text{if } s_p(v,t) = -1 \\ false & \text{if } s_p(v,t) = 0 \wedge s(v,t) = 0 \\ true & \text{otherwise} \end{cases} \tag{6.1}$$

The state of the node is updated and either decreased or increased depending on the Boolean value of a logical function f_v, defining the new state at time $t+1$ as:

$$s(v,t+1) = \min\left(s_{max}(v), \max\left(0, s(v,t) + \begin{cases} 1, & \text{if } f_v(v,t) = true \\ -1, & \text{otherwise} \end{cases}\right)\right) \tag{6.2}$$

The logical rule defining the new state of a node at a given time is defined as the function f_v depending on logical functions f_e of edges connected to v:

$$f_v(v,t) = \left(\bigwedge_{e \in E_{in}^-(v)} \neg f_e(e,t)\right) \wedge \left((E_{in}^+(v) = \emptyset) \vee \bigvee_{e \in E_{in}^+(v)} f_e(e,t)\right) \tag{6.3}$$

with $E_{in}^+(v)$ and $E_{in}^-(v)$ being the set of incoming edges with a positive and negative interaction type, respectively, in which v is a target node.

$$E_{in}^+(v) = \{e \in E(G) | v \in S_v(e) \wedge \tau_e(e) = 1\}$$
$$E_{in}^-(v) = \{e \in E(G) | v \in S_v(e) \wedge \tau_e(e) = -1\}$$

The functions f_e themselves are defined by logical rules evaluating the states of source nodes and modifications in the edge. The logical rule of an edge $e \in E(G)$ is evaluated to **`true`** only if all source nodes $S_u(e)$ are ON. Secondly, each modifier $m \in \mathbf{S}_m$ for each $(\mathbf{S}_m, \tau_m) \in M_e(e)$ associated with e has a type τ_m, either -1 or 1. Let $\mathbf{S}_m^-(e) = \cup\{\mathbf{S}_m \mid (\mathbf{S}_m, -1) \in \mathbf{M}_e(e)\}$ represent the set of modifiers with a negative type and $\mathbf{S}_m^+(e) = \cup\{\mathbf{S}_m \mid (\mathbf{S}_m, 1, e) \in \mathbf{M}_e(e)\}$ represent the set of modifiers with a positive type for the edge e. All modifiers m in $\mathbf{S}_m^-(e)$ must be false, and, if one exists, at least one modifier m in $\mathbf{S}_m^+(e)$ must be true for the interaction e to occur. The Boolean function $f_e(t)$ for edge e at time t, considering multiple source nodes $\mathrm{S_u}(e)$ and the modifier conditions are thus defined as:

$$f_e(e,t) = \left(\bigwedge_{u \in S_u(e)} s'(u,t)\right) \wedge \left(\bigwedge_{m \in \boldsymbol{S}_m^-(e)} \neg s'(m,t)\right) \wedge \left((\boldsymbol{S}_m^+(e) = \emptyset) \vee \bigvee_{m \in \boldsymbol{S}_m^+(e)} s'(m,t)\right) \tag{6.4}$$

For any node v, the percentage of time steps where the node is ON over a total number of steps T, in the following referred to as their **activity**, is defined as:

$$\bar{s}(v) = \sum_{t=1}^{T} \frac{s'(v,t)}{s_{max}(v)} \tag{6.5}$$

Measuring the activity $\bar{s}(v)$ under varying conditions allows to assess the sensitivity v towards these perturbations. Let v_p be a node that is being perturbed, and for n number of Boolean simulations of T steps each, vary the perturbation strength, denoted as $\alpha \in [0,1]$, from 0% to 100%. Thus, $\alpha \in \{\alpha_1, \alpha_2, \dots, \alpha_N\}$ where each $\alpha_i = \frac{i-1}{n-1}$ for $i = 1,2, \dots, n$. During each simulation, v_p is perturbed for a total of $\lceil \alpha_i \cdot T \rceil$ steps (rounded to the nearest higher integer) equally distributed over all T defined as:

$$s_p(v,t) = \begin{cases} C_{p,v} & \text{if } t \; mod \left\lceil \frac{T}{\alpha_i} \right\rceil = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

with the perturbation value $C_{p,v} = 1$ or $C_{p,v} = -1$, depending on the scientific question. For any node v, $\bar{s}(v, \alpha_i)$ represents the state activity of the current perturbation strength α_i. The correlation between $\bar{s}(v_p, \alpha_i)$ and $\bar{s}(v, \alpha_i)$ for every other node v in the KG for all α_i can be calculated using the Pearson correlation coefficient r. Assuming the two arrays of activities $\bar{s}$ for v_p and v are defined as $A = [\bar{s}(v_p, \alpha_1), \dots, \bar{s}(v_p, \alpha_N)]$ and $B = [\bar{s}(v, \alpha_1), \dots, \bar{s}(v, \alpha_N)]$, respectively, the correlation is defined as:

$$r\,(v_p, v) = \frac{n \sum A_i B_i - \sum A_i \sum B_i}{\sqrt{\left(n \sum A_i^2 - (\sum A_i)^2\right)\left(n \sum B_i^2 - (\sum B_i)^2\right)}} \tag{6.7}$$

6.4 Computational Implementation

A boolean simulation is run by calling the **run_boolean** from a **`Model`** object (Figure 37A). The function accepts the number of steps T and any node perturbations as an input. In the **`Model`** object, Node states and perturbations are each stored in a 2D Boolean numpy array of shape $|V| \times T$ that are initiated at the start of **run_boolean** to preallocate RAM. For each step t of T, a function **`boolean_step`** of the same **`Model`** object is called. In **`boolean_step`**, the states of each **`Node`** object v in the model are updated, and their new states are added to the 2D array by the node's id and the current step. The **`Node`** function **update_state** equals in Equation (**6.2**), updating its state $s(v,t)$ based on $f_v(v, t-1)$ from Equation (6.3). The function f_v is evaluated through the the **boolean_rule** function of a Node, which logically connects the **active** function of incoming nodes u, representing $s'(u, t-1)$ from Equation (6.1) and returning a Boolean value. The **boolean_rule** function was created dynamically on initialization of the **`Model`** object by combining strings of the logical rules for each incoming edge of the node and evaluating the string into a lambda function through the

Python **eval** method (Figure 37C). The **Edge** and **Modification** classes have an as_Boolean_string() function that returns a string representation of logical links between nodes (Figure 37D). In these Boolean strings, each node is denoted as "node#", where "#" is a unique node ID derived from the node's index within the model's node dictionary. This way, the Boolean strings can be created dynamically by nesting the strings of edge modifications into the edge's Boolean string and finally into the Boolean string of the node. Finally, in an eval function, the string representations of the node IDs are mapped to the node objects and, consequently, their **active** function.

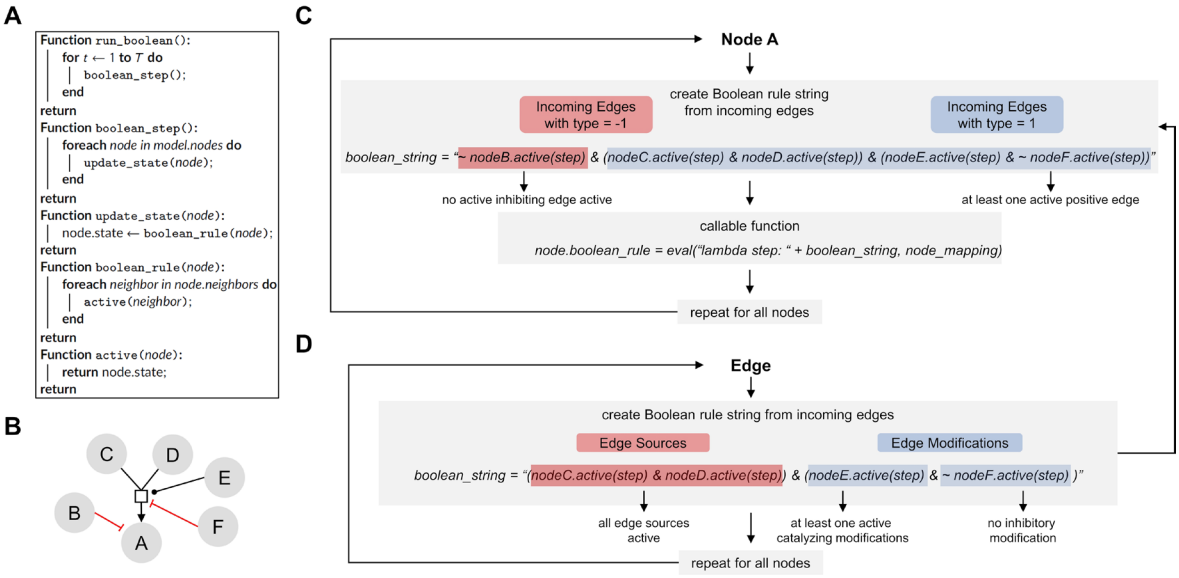

Figure 37: Computational workflow to dynamically generate Boolean logic from knowledge graphs (KG). (A) Boolean simulations from the Model class are performed iteratively in each step, updating each node's state by calling its Boolean_rule() function. (B) Exemplary KG in process description from which a Boolean rule is generated in (C) and (D). (C) For each node, the Boolean logic is combined with the Boolean logic of all incoming edges in a string. The string is then evaluated into a Python lambda function, and node placeholders are mapped to corresponding node objects. (D) A Boolean logic string is created for each edge by logically combining the source state and modification nodes' state through their active() function

6.5 Simulating Systemic Molecular Processes

To test the Boolean model, I studied the behavior of the carbohydrate system under different nutrition states, i.e., different active frequencies of the food intake node. Carbohydrate metabolism is a tightly regulated system and the central part of the energy cycle that controls muscle function. Therefore, the carbohydrate system is a crucial pathway linking LC and ID to sarcopenia, as carbohydrate resorption, storage, and usage are impaired in these diseases [333], [334]. In clinical settings, glucose supplementation has been shown to reduce muscle mass loss, while glycogen depletion has been identified as a major cause of the development of sarcopenia in LC patients [335], [336]. A sufficient model

of the Sarcopenia Map must, therefore, ensure that carbohydrate activities respond correctly to changing nutritional conditions and disturbances.

First, I measured the response of glucose and glycogen to altered nutritional stimuli. Figure 38A shows the extent of hepatic glycogen storage (blue dots) and blood glucose (red dots) in response to increasing food intake (y-axis, black dots). I observed increasing hepatic glycogen activity and its prolonged conversion to blood glucose after food intake was switched off. Blood glucose is continuously active as long as food intake occurs and oscillates during glycogen depletion. These results show that our model can simulate the conversion of glycogen to glucose and its release into the bloodstream in fasting situations.

Next, I measured carbohydrate behavior again, but with different combinations of ON and OFF food intake, representing changing frequency and quantity, but not quality, of diet. I identified three specific **nutrition states**, which will act as input for the model to simulate (patho-)physiological behavior. Importantly, Boolean models use steps as a discrete and arbitrary time measurement and cannot simulate real time-scale. Here, we defined the nutrition states by their impact on the carbohydrate system (Figure 38B): (i) **undernourished**, i.e., long fasting periods with complete depletion of glycogen storage (5 ON-steps and 25 OFF-steps), (ii) **well-nourished**, with continuous glycogen storage (5 ON-steps and 10 OFF-steps), and (iii) **overnourished**, with continuously increasing glycogen (5 ON-steps and 2 OFF-steps). I incorporated these states into the UI of the Sarcopenia Map plugin to facilitate their comparison when running different simulations. These nutritional states differ only in the quantity of food, not its composition, and are assumed to contain all macro- and micronutrients. However, users of the map can perturb nodes in the intestine submap to change the composition of the diet individually.

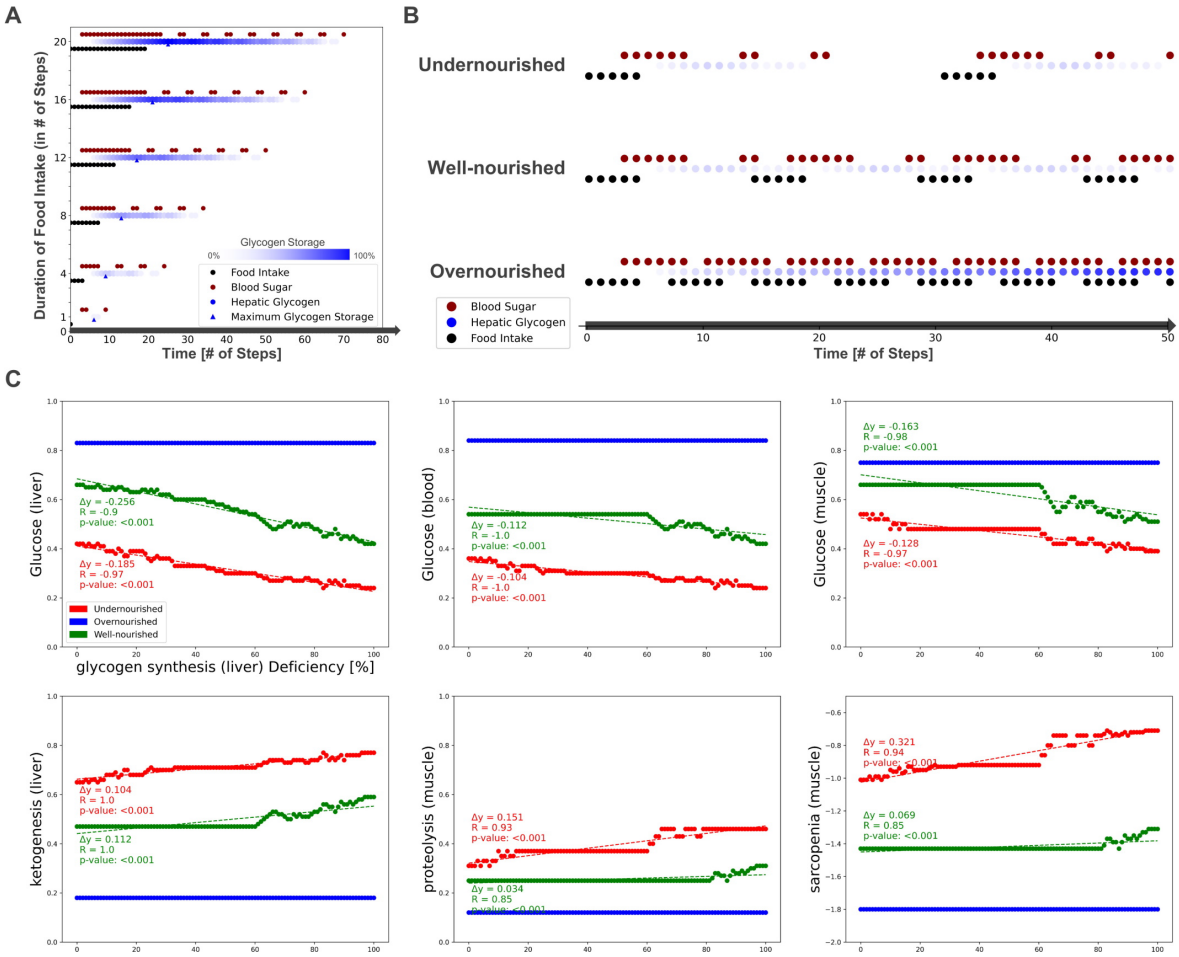

Figure 38: Testing the model by simulating different nutrition states and carbohydrate availability. (A) The activity of hepatic glycogen storage and extracellular glucose depends on the duration of the food intake stimulus. (B) Definition of three nutrition states by their food intake frequency and the resulting activities of hepatic glycogen storage and extracellular glucose. (C) Predicted activities of selected nodes in response to an increasing hepatic glycogen synthase deficiency.

After testing the model under physiological conditions, I simulated pathophysiological disease states by molecular perturbations. I investigated how a deficiency of glycogen synthase (GS) in the liver correlates with the activity of glucose in the liver, blood, and muscle, ketogenesis in the liver, and proteolysis and sarcopenia in the muscle (Figure 38C). Most noticeably, in the overnourished state, GS deficiency does not correlate with any of the nodes, and in the well-nourished state, a correlation becomes visible only at high inactivation. The latter is probably caused by the compensation of a lower GS deficiency due to increased blood glucose due to a more frequent food intake than the undernourished state. GS deficiency correlates negatively with glucose activity, which is most prominent in the liver compartment and less in the blood and muscle compartments. In the well-nourished state, glucose activity in the muscle shows a large plateau at medium GS deficiencies (20-60%), possibly due to compensation by muscle glycogen. A positive correlation is visible for the 'ketogenesis' phenotype in the liver and

'proteolysis' in the muscle, both physiological responses to hypoglycemic states [336], [337]. Interestingly, the plot for "sarcopenia" also shows a positive correlation and is very similar to that for "proteolysis," suggesting that sarcopenia in GS deficiency is most likely mediated by increased activity of muscle proteolysis.

I conducted additional simulations for deficient glycogenolysis (Figure 39A), deficient glucose uptake in the muscle (GLUT4/SLC2A4, Figure 39B), and deficient glucose resorption in the intestine (SGLT1/SLC5A1, Figure 39C). All three cases positively correlate with sarcopenia. Although both GLUT4 and SGLT1 deficiencies lead to glucose depletion in muscle, the effect of SGLT1 on sarcopenia is much stronger, especially in well- and over-nourished states. This effect is most likely due to the negative impact of SGLT1 deficiency on blood sugar. Conversely, disruption of GLUT4 does not lead to a decrease in blood glucose levels. Thus, anabolic hormones such as insulin remain elevated. In the Sarcopenia Map, energy loss is compensated by other nutrients, such as fatty acid oxidation, comparable to a resting state. During exercise, the effects of reduced glucose uptake in muscle would be more significant.

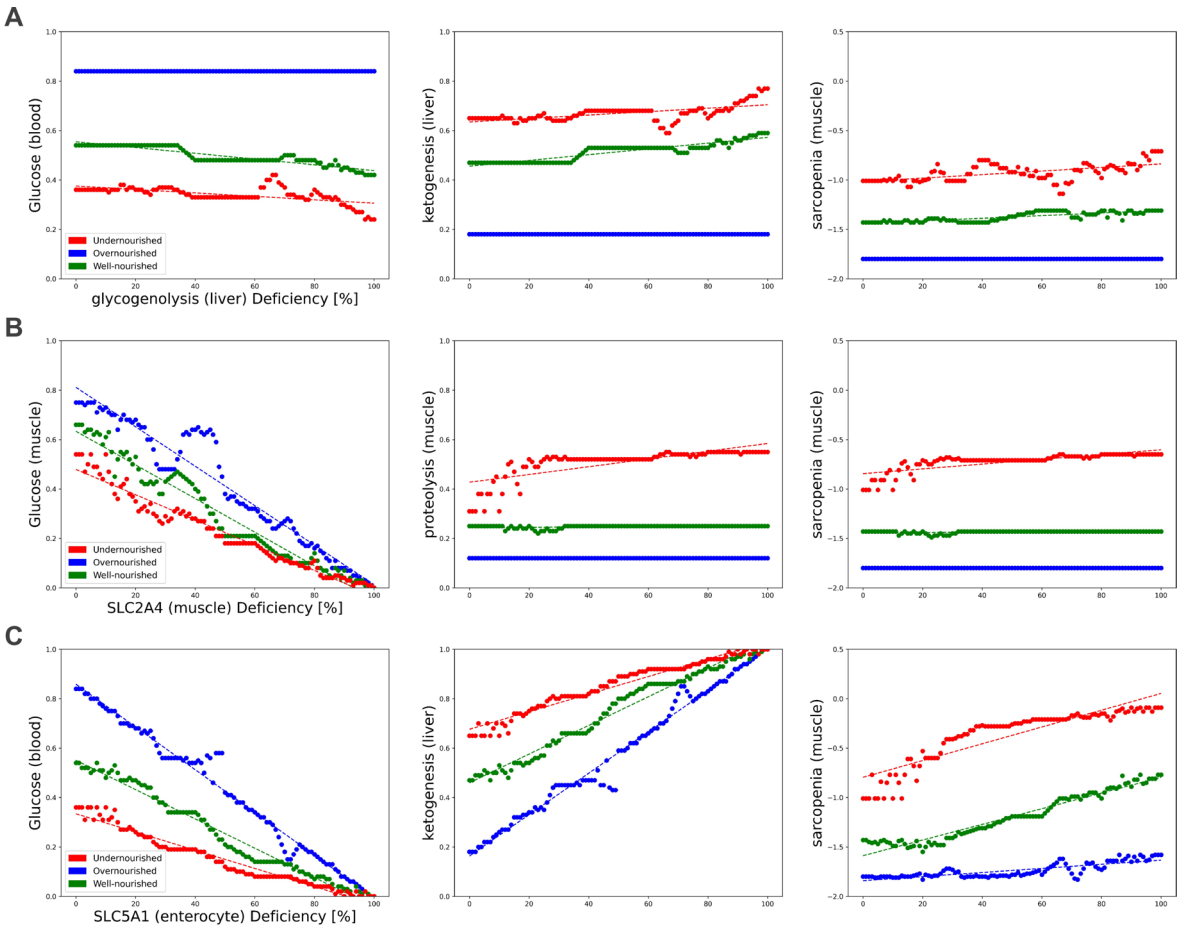

Figure 39: Simulations of molecular perturbations and their observed correlation with other nodes in the map. Each point represents a simulation experiment in which the respective nutritional state was simulated

over 100 steps. During the simulation, the input node was perturbed by setting its state to 0 at a specific frequency (x-axis), and the activity of the observed node (y-axis) was measured. (A) Deficient glycogenolysis in the liver. (B) Deficient glucose uptake in the muscle through SLC2A4 (GLUT4) (C) Deficient glucose absorption in the intestine through SLC5A1 (SGLT1) without sucrose/fructose supplementation.

Next, I investigated the correlations between the activities of LC and ID on the muscle phenotypes 'anabolism', 'catabolism', and 'sarcopenia' dependent on the nutrition state (Figure 40). Although both diseases are present as phenotype nodes in the map, the perturbations are not made to these but to their targets. Given a disease phenotype v, all nodes $u \in V$ with $(u, v) \in E$ are perturbed as $s_p(u, t) = \tau_e(u, v)$ at all steps t where $s_p(v, t) = 1$. This ensures that the disease effects are always prioritized, i.e., even if a target node positively affected by the disease node has another negative input. Both diseases show a strong positive correlation with catabolism (blue) and a negative correlation with anabolism (red). Thus, both disease states also correlate positively with sarcopenia. No major differences are observed between the nutrition states. However, the contribution of both diseases to anabolism appears to be lower in the malnourished state than in the other states. Presumably, this is due to the generally lower activity of anabolism in the undernourished state. The correlation in the overnourished state tends to be constant, whereas the correlations in the nourished and undernourished states are more divergent. In these undernourished states, a greater increase in catabolic and sarcopenic activity is observed even at low LC activities (<0.2). Conversely, the sarcopenia phenotype in ID shows an almost plateau-like behavior at lower disease activities (<0.5), especially in malnourished states, and only then starts to increase. This behavior is expected because at a low frequency of food intake, the baseline activity of sarcopenia is increased, and the effects of ID, which is mainly related to food absorption, are minimal.

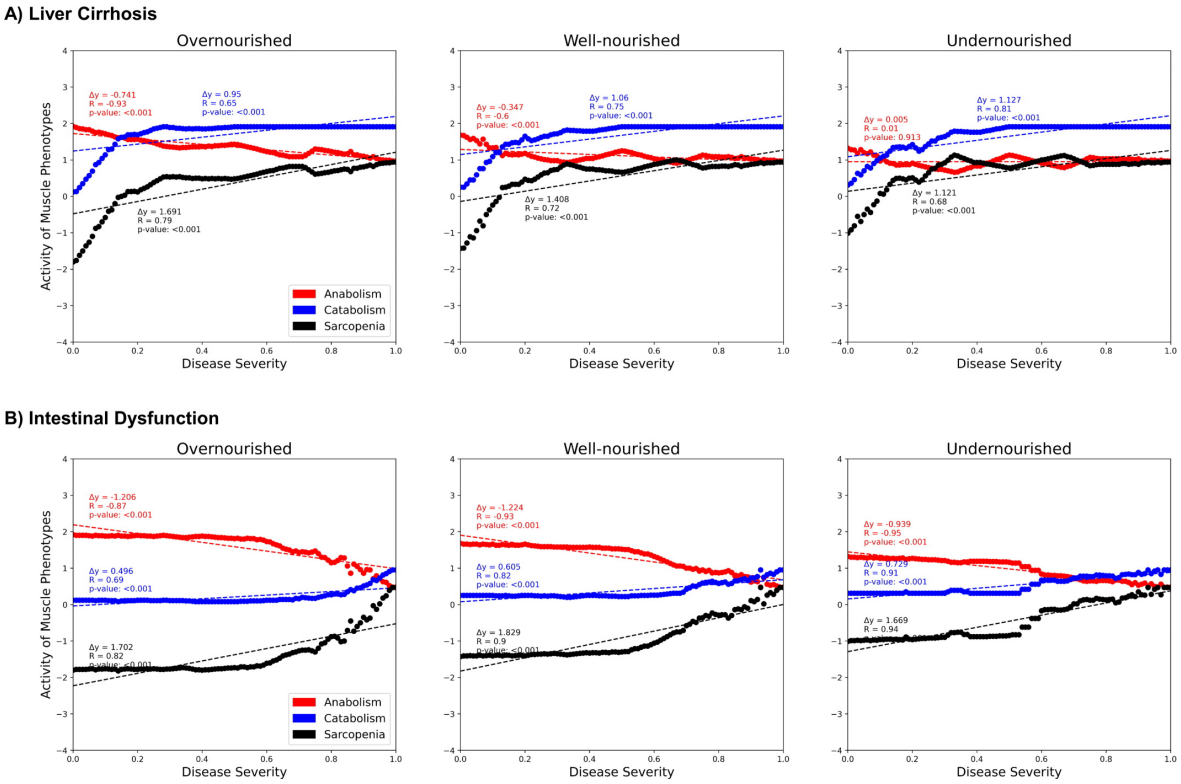

Figure 40: Predicted activities of three muscle phenotypes in response to increasing severity of liver cirrhosis (LC, A) and intestinal dysfunction (ID, B) in three different nutrition states. Each point represents a simulation in which signal transduction is iterated over 100 consecutive steps starting from an initial state. During these steps, the state of LC or ID is set to active with a defined frequency representing their severity.

6.6 Implementation into Disease Maps

The first part of the developed tool provides network topology functions to investigate paths between user-specified nodes to explore their underlying molecular signal transduction. Users can select a source node (“From”) and a target node (“To”) whose paths are to be identified in the KG (Figure 41A). In addition, a further node can be specified to filter paths that lead "Through" it. The output is presented as a table that shows all identified paths, their length, the total impact on the target, and all individual steps within the path (Figure 41B). A PubMed identifier references each interaction, and clicking on the icon takes the user to the location on the submaps. In addition, a bar chart lists the percentage of these paths in which each node occurs, separated into positive and negative paths (Figure 41C). Because of the limitations of topological models (see Introduction), assumptions about functional relationships and mechanisms should not be inferred from the distribution of positive and negative interaction paths alone. Nevertheless, they provide an intuitive overview of the design of molecular pathways and the flow of information.

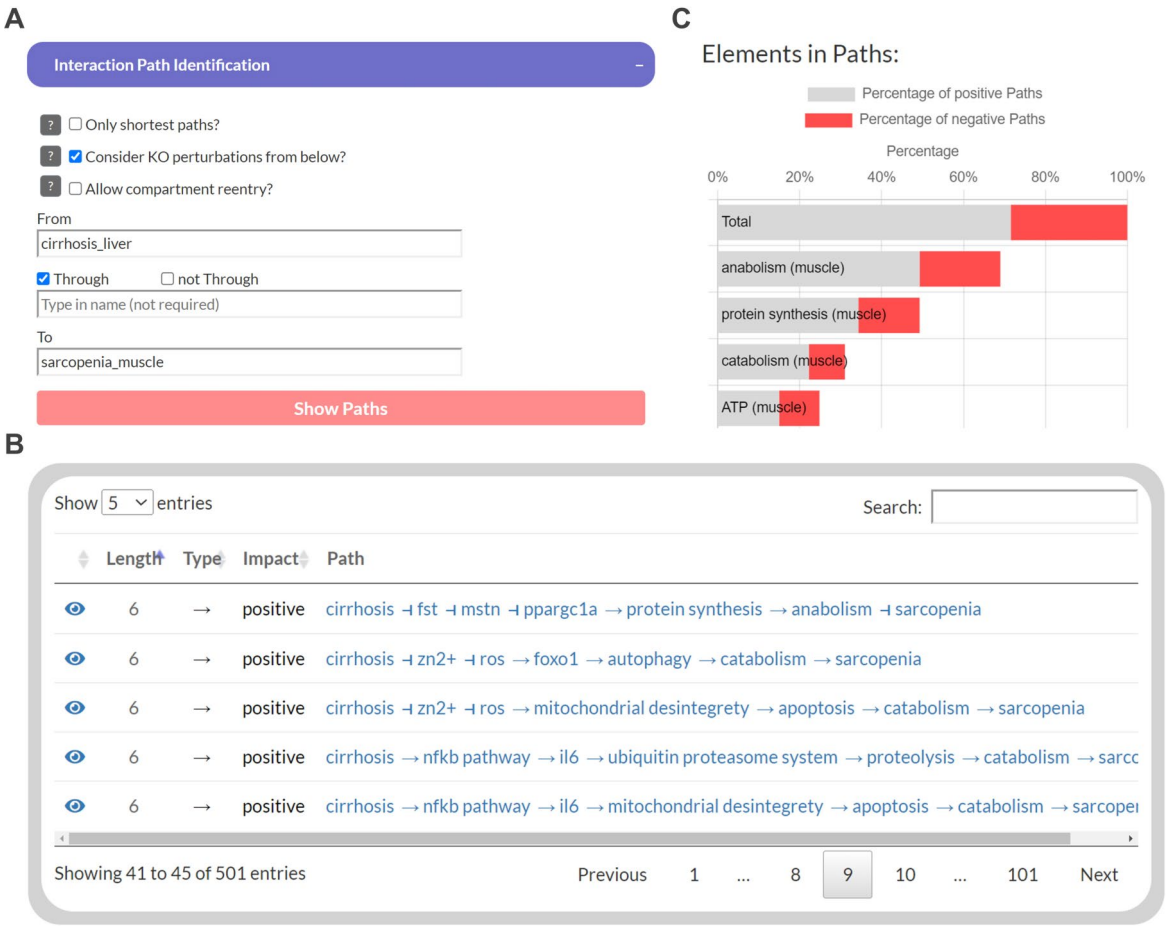

Figure 41: The user interface to identify interaction paths between selected nodes in the Sarcopenia Map. For selected nodes (A), their interaction pathways are listed in a table. (B) Additionally, nodes along the paths are ranked by their percentages of appearance separated by the type of interaction (C).

The second part of the tool enables Boolean simulations on the Sarcopenia Map via a simple UI and colored map overlays (Figure 42A-B). One of its functions is correlation analysis, which provides insights into the mechanistic relationship between nodes. For a selected node (source) and nutrition states, multiple simulations are iteratively and automatically performed with increasing activity or deficiency of the source node (Figure 42C). The correlations of the source and all target nodes, represented by the Pearson correlation coefficient r from Eq. (**6.7**), are then summarized in a table. Scatter plots of the activities of the two nodes provide further information by showing the detailed correlation course at different nutrition states (Figure 42D). For every target node, the table also ranks other nodes in the KG according to the similarities of their activity distributions toward the source and target. Nodes that correlate with both could potentially be responsible for transmitting the signals. In addition, based on the type of correlation (positive or negative), users can investigate the role of the transmitting node, i.e., whether inhibition/activation of an inhibitory/activating signal has occurred or vice versa.

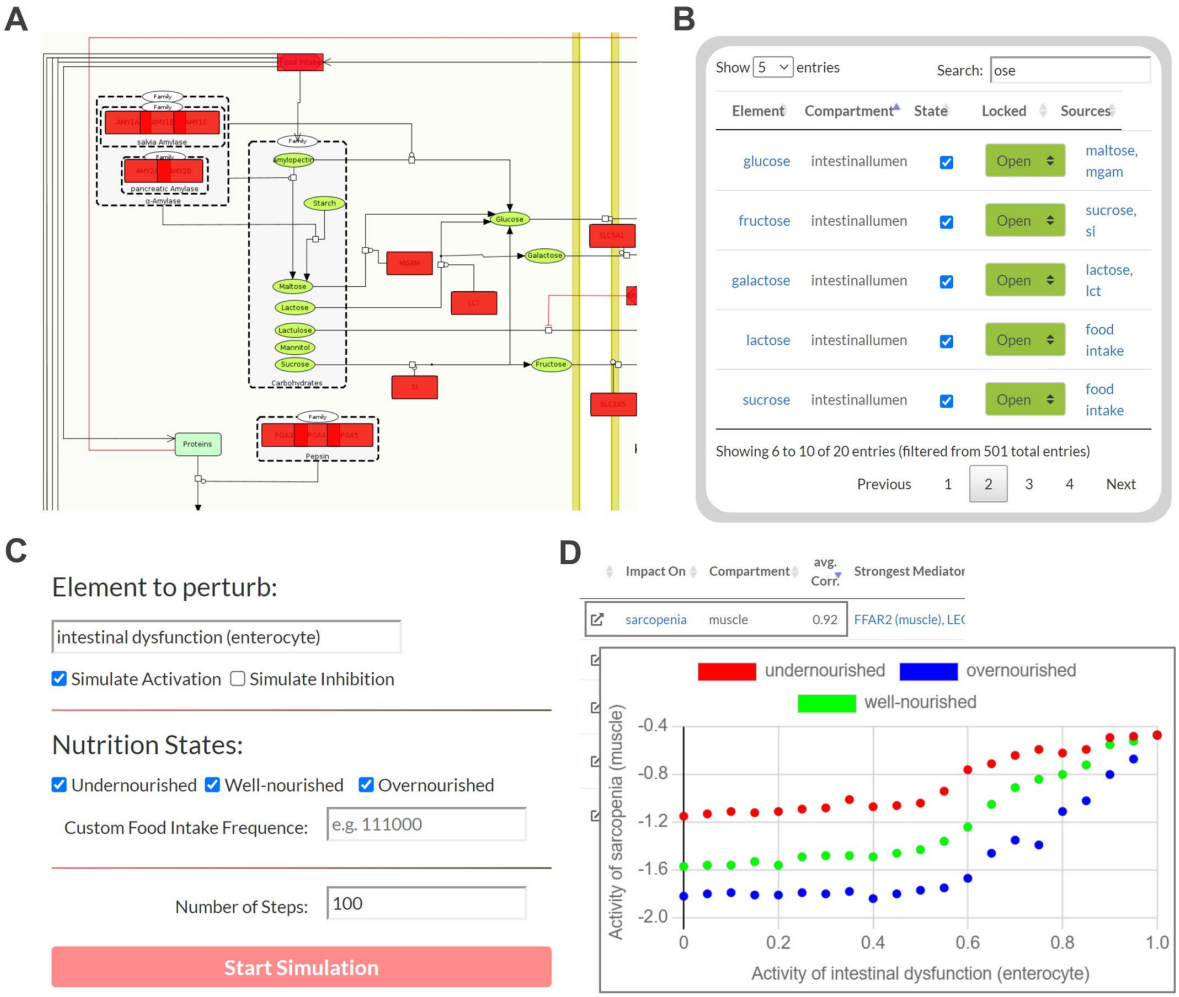

Figure 42: User interface to perform Boolean simulations using the Sarcopenia Map. (A) The active (red) or perturbed (gray) nodes in the network are highlighted for each step. (B) An interactive table provides an overview of all nodes in the KG and allows perturbations by activating or inhibiting their state. (C) Automated perturbation experiments allow the simulation of an increased activation or inhibition of a selected node. (D) The correlation of the activities of the other nodes in response to the perturbation is then presented in a table and diagrams.

6.7 Summary

The Sarcopenia Map employs KG approaches to investigate the molecular signaling linking gastrointestinal diseases and sarcopenia. Understanding which mechanisms are involved can often be very intuitive for the researcher. While there may be one main pathway from one node to another, other mechanisms can also involve non-obvious or not yet directly associated links. With the Sarcopenia Map, I aimed to bring the complex molecular interaction pathways in sarcopenia into a comprehensive and standardized format. It is a knowledge base that (i) gathers molecular information annotated with database references, (ii) intuitively visualizes causalities of molecular mechanisms, and (iii) provides tools for *in silico* simulations. By topologically evaluating the highly interconnected KG, users can

utilize the tools I developed to identify interaction paths between molecules and processes of interest. Using Boolean simulations, the tool allows observing how changes in molecular activities propagate through the system and affect processes on a systemic level, i.e., in other compartments.

However, it should be noted that Boolean models are divided into successive steps of discrete values and cannot analyze continuous changes or molecular quantities. It is, therefore, not well suited to deriving information from a single simulation of signal transmission. However, by carrying out numerous simulations under varying conditions, correlations can be identified across the entire KG, and insights into the underlying mechanisms can be obtained. Another limitation is the abstraction of pathways of multiple reaction steps into a single node. This simplification allows a structured visualization and makes the model more robust towards feedback signaling. Also, retaining all reactions would distort the temporal perception of signal transduction. In a synchronously updated Boolean model, the time scales of all biological events are considered equally. Thus, more steps would have been required for pathways with more intermediate reactions. In reality, however, most reactions occur simultaneously due to the large quantities of molecules involved. For example, in the muscle submap, only the ubiquitin-proteasome system is included as a junction of catabolic signals. Thus, in developing this model, I aimed to strike a balance with the abstraction between the feasibility and informativeness of the complex KG.

By successfully reproducing existing knowledge of the carbohydrate system in (patho-)physiological conditions, I showed that the model can simulate such molecular processes. While my work focused on the effects of gastrointestinal diseases, such as LC or ID, in sarcopenia, the map provides a comprehensive knowledge base linking nutrition and muscle metabolism that can also be useful for other research areas. The hierarchical format of the map and the standardized SBML representation of molecular interactions facilitate the extension to other related diseases or integration of new information, such as malnutrition in relation to other tissues, in the future.

Chapter 7

Multi-compartmental Modeling of Spatial Disease Mechanisms

7.1 Steatotic Liver Disease

Metabolic dysfunction-associated steatosis (MASLD), formerly known as non-alcoholic fatty liver disease (NAFLD) [338], exhibits a complex pathogenesis that requires an in-depth understanding of diagnostic and prognostic molecular patterns to develop effective clinical treatment strategies [339]. MASLD is classified as liver steatosis under cardiometabolic criteria without any other identified factors. Advanced fibrosis and the initiation of inflammatory processes can lead to the progression into metabolic dysfunction-associated steatohepatitis (MASH), cirrhosis, and hepatocellular carcinoma [338], [340]. The liver is a complex organ that performs many metabolic processes and adjusts the blood metabolome to current needs. Consequently, SLDs have a high systemic relevance. In this context, many of the currently investigated drugs do not target the liver directly but hormonal or metabolic processes in general. GLP-1 agonists, such as semaglutide, target the GLP-1 receptor, directly affecting many tissues such as the pancreas, adipose tissue, intestine, muscle, heart, and brain [341]. The recently FDA-approved drug resmetirom uniquely targets the thyroid receptor beta to regulate metabolic dysfunction [342].

However, the wide variety of molecular and cellular processes in MASLD and their complexity make it difficult to analyze experimental and clinical data due to the non-intuitive causalities at the systemic level [343], [344]. Many computational models have been developed aiming to simulate specific mechanisms on a specific spatiotemporal scale [142], [345], [346], [347], [348], [349]. These models focus on specific disease aspects, e.g., modeling signaling or metabolic pathways in hepatocytes [345], [350]. Thus, they are limited to small-scale processes and cannot describe MASLD as a systemic disease. Multi-scale aspects of complex diseases like MASLD often cannot be adequately represented by such approaches limited to small-scale pathways. A large-scale approach that connects

processes across multiple tissues and organs is needed to capture the dynamic complexity of the disease caused by multiple factors throughout time and space.

Given our previous work on larger-scale KG models, i.e., Disease Maps of inflammation (Chapter 2) and gastrointestinal diseases (Chapter 6), in collaboration with *HEEL GmbH*, we developed a novel large-scale Disease Map on systemic processes in MASLD. The project can be divided into two parts described in the following two sections. First, Section 7.2 details the creation of a novel Disease Map on MASLD/MASH-relevant molecular processes underlying the phenotypic changes of various tissues and organs. It showcases the exploration of clinical information in the map and the visualization of experimental data and their analysis using the 2DEA approach presented in Chapter 3. Section 7.2.5 explores modeling strategies to simulate the progression of MASLD in spatial and temporal contexts. It describes the current state of development of a multi-compartmental Boolean model and its implementation into a MINERVA Disease Map.

7.2 A Systemic Disease Map Approach to MASLD

We connected with renowned clinicians from gastroenterology departments of university hospitals in Germany with extensive expertise in MASLD, namely Ali Canbay (Bochum, Germany), Andreas Geier (Würzburg, Germany), and Jörn Schattenberg (Mainz, Germany). The clinical experts support the project by validating information curated in the MASLD Map, advising the model development from a biomedical viewpoint, and providing feedback on the map's usability and the plugin from a clinical user perspective. Similar to the AIR, we constructed a Disease Map and compiled experimentally validated molecular interactions in standardized submaps. We published the MASLD Map as an interactive web-based platform on MINERVA (https://www.sbi.uni-rostock.de/MASLD). The MASLD Map aims to provide a detailed computational representation of MASLD pathogenesis and progression, offering users insights into the dynamic state changes of nodes over time and based on input parameters.

7.2.1 Designing and Curating the MASLD Map

We present the MASLD Map as a web-based platform providing a comprehensive overview of disease mechanisms in MASLD and MASH. We published the resource on the MINERVA platform, which provides many features for exploring its content and accessing

the data. We designed the MASLD Map in a multi-level layout, allowing the users to explore knowledge at different resolutions, ranging from higher-level systemic processes to their underlying molecular pathways. Similar to the AIR [187], we summarized the most relevant processes involved in MALSD in an overview image summarizing the most important molecules, higher-level biological processes, cell types, and tissues (Figure 44A). We considered extrahepatic tissues such as intestinal digestive processes, the endo- and exocrine pancreas, and adipose tissue. At the cellular level, the MASLD Map includes various cell types of the hepatic microenvironment, innate and adaptive immune cells, and gastrointestinal cell types, among others. From the overview image, manually curated molecular mechanisms that modulate the higher level are accessible as standardized SBML submaps created in CellDesigner [64], [351]. Currently, the MASLD Map comprises 35 submaps summarized in Table 7, integrating a total of 4288 interactions between 3827 tissue-specific nodes, of which 2140 relate to unique genes, proteins, metabolites, and phenotypes.

Table 7: Overview of the submaps currently included in the MASLD Map.

Submap	Nodes	Unique Nodes	Edges
Adipose Tissue	49	10	37
B Cell	98	34	44
Bile Acids	113	75	87
Coordinated Lysosomal Expression and Regulation (CLEAR) network	440	161	393
Chaperone mediated autophagy	44	23	27
Cholesterol Synthesis and Effects	73	42	59
Digestion and Absorption	455	136	1271
Endocrine Pancreas	103	34	120
GLP-1	111	66	82
Hepatocyte Apoptosis	131	66	123
Hepatocyte	184	82	145
IL17	96	33	65
Immune Cascade	40	16	35
Kupffer Cell	35	10	32
Lipid Droplets	40	23	29
Lysosomal biogenesis	346	144	261
M1 Macrophage	111	39	89
M2 Macrophage	75	18	88
Macroautophagy	118	46	59
Macrophages phagocytosis	211	112	174
Metabolic Pathways	368	288	173
Microautophagy	73	43	15
Natural killer (NK) cell chemotaxis	59	22	55
Natural killer cell	106	55	55
Neutrophil chemotaxis	52	17	63
Stellate Cells	127	49	133
T cell activation	106	31	44
T follicular helper (Tfh) cell	24	3	19

T Helper 1 cell	40	6	31
T Helper 17 cell	70	0	51
T Helper 2 cell	59	10	45
T Helper 22 cell	25	5	17
T Helper 9 cell	46	11	27
Regulatory T (Treg) cell	68	6	49
Vitamins and Trace Elements	56	16	43

Using the OmniPath resource (https://omnipathdb.org/) [352], we created a KG on interactions between transcription factors and their gene targets, including protein-coding genes and miRNAs. OmniPath provides a computational portal with implementations for Python and R that combines data from various public resources in comprehensive KG formats. We have compiled a dataset on the interactions of transcription factors with their gene targets from the Omnipath datasets "dorothea", "tf_target" and "collectri" [353], [354]. Additional data from OmniPath include the dataset "mirnatarget", which contains the interactions of miRNAs with gene products, and the dataset "tf_mirna", which describes the interactions of TFs with miRNA genes. We also integrated the latest dataset on lncRNA-protein interactions from lncTarD (https://lnctard.bio-database.com/). The information from the submaps and the regulatory interactions was merged into a large-scale KG of 130,920 edges between 19,180 nodes. Similar to the AIR, the complete KG is not explorable through an SBML diagram but was made accessible through the plugin tool suite we developed. The plugin files were adapted for the MASLD Map and are available at https://github.com/sbi-rostock/AIR/tree/master/MASLD.

7.2.2 Highlighting Clinical Knowledge

When processing the KG using the computational framework described in Section 2.5, the nodes are uniquely identified by their compartment in addition to their name and type. This way, I created a modularized version of the MASLD Map in which entities in different compartments represent separate nodes. The number of directed connections between two compartments is defined as the number of elements without a compartment for which there are two interactions: (i) one with the element as the target where the source element is from one compartment, and another interaction (ii) with the element as the source where the target is from the other compartment. Figure 43A shows the directed interactions between compartments in the MASLD Map weighted by the number of mediators and nutrients secreted by or acting on the tissue. It shows that a large part of the interactions happens between the intestine, adipose tissue, and liver, representing the vast amount of nutrients

exchanged between the tissues. Other interactions, especially between different immune cell types, are more fine-grained. MINERVA offers a comprehensive API that enables seamless integration of the MASLD Map into analysis workflows [205]. For example, the MINERVA platform can also be used to research functional processes in a clinical context by analyzing processes affected by clinical variants or drug targets. I mapped information on disease-associated gene variants from the DisGeNET database to the MASLD Map. In Figure 43B, genes are ranked by the DisGeNET gene-disease association score for MASLD, and the cumulative percentage of genes included in the map, starting with the top rank, is indicated. The manually curated KG of the MASLD Map contains all the top 16 genes, 50% of the top 100, and around 30% of all genes. However, many of the genetic variants are identified by GWASs, and the function of their gene products may not yet be known, preventing their incorporation. The complete KG of the MASLD Map covers >90% of all disease-related genes.

The MINERVA platform further provides UI and API capabilities to identify protein-targeting drugs, chemicals, or miRNAs. To illustrate these functionalities, I evaluated the drug targets of 619 drugs on the MASLD map that are listed in the MASLDkb database as being under investigation for the treatment of MASLD [355]. Of these drugs, 267 had at least one identified target on the map (3.6± 9.9 targets on average). Figure 43C shows in which parts of the map the target proteins of each drug are involved. Due to the easy integration of the map with other resources, the context-specific KG of the MASLD map provides an excellent platform for systems pharmacology approaches. The map enables the investigation of the effects of drugs, genetic variants, or a combination of both on molecular and cellular processes in MASLD and MASH.

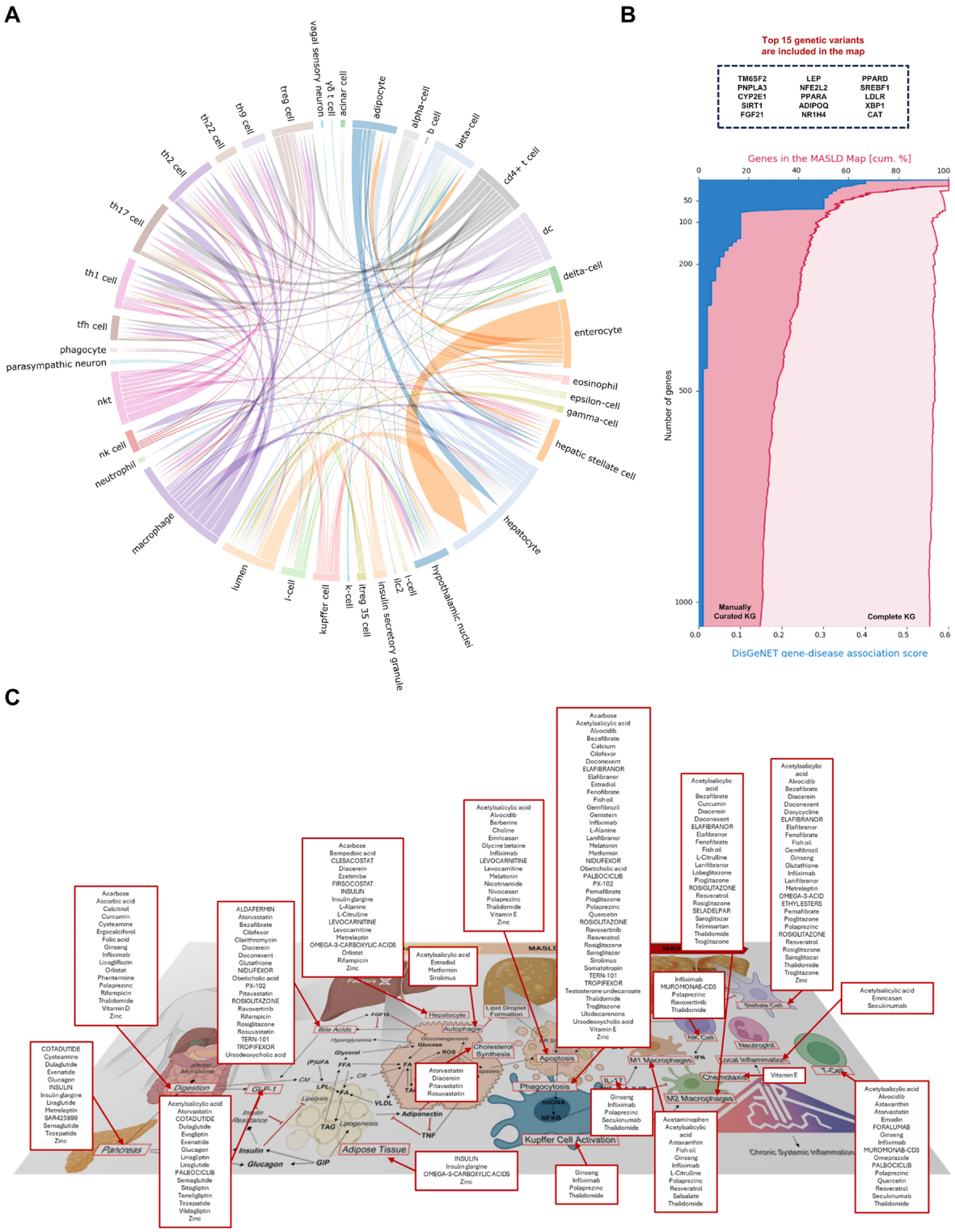

Figure 43: (A) Chord diagram highlighting the interactions between compartments in the MASLD Map. The width of an arrow reflects the number of unique hormones and metabolites exchanged between two compartments, while its direction is the signaling flow or transport of substances. (B) Genes associated with MASLD from the DisGeNET databases were sorted by their gene-disease association score, and the cumulative percentage of their inclusion in the MASLD Map was calculated. (C) Visualization of which processes in the MASLD Map are targeted by drugs under investigation for the treatment of MASLD based on the inclusion of their potential protein targets.

7.2.3 Data Visualization and Analysis

I employed a dataset originally published by Hoang *et al*. in 2019 (GEO accession number GSE130970) as a case study showcasing the functionalities of the MASLD Map [356]. The data includes bulk-tissue RNA-Seq profiles from hepatic biopsy tissues of 78 MASLD patients at clinical disease stages, clinically defined by NAFLD fibrosis score (NAS), and scores of fibrosis, inflammation, and ballooning. I identified the differential gene expression between subsequent disease stages and for each stage compared to the respective control (stage 0) using the GEO2R tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/). Additionally, I downloaded the statistical associations and differential expression of genes associated with NAS and fibrosis progression as identified by the original authors through ordinal regression. The individual files were combined into tabular files, with the first column being the gene name and, subsequently, the FC value and adj. p-value for each sample. Results were uploaded to MINERVA as publicly available overlays highlighting the FC values of DEGs. We used MINERVA's functionality to link submaps with SBML nodes and create a nested design for related processes. In these cases, not only are overlays on molecules displayed in the KG, but the data values of all nodes in a submap are projected onto the nodes linked to it. This way, large data sets with multiple mapped probes can be compressed into an intuitive representation of the affected processes on the overview image (Figure 44A).

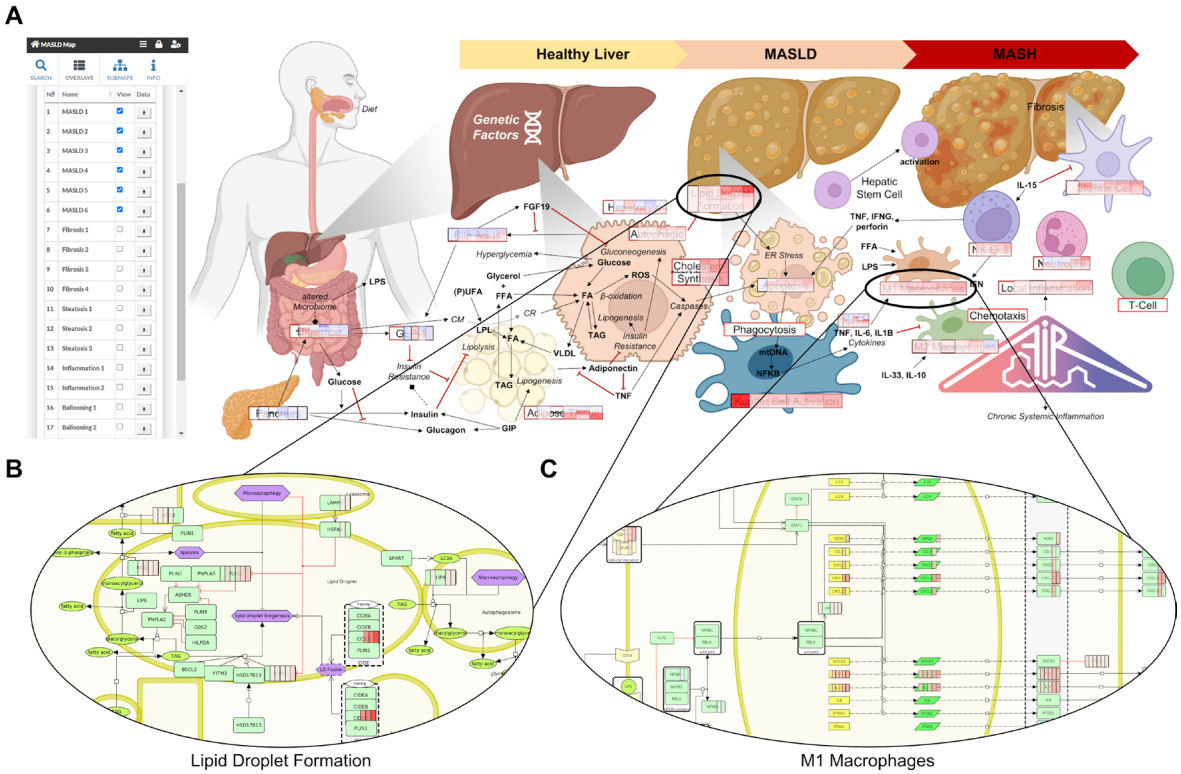

Figure 44: Visualization of the molecular data on the MASLD Map. The differential gene expression of different MASLD stages from Hoang *et al*. is available as public overlays on the map (A). Selecting an overlay automatically highlights the representations of the corresponding gene products on the map (B-C). The overall expression profile of genes from a single submap is displayed in compressed form on the overview image, facilitating the interpretation of the data.

7.2.4 Phenotypic Activity at MASLD Stages

The plugins of the MASLD Map also include the 2DEA, which I employed to analyze the RNA-Seq by Hoang *et al*. that was uploaded to the map [356]. Figure 45A-D shows the results for phenotypes in samples comparing gene expression at different stages of fibrosis, NAS, inflammation, and steatosis compared to their respective controls. The most overlap, especially in significant predictions, is visible for apoptosis, neutrophils, macrophages, B cells, and T-cell receptor (TCR) response, indicating that the activation of these processes might be relevant across all phenotypes and stages in MASLD and MASH. A functional difference we see in the processes of fibrosis and stellate cell activations, which is downregulated specifically at steatosis grade 3 compared to no steatosis. These processes are highly upregulated in all other disease stages, especially in high-grade fibrosis. Visualizing the differential gene expression on the stellate cell submap showed that in fibrosis, as expected, pro-fibrotic genes such as collagens, CCL2, CCN2, and compensatory MMPs are upregulated. In contrast, late-stage steatosis shows downregulation of ACTA2 and CCN2, unaffected collagen expression, and higher MMP9 levels. Other indications are

reduced leptin receptor expression in late-stage steatosis and retinoid receptor in fibrosis. Available evidence suggests that CCN2 causes a non-fibrotic phenotype at high levels of steatosis. However, while inhibition of CCN2 has been associated with reduced fibrosis, current research does not suggest that CCN2 is downregulated in steatosis. Evidence in mouse models indicates that CCN2 expression is not affected in response to high-fat diets [357], [358], [359].

In addition to the differential analysis between the individual stages, Hoang *et al.* identified genes whose total expression correlated with NAS and fibrosis stages by ordinal regression and calculated their total change across disease scores. I performed the downstream enrichment analysis on this data and compared how individual genes contributed to the 2DEA for both NAS and fibrosis (Figure 45E-F). The top five genes with positive fold change values associated with increasing NAS but not fibrosis levels were FGF21, GAPDH, PFKFB4, PGK1, and GPI, while the genes predicting fibrosis stage but not NAS were STAT1, KLF2, CCL19, SLC51B, and RPIA. Four genes from NAS data are enzymes from carbohydrate metabolism, indicating a strong relevance of these processes in non-fibrotic MASLD progression.

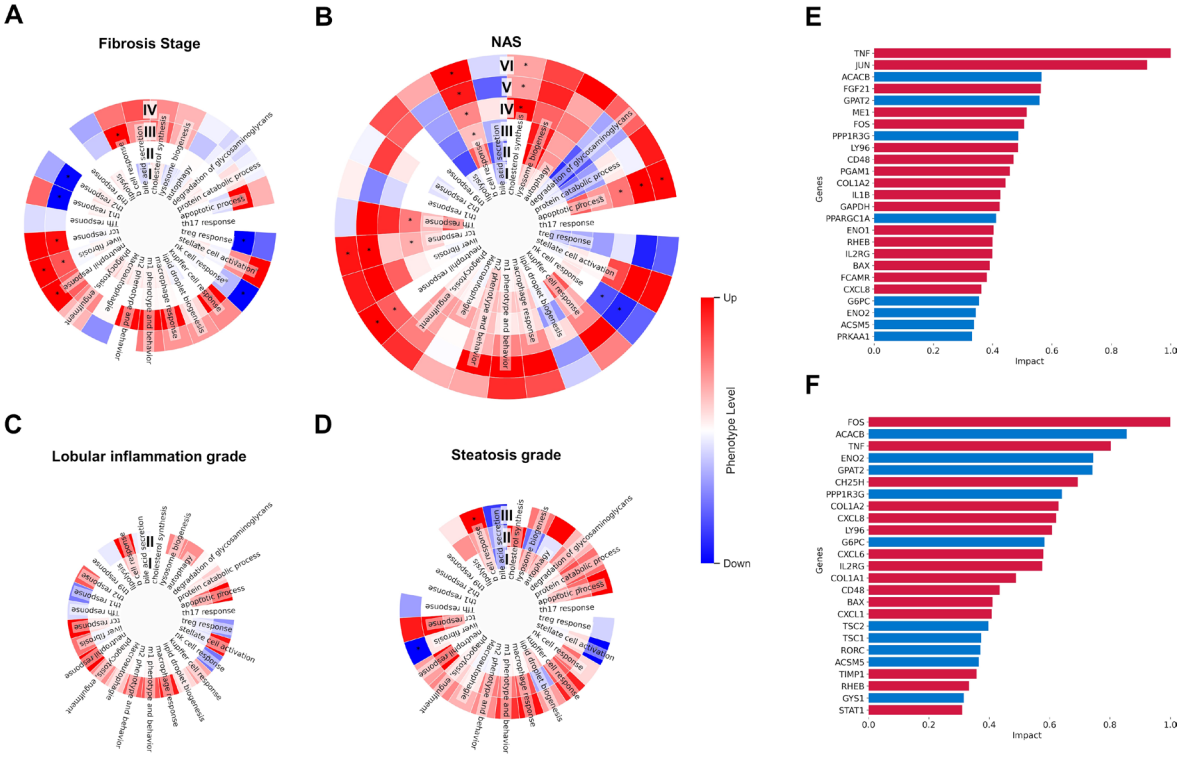

Figure 45: Network-based enrichment analysis of bulk RNA-Seq data from liver biopsies in different MASLD stages. (A-D) Predicted levels of biological processes at various stages of fibrosis (A), NAFLD activity score (B), inflammation (C), and steatosis (D) using network-weighted gene set enrichment. Circles represent the stage from lowest (inner) to highest (outer) compared to the stage zero control. * adj. p-value < 0.05 (E-F) Impact of genes on the enrichment analysis of differential gene expression identified from ordinal regression along fibrosis stages and NAS values by Hoan *et al.* The genes are colored by the fold change in the original data (blue – downregulation, red – upregulation).

7.2.5 Summary

With the MASLD Map, we aim to tackle the challenges of describing and simulating disease processes in MASLD caused by its spatiotemporal heterogeneity and dynamic nature. The map contains manually curated, evidence-based knowledge from research and connects it to other resources that link clinical questions to understanding disease mechanisms from basic research. Multiple anchor points for data processing, such as API capabilities of MINERVA, with the main functions: (i) interactive exploration of available knowledge, (ii) dynamic data visualization, (iii) identification of targets for drugs, chemicals, miRNA, (iv) investigation of genome and protein structures including genetic variants, and (v) functional analysis of large-scale datasets. Therefore, we developed the MASLD Map to provide the research community with a publicly accessible platform that accelerates clinical and experimental research of systemic disease processes in MASLD and MASH through improved data accessibility and analysis tools.

7.3 Multi-compartmental Modeling of Spatial Disease Mechanisms

Research has shown that liver function, and therefore the progression of MASLD, is influenced by spatial heterogeneity in the liver. This heterogeneity can be divided into a "macroscale," i.e., different macroscopic regions have higher or lower metabolic rates caused by the shape of the liver and blood vessel coverage, resulting in heterogeneous availability of metabolites and oxygen (Figure 2A) [360], [361], [362], [363]. Secondly, there is the general microscale heterogeneity caused by high variability in cellular composition and a genetic and transcriptional mosaic of liver tissue. Therefore, considering spatial aspects in computer models is of central importance for understanding disease progression. Due to the heterogeneity of the liver, approaches that focus on small-scale models of a few signaling pathways may not adequately represent the spatial aspects of disease progression [364]. The wide variety of molecular and cellular processes in MASLD and its multifaceted nature complicates data interpretation [343], [344].

Boolean models have proven helpful in simulating multi-level disease mechanisms. With the Sarcopenia Map described in Chapter 6, I showed that Boolean models can model systemic and gastrointestinal processes and even simulate quantities of metabolites by integrating multivalued Boolean logic. A similar approach can be applied to the MASLD

Map. However, the Sarcopenia Map was developed using a one-compartment approach, where each cell/tissue is represented by only one SBML compartment. Consequently, each molecule per cell/tissue is represented by only one node and can assume only one state. This allows the molecular mechanisms to be analyzed per se, but the spatial interactions and their effects on disease phenotypes at the tissue level cannot be adequately described. Therefore, I integrated a Boolean model similar to the Sarcopenia Map into a multi-compartmental model of the liver to describe the pathogenesis of MASLD.

The model is divided into two submodels: a single-compartmental Boolean model of the extrahepatic tissues, such as the intestine with all digestive processes, pancreas, and adipose tissue, and a multi-compartmental, ABM-like model of the hepatic tissues (Figure 46B). The individual compartments of the model are, in the following, referred to as **agents** to avoid confusion with compartments in the biological sense, even if they do not exactly correspond to the definitions of agents in ABMs. In the model, each agent represents separate Boolean models that communicate via the signals from metabolic and hormonal processes (further detailed in Section 7.3.2).

The individual agents differ slightly in the activity of the metabolic processes and the amount of metabolite signaling received from the extrahepatic model. The stochastic distribution of nutrient supply and metabolic activities between compartments allows the simulation of variations in liver metabolism under different conditions. KG models of digestive processes enable the study of disease progression following changes in nutritional status. Using a multi-compartmental model of the liver to simulate macro- and micro-heterogeneity in hepatic processes, I aim to identify processes favoring MASLD progression in specific regions. This approach will allow to simulate molecular processes in heterogeneous environments and to investigate spatial aspects of MASLD progression.

Integrating extrahepatic compartments, particularly digestive processes, allows the study of MASLD pathogenesis in relation to diet and, later, the integration of systemic clinical data (see Section 7.3.4). In its current state, the agents' parameters are randomized. In the future, I plan to adapt the layout and parameters of the multi-compartmental model to the clinical data. Given the quasi-hierarchical structure of the model, the model is introduced by defining the quantity of nutrition (frequency of the node "food intake") and the quality (state of the individual nutrients), starting from the food intake. The state of the nodes in the model is evaluated step by step for a certain number of steps. The response can be measured as the frequency of active states in specific nodes in a single model or aggregated for multiple agents (Figure 46C).

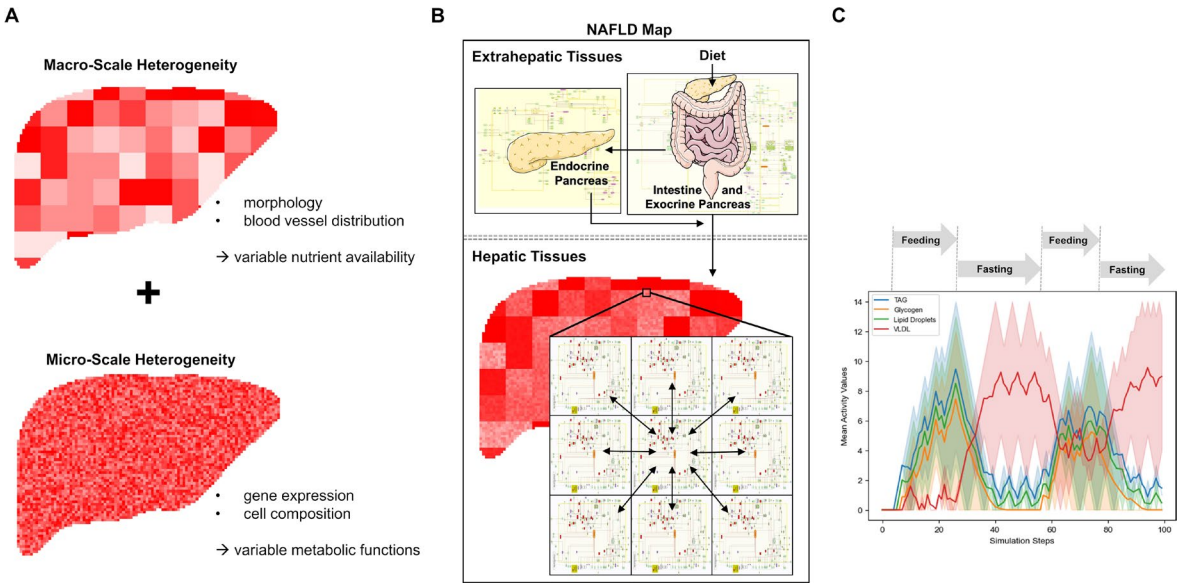

Figure 46: The multi-compartmental Boolean modeling in the MASLD Map. (A-B) The map is divided into two submodels: an extrahepatic Boolean model and a compartmental/agent-based model for the liver. The liver compartments are distributed with different parameters to represent the heterogeneity of the liver. (C) The Boolean model enables mechanistic simulations of molecular signaling pathways. Integrating extrahepatic processes, such as food intake, enables the simulation of the hepatic metabolic response to nutrition.

7.3.1 A Multivalued Probabilistic Boolean Model

To dynamically simulate spatial and temporal mechanisms in a qualitative model, I adapted the Boolean definitions from Section 6.3 with parameters that reflect complex biological behavior. First, I incorporated a delay property t_d into some nodes that causes their state to be interpreted as their state t_d steps earlier, redefining $s'(v,t)$ as:

$$s'(v,t) = \begin{cases} \text{true} & if\ s_p(v,t) = 1 \\ \text{false} & if\ s_p(v,t) = -1 \\ \text{false} & if\ s_p(v,t) = 0 \wedge s(v, t - t_d(v)) = 0 \\ \text{true} & otherwise \end{cases} \tag{7.1}$$

Next, I implement a stochastic modeling approach by evaluating the states of source nodes of an edge not deterministically but probabilistically based on the percentage of their maximum states. When evaluating an edge function f, a Boolean variable p_e is added to the query of the pseudo-state of source and modifier nodes, which indicates whether it is interpreted probabilistically. The value of p_e is **`false`** by default but is set to **`true`** if the edge note in SBML has the "probabilistic" tag. The pseudo-state s' is now defined as following with $r \in \mathbb{R}, r \in [0,1]$ being a randomly generated value:

$$s'(v,t,p_e) = \begin{cases} true & \text{if } s_p(v,t) = 1 \\ false & \text{if } s_p(v,t) = -1 \\ false & \text{if } s_p(v,t) = 0 \wedge s(v,t) = 0 \\ false & \text{if } s_p(v,t) = 0 \wedge p_e \wedge r > \dfrac{s(v,t)}{s_{max}(v)} \\ true & \text{otherwise} \end{cases} \tag{7.2}$$

Introducing probabilistic logic allows for dynamic representations of discrete node states. However, there is still the issue that a node is continuously inhibited by a negative input if one of the inputs' states is $s(v,t) = s_{max}(v)$. From a biological perspective, negative inputs should sometimes only partially affect their target's state. Such behavior can be modeled by adding a second limit to the node state, which is lower than s_{max}. This way $\frac{s(v,t)}{s_{max}(v)}$ can be set so that it is always below 1 and, if defined for nodes with inhibiting outgoing edges, prevent 100% inhibition by them. I define this upper state limit of a node v as $s_{lim,up}(v)$, with $s_{lim,up}(v) = s_{max}(v)$ by default, which replaces the upper limit when updating the node state. Conversely, nodes should be continuously active in other cases but increase their activity with input. Thus, I also define a lower state limit of a node v as $s_{lim,low}(v)$, with $s_{lim,low}(v) = 0$ by default. The Equation (6.2) that updates a node's state changes to:

$$s(v,t+1) = \min\left(s_{lim,up}(v), \max\left(s_{lim,low}(v), s(v,t) + \begin{cases} 1, & \text{if } f_v(v,t) = true \\ -1, & \text{otherwise} \end{cases}\right)\right) \tag{7.3}$$

Because the state of multivalued nodes decreases by one for every step where there is no input, an automatic decay function is integrated. The decay $C_d(v)$ (1 by default) defines the amount by which the state of a node is reduced when there is no input at a step.

$$s(v,t+1) = \min\left(s_{lim}(v), \max\left(0, s(v,t) + \begin{cases} 1, & \text{if } f_v(v,t) = true \\ -\boldsymbol{C_d(v)}, & \text{otherwise} \end{cases}\right)\right) \tag{7.4}$$

To represent the consumption of molecules through metabolic reactions more accurately, an edge-specific decay was introduced, which, if active, reduces the state of its source nodes by a specific number. Let's assume that $E_{out}(v) = \{e \in E(G) | v \in S_u(e)\}$ are the outgoing edges of a node $v \in V(G)$ each with a decay constant $C_c(e) = 0$ by default. The aggregated decay of node v is defined as the following and integrated into the state function shown in Equation (7.7).

$$\overline{C_c}(v) = \sum_{\substack{e \in E_{out}(v) \\ f_e(e,t,v) = true}} C_c(e) \tag{7.5}$$

In the Sarcopenia Map, the storage of glucose in glycogen and subsequent release of glucose from glycogen is modeled through a regulatory feedback loop, including insulin, glucagon, and the glucose transporter. In the MASLD Map, more stored metabolites are integrated than in the Sarcopenia Map, including fat tissue, vitamins, and other byproducts, such as reactive oxygen species (ROS). A more efficient solution is required to avoid curating a complex regulatory system for each node in the KG to prevent positive feedback loops. I added a "refill" tag to reactions to edges outgoing from storage nodes (inside of compartments) to their released form (outside of compartments). In these cases, $s'(v,t)$ of the released node v is extended with the state of any storage nodes $u \in N_r(v)$ of all storage nodes $N_r(v) \subseteq N_{in}(v)$ releasing v. To prevent feedback loops, the original node v_0 which is requesting $s'(v,t)$ is being parsed through the edge logic $f_e(e,t,v_0)$ for all $e \in E_{out}(v_0)$. Consequently, if the state of the storage node v is being updated from all incoming edges, the state of the released node does not project the state of v because $v = v_0$. In summary, u is projecting the state of its storage nodes to any outgoing edges, except for the storage nodes themselves, described by the following Boolean logic and integrated into the final equation for $s'(v,t)$ in Equation (7.10).

$$f_r(v,t,v_0) = \bigvee_{u \in N_r(v)} \left(v_0 \neq u \wedge s'(u,t)\right) \tag{7.6}$$

Considering the adaptations to the Boolean definitions in Section 1.4.4 and Section 6.3, Equations (7.7) to (7.10) below summarize the final formulations for the Boolean model of the MASLD Map. The new state of a node at a step $t+1$, as discrete values between 0 and s_{max}, is based on a Boolean function f_v of incoming edges and a decay function of outgoing edges.

$$s(v,t+1) = \min\left(s_{lim,up}(v), \max\left(s_{lim,low}(v), s(v,t) + \begin{cases} 1, & \text{if } f_v(v,t) = true \\ -C_d, & \text{otherwise} \end{cases} - \left(\sum_{e \in E_{out}(v)} \begin{cases} C_C(e) & \text{if } f_e(e,t,v) = true \\ 0 & \text{otherwise} \end{cases}\right)\right)\right) \tag{7.7}$$

The Boolean function f_v itself is defined in such a way that at least one of the incoming edges with a positive type is ON, while all negative types are OFF, defined by the Boolean function f_e of the edges.

$$f_v(v,t) = \left(\bigwedge_{e \in E_{in}^-(v)} \neg f_e(e,t,v)\right) \wedge \left((E_{in}^+(v) = \emptyset) \vee \bigvee_{e \in E_{in}^+(v)} f_e(e,t,v)\right) \tag{7.8}$$

The Boolean function f_e itself is defined in such a way that all source nodes of the edge and all modifier nodes of at least one positive modification need to be ON, while all negative modifications are OFF, interpreted through the pseudo-state s' of the nodes.

$$f_e(e,t,v_0) = \left(\bigwedge_{u \in V_s(e)} s'(u,t,v_0) \right) \wedge \left(\bigwedge_{m \in \mathbf{S}_m^-(e)} \neg s'(m,t,v_0) \right) \wedge \left((\mathbf{S}_m^+ = \emptyset) \vee \bigvee_{m \in \mathbf{S}_m^+(e)} s'(m,t,v_0) \right) \tag{7.9}$$

The state of a node as 'seen' by the Boolean logic of edges is interpreted as a Boolean value that is true on a positive state s of the node or false on $s = 0$, and be perturbed, be probabilistic based on a percentage of the max state value or display the state of a connected storage node.

$$s'(v,t,p_e,v_0) = \bigvee_{u \in V_r(v)} \left(v_0 \neq u \wedge s'(u,t)\right) \vee \begin{cases} true & \text{if } s_p(v,t) = 1 \\ false & \text{if } s_p(v,t) = -1 \\ false & \text{if } s_p(v,t) = 0 \wedge s\big(v, t - t_d(v)\big) = 0 \\ false & \text{if } s_p(v,t) = 0 \wedge p_e \wedge r < \dfrac{s\big(v, t - t_d(v)\big)}{s_{max}(v)} \\ true & \text{otherwise} \end{cases} \tag{7.10}$$

7.3.2 Compartmentalization of the Boolean Model

In the next step, the KG and Boolean formulations must be converted into a multi-compartmental model. From a theoretical perspective, every agent in the model is a copy of a graph G defining the multi-compartmental model of size n as a matrix:

$$M_G = \begin{bmatrix} G_{1,1} & G_{1,2} & \cdots & G_{1,n} \\ G_{2,1} & G_{2,2} & \cdots & G_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n,1} & G_{n,2} & \cdots & G_{n,n} \end{bmatrix} \tag{7.11}$$

In all graphs of M_G, a Boolean model is evaluated in parallel with the description in Section 6.3. Given a node $v_{i,j}$ in a graph $G_{i,j}$ of an agent at position (i,j), the set of its equivalents in the neighboring agents is defined as:

$$\mathcal{N}\big(v_{i,j}\big) = \{v_{k,l} \mid k \in \{i-1, i, i+1\}, l \in \{j-1, j, j+1\}, (k,l) \neq (i,j)\} \tag{7.12}$$

Interactions between v and neighboring agents were integrated into the model by perturbing a node $v_{i,j}$ in an agent (i,j) with a probability based on the state of a node u in neighboring agents $\mathcal{N}\big(u_{i,j}\big)$. The aggregation and probabilistic perturbation of a node v in any agent given a random number $r \in [0,1] \subseteq \mathbb{R}$ is defined as

$$s_p(v,u) = \begin{cases} 1 & \text{if } r \leq \dfrac{\sum_{u' \in \mathcal{N}(u)} s'(u')}{|\mathcal{N}(v)| \times s_{max}(v)} \\ 0 & \text{otherwise} \end{cases} \tag{7.13}$$

At the time of this thesis, the implemented interactions include activating fatty acids in an agent depending on the VLDL state of neighboring agents, representing the paracrine exchange of fatty acids.

7.3.3 Computational Implementation

Assigning a new model object to each agent as described in Equation (**7.11**) would be computationally unfeasible. Instead, the states in the different compartments since the KG itself does not change between managed through arrays of the numpy Python package for each node in the overall model. The model is implemented into the computational framework from Section 2.5 in the same way as described in Section 6.3 but instead performs logical operations on 2D **`numpy`** arrays. This way, a multi-compartmental model can be run completely on its own by a **`Model`** object. However, the specifications of the MASLD model, namely separating an extrahepatic single-compartment and hepatic multi-compartmental model, require further specifications. Therefore, I added a new **`ABM`** class to the computational framework, which handles the interactions between the two models. The initialization function requires two lists of submaps files for each model, respectively, the grid size n, and a seed value to ensure reproducability in probabilistic functions. In addition, I offer the option of adjusting the compartments to an input image by completely removing compartments whose position in the agent matrix corresponds to the white pixels in the image after adjusting the scales. This way the MASLD Map multi-compartmental model takes the shape of a liver, as shown in Figure 47C.

During the initialization of an **`ABM`** object, a **`Model`** object is generated for the extrahepatic and hepatic models, respectively. For the extrahepatic model, a 2D array $A = (a_{ij})_{1 \leq i,j \leq n}, 0 \leq a_{ij} \leq 1$ is being created, representing the differences between the agents due to heterogeneity. The parameter a_{ij} equals the probability of an active nutrient from the extrahepatic model being activated in the hepatic model. The matrix A is being created by first dividing the grid into $m \cdot m$ sections, each including $\frac{m}{n} \cdot \frac{m}{n}$ agents, that represent macroscale heterogeneity. For each section, a mean value μ_{micro} is being generated from a normal distribution of $\mu_{macro} = 0.6$ and standard deviation of $\sigma_{macro} = 0.3$, limited

between 0.2 and 1. Then, for the agents in each section, a_{ij} is drawn from a normal distribution distribution of μ_{micro} and $\sigma_{micro} = 0.01$.

The **`run`** function of the **`ABM`** class then runs the boolean simulation, accepting the number of steps T, node perturbations, such as the food intake frequency, as parameters, and a list of nodes that will be aggregated across neighboring agents in the hepatic model. During the simulations, at each step of T, both the extrahepatic and hepatic model execute their **`boolean_step`** function after another, as described in Section 6.4. At the beginning of each step, the nodes that are transported out of the intestinal compartment in the extrahepatic model project their state to the hepatic counterparts (of the same hash) with a probability a_{ij}. Additionally, the aggregation from Equation (**7.13**) is applied to all node pairs (v, u) supplied as parameters to the **`run`** function.

The computational complexity of the multi-compartmental model is approximated as $O(n^2 \cdot |V(G)| \cdot T)$, where n is the size of the agent grid, $|V(G)|$ represents the number of nodes in the graph G in all agents, and T is the number of steps in a Boolean experiment. One significant challenge is the computational storage requirement for maintaining the states of all nodes across all agents and time steps. After running the simulation, the MASLD Map should be able to visualize the state of each node in each step to give users the most detailed insight possible. Therefore, each evaluation in the Boolean model must be traceable afterward and can be implemented in two ways: Either saving the results of a single simulation for later access or recalculating the simulation for each new user request. The latter becomes impractical with the expected expansion of the KG, as the calculation time for a standard model could become too long. Increasing the resolution of the model, e.g., by increasing the number of steps to describe physiological processes such as food intake or the number of agents, will also increase computing times exponentially. It is to be expected that users will perform a single simulation under defined conditions and then focus on analyzing the results. From the point of view of user-friendliness, a longer duration simulation is much more tolerable for the user than longer loading times for each new visualization. We thus prioritized the efficient storage of intermediate results and rapid access to the data for custom requests.

A Boolean model with 500 nodes on a 100x100 agent grid over 1000 steps requires 5 GB of storage, excluding additional data like perturbations or multivalued states. This storage demand can be substantially reduced by recording changes in node states rather than storing every result. Storage needs can be reduced further by utilizing sparse arrays, as offered by the statsmodel library in Python, which stores only non-zero values. However,

sparse arrays are limited to two dimensions and require more initialization time. To address this issue, the 2D agent grid is reshaped into 1D, allowing the final storage for simulation results to be a list of sparse arrays, each representing node state changes in the agents at each step. The shape of these sparse arrays is $(|V(G)|, n^2)$. The most efficient method for creating a sparse matrix is using a coordinate (coo) matrix formed from three arrays indicating the rows, columns, and values of non-zero nodes. Pre-allocating these arrays to a size of $|V(G)| \times n^2$. At each step in the model, the difference between the numpy 2D arrays of its state in the current versus in the previous step is calculated for each node. For those values, where $\triangle s(v)_{i,j} \neq 0$, the difference itself, the node id, and the (i, j) position converted to the 1D space are attached to the next position in the array using a running index variable. The implementation of the storage of node states in sparse arrays is shown in Algorithm 3.

Notably, not all nodes in the model require re-evaluation at every step. A node's state is unlikely to change if there has been no change in the states of its incoming nodes. Exceptions exist for multivalued nodes or nodes with probabilistic edges. The model class maintains a set of nodes expected to change state, termed **`nodes_to_evaluate`**. This set includes multivalued nodes, nodes with outgoing probabilistic edges, and targets of nodes that underwent state changes in the previous step. In the MASLD Map, the food intake node is iterating between ON and OFF over intervals of several hundred steps. In these steps, the states of many nodes are not expected to change frequently, especially those of nutrients in response to digestive processes. On average, only about 20% of nodes require re-evaluation in the MASLD Map. Secondly, the **active()** function is invoked multiple times per step but does not change its output. Therefore, its output can be cached to prevent recalculations when no state change is anticipated. I employed Python's **functools.lru_cache** module to store the function's output and reset the cache for all nodes with changed states at the end of each step.

7.3.4 Multi-compartmental Simulations on Disease Maps

The complexity of the boolean model described in the previous makes an implementation in JavaScript unfeasible. I, therefore, developed a plugin for the MASLD Map, which utilizes the server-based plugin design described in Section 3.3.3. This way, users can perform the multi-compartmental simulations under user-defined parameters and present the results interactively on the MASDL Map. The following section describes the plugin's

design, focusing on the UI and main functionalities. The JavaScript and CSS files of the plugin are available at https://github.com/sbi-rostock/AIR/tree/master/AirPlugins_Server. The plugin UI is divided into three parts, as shown in Figure 47. The first step is initializing the model (Figure 47A), requesting the server to initialize an **`ABM`** object from a fixed list of submaps, which in its current version includes the "Digestion and Absorption.xml", "Adipose Tissue.xml", and "Endocrine Pancreas.xml" submaps for the extrahepatic model, and the "Hepatocyte.xml", "Hepatocyte Apoptosis.xml", "Vitamins and Trace Elements.xml", "Bile Acids.xml", and "Cholesterol Synthesis and Effects.xml" submaps for the hepatic model. The grid size of the multi-compartmental model is fixed, currently to 50*50 agents. The server returns a list of node names and hashes. The user can display and perturb the former in the UI (see below), while the latter allows the communications of (perturbed) nodes between the frontend and backend. Before the initialization of the ABM, users are given the possibility to enter a defined seed ID in the UI to replicate previous results.

The next part of the UI includes the configuration of model parameters (Figure 47B). Currently, the user can set the total number of steps the model will take and the quantity and quality of food. The quantity can be set by entering a user-defined frequency of the number of steps for the active or inactive state of the node for food intake. In this way, the model can also be defined for a specific dietary style, e.g., longer periods of fasting with meals of 1 or vice versa. Changing the duration of the frequency of food intake requires adjusting the total number of steps of the model to allow for more than one repetition of the sequence. The diet quality can be defined by adjusting the relative amount of each nutrient node as values between zero and one using a slider. Clicking on the “Run” button calls the run function on the server for the set number of steps and the conditions defined by the user. The value for the nutrient nodes represents the percentage of the active step in the execution of the model, i.e., if the value is set to 0.5, the node will be perturbed every second step, similar to Equation (**6.6**). The last field is the visualization of the results (Figure 47C).

After running the simulation, the activity for a selected map node can be visualized across all agents. For this purpose, I created a submap for the MALSD Map representing the multi-compartmental grid through gene SBML elements for each agent overlayed on a liver image, as shown in Figure 47C. Clicking on “show” will request the state values of the selected node for all agents and steps from the server employing the **restore_matrix_at_node** function of the extrahepatic **`Model`** object that recreates the values

from the sparse matrix. In the frontend, the values are mapped to color hex codes between white (0) and red (1) and highlighted on the respective agents of the liver image submap through the MINERVA plugin API. The simulation can also be performed comparatively by setting two different conditions. In this case, the simulation is run two times with different conditions, but the same seeds and the state values of nodes are subtracted. The visualization shows the difference between both conditions from downregulated (blue) to upregulated (red).

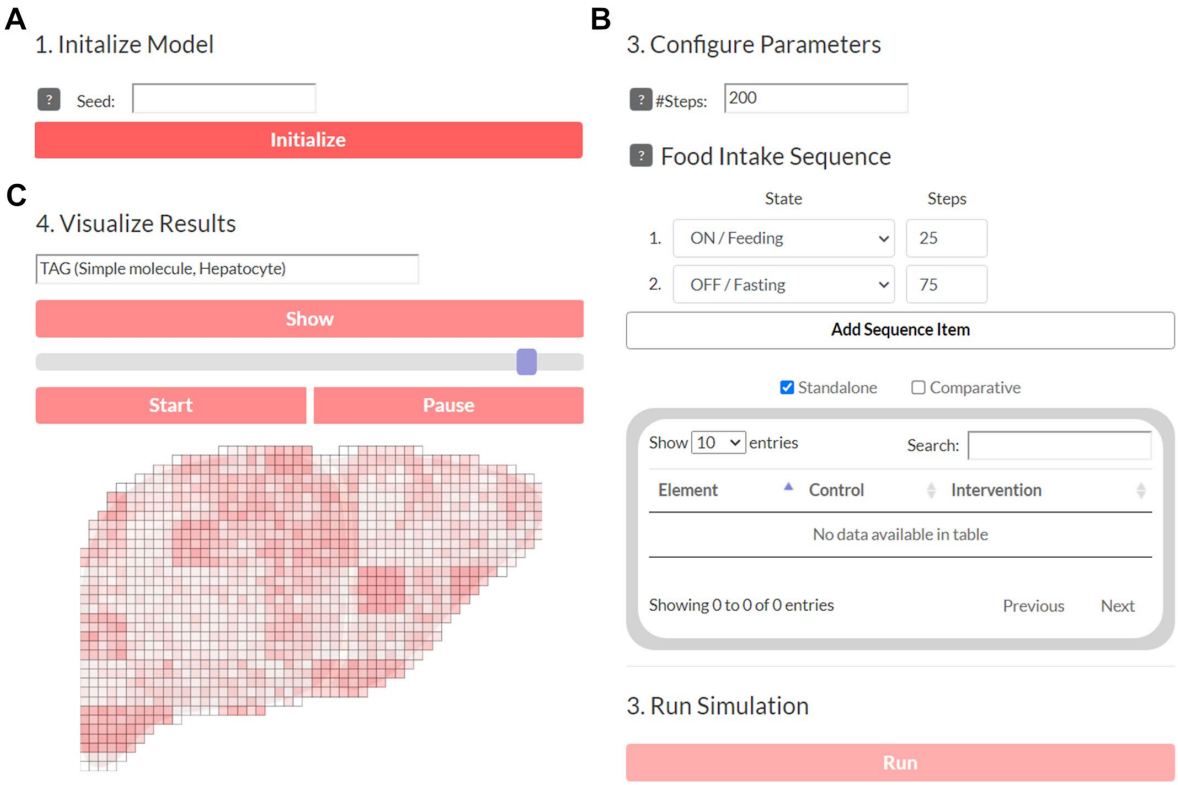

Figure 47: The user interface of the MASLD Map plugins. (A-B) In the current state of the plugin, users can parameterize the simulation by defining the quantity and quality of the diet. (C) Interactive, colored visualizations of selected nodes enable the interpretation of their spatiotemporal dynamics.

7.3.5 Strategies for Model Validation

In its current version of the model, the differences between compartments in the MASLD Map model are generated randomly. In this state, they should be viewed as a randomization of parameters to assess which disease phenotypes can be expected under certain conditions. The spatial modeling currently focuses solely on how the exchange between agents can spread across the compartments and impact disease progression. Therefore, the orientation of the sections in the multi-compartmental model itself does not yet have a clinical value.

The first step in validating the MASLD Map will be the validation of the Boolean model by testing the simulations by the model under changing conditions and in different agents, i.e., parameterizations. The response of the model to basic input specifications and the associated local and global attractor states should remain within (physiologically) reasonable limits. For example, in the absence of food intake and nutrient stores, no nutrient processes should be activated, and, if at all, only responses should be shown in (endocrine) processes related to starvation. The actual predictions from the mode should ideally be validated by reproducing empirical data from clinical studies using MASLD patients, if possible. To achieve the most comprehensive testing scenario, the evidence from the validation data should be related to many of the processes included in the MASLD Map. For example, Velanosi *et al*. measured pre- and postprandial lipidomic profiles of MASLD patients and healthy subjects [365]. Given that lipid metabolic pathways are integrated in detail in the MASLD Map, reproducing such measurements would substantially validate the model in a temporal and clinical context.

After validation of the Boolean model, the multi-compartmental design can be adjusted to actual clinical phenotypes. Many datasets are available that measure spatial properties of the liver, such as blood flow or metabolic activity [360], [361], [362], [363]. The compartmental design could be mapped to such imaging data, designing the model in a more realistic and clinically relevant way. Another possible application of the MASLD Map relates to moving from a macro- into the microscale. Spatial transcriptomics data provides high-resolution insights into single-cell molecular mechanisms and cellular phenotypes. Mapping the agents to individual cells in such data and modeling their communication processes could provide valuable insights into the processes in the tissue microenvironment of liver functions.

Concluding remarks

Around the time systems biology emerged, Ardeshir Bayat predicted in 2002: "The clinical research teams that will be most successful in the coming decades will be those that can switch effortlessly between the laboratory bench, clinical practice, and the use of these sophisticated computational tools" [366]. Indeed, 20 years later, systems biology plays an important role in biomolecular and, in particular, disease research [367], [368]. Large-scale KG models of diseases are employed to connect fields of research, predicting clinical effects from experimental data and therapeutic or diagnostic molecular mechanisms from clinical information. However, from my own experience, their general application in experimental research and clinical practice is not yet sufficiently established. In areas where large amounts of heterogeneous data are generated at high frequency, the effort required to create, curate, and adapt methods to the research question is too immense. In general, laboratory work and basic research are often more exploratory, working with various types of in vitro and in vivo samples requiring separate model specifications. Still, there should be a high demand for adaptable computational approaches, given that a large portion of generated data is still unpublished [147]. Of course, a completely generalized solution is not possible due to the far too great differences in various aspects relevant to all the areas of biomedical research. Instead, the solutions can be limited to a specific disease and its relevant scales, which corresponds to the principle of Disease Maps and is constantly confirmed by their success stories.

With the work described in this thesis, I want to advance computational models of diseases, particularly in the area of Disease Maps, by developing accessible data analysis platforms. During my PhD, I presented three Disease Maps, for each of which I developed specific approaches that can cover a wide range of potential disease-specific research questions. In this process, I followed three goals in particular: (i) curating KGs specifically tailored to the disease processes, (ii) developing approaches for KG and data analysis in the disease context, (iii) integrating these approaches into easily accessible Disease Map tools with intuitive presentation of the results. A central component during my research was the development of the 2DEA for KG-based data enrichment on Disease Maps. The 2DEA considers more information from the data and the KG than existing approaches, and its design allows for more valuable insights. Integrated into a plugin for MINERVA, the

approach has been employed for several Disease Map projects as a universal tool for exploratory data analysis. Using the AIR, I demonstrated the potential of the 2DEA to compare pharmacological mechanisms between ani-inflammatory drugs from time series data. Furthermore, I investigated cell type-specific metabolic processes and their gene regulations using cellular KGs generated from single-cell data. The 2DEA was also applied to the MASLD Map to infer differences in biological processes between groups of patients. Additionally, in the Sarcopenia and MASLD Map, I employed Boolean modeling approaches to simulate molecular disease mechanisms on a systemic scale. I have developed new tools for the MINERVA plugins for both maps that enable an advanced visualization of the simulations to facilitate their interpretation directly on the Disease Map in the web browser.

During these projects, I realized that the manual curation of KGs is still a major bottleneck. Machine learning-based text mining techniques are on the rise, but so far, no one has been able to replace manual curation in terms of extracting information from the literature and creating the layout. The AutoMap tool by the MINERVA team offers a first solution, automatically creating Disease Map projects by merging disease-relevant pathways from different resources. While this workflow still has its limitations, if Disease Maps could be created and kept up to date automatically, their applicability in everyday research would be greatly improved. I envision that such models can be created effortlessly for a specific research context and automatically integrated into experimental workflows to visualize and analyze data generated in the lab in real time, specific to the species, tissue, disease, or process being studied.

The efforts presented in this thesis support this development and offer solutions for a more in-depth and user-oriented data analysis on large-scale KGs. Together with the associated Disease Maps developed in this work, they can improve the accessibility and usability of computer-aided biological methods and subsequently accelerate biomedical research through facilitated knowledge generation and utilization.

Bibliography

[1] M. Hoch, S. Gupta, and O. Wolkenhauer, "Large-scale knowledge graph representations of disease processes," *Curr Opin Syst Biol*, vol. 38, p. 100517, Jun. 2024, doi: 10.1016/J.COISB.2024.100517.

[2] C. N. Serhan *et al.*, "The Atlas of Inflammation Resolution (AIR)," *Mol Aspects Med*, vol. 74, p. 100894, Aug. 2020, doi: 10.1016/j.mam.2020.100894.

[3] M. Hoch *et al.*, "Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps," *NPJ Syst Biol Appl*, vol. 8, no. 1, p. 13, Apr. 2022, doi: 10.1038/s41540-022-00222-z.

[4] M. Hoch *et al.*, "Cell-Type-Specific Gene Regulatory Networks of Pro-Inflammatory and Pro-Resolving Lipid Mediator Biosynthesis in the Immune System," *Int J Mol Sci*, vol. 24, no. 5, p. 4342, Mar. 2023, doi: 10.3390/IJMS24054342/S1.

[5] M. Hoch *et al.*, "Network analyses reveal new insights into the effect of multicomponent Tr14 compared to single-component diclofenac in an acute inflammation model," *Journal of Inflammation 2023 20:1*, vol. 20, no. 1, pp. 1–15, Mar. 2023, doi: 10.1186/S12950-023-00335-0.

[6] M. Hoch *et al.*, "In silico investigation of molecular networks linking gastrointestinal diseases, malnutrition, and sarcopenia," *Front Nutr*, vol. 9, Nov. 2022, doi: 10.3389/fnut.2022.989453.

[7] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology 2017 18:1*, vol. 18, no. 1, pp. 1–15, May 2017, doi: 10.1186/S13059-017-1215-1.

[8] C. Alcocer-Cuarón, A. L. Rivera, and V. M. Castaño, "Hierarchical structure of biological systems: A bioengineering approach," *Bioengineered*, vol. 5, no. 2, p. 73, Mar. 2014, doi: 10.4161/BIOE.26570.

[9] A. L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics 2004 5:2*, vol. 5, no. 2, pp. 101–113, Feb. 2004, doi: 10.1038/nrg1272.

[10] A. Gough *et al.*, "Biologically Relevant Heterogeneity: Metrics and Practical Insights," *SLAS discovery*, vol. 22, no. 3, p. 213, Mar. 2017, doi: 10.1177/2472555216682725.

[11] J. J. Tyson, T. Laomettachit, and P. Kraikivski, "Modeling the Dynamic Behavior of Biochemical Regulatory Networks," *J Theor Biol*, vol. 462, p. 514, Feb. 2019, doi: 10.1016/J.JTBI.2018.11.034.

[12] J. E. Purvis and G. Lahav, "Encoding and Decoding Cellular Information through Signaling Dynamics," *Cell*, vol. 152, no. 5, pp. 945–956, Feb. 2013, doi: 10.1016/J.CELL.2013.02.005.

[13] H. Kitano, "Computational systems biology," *Nature 2002 420:6912*, vol. 420, no. 6912, pp. 206–210, Nov. 2002, doi: 10.1038/nature01254.

[14] S. Barsi and B. Szalai, "Modeling in systems biology: Causal understanding before prediction?," *Patterns*, vol. 2, no. 6, p. 100280, Jun. 2021, doi: 10.1016/j.patter.2021.100280.

[15] Y. I. Wolf, M. I. Katsnelson, and E. V. Koonin, "Physical foundations of biological complexity," *Proc Natl Acad Sci U S A*, vol. 115, no. 37, pp. E8678–E8687, Sep. 2018, doi: 10.1073/PNAS.1807890115/ASSET/5B2DEDCA-0EAA-4FB2-8A46-8EE67C6F8DA7/ASSETS/GRAPHIC/PNAS.1807890115FIG02.JPEG.

[16] F. Mazzocchi, "Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory," *EMBO Rep*, vol. 9, no. 1, p. 10, Jan. 2008, doi: 10.1038/SJ.EMBOR.7401147.

[17] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, "A Modern Introduction to Probability and Statistics," 2005, doi: 10.1007/1-84628-168-7.

[18] R. A. Weinberg, "Coming full circle - From endless complexity to simplicity and back again," *Cell*, vol. 157, no. 1, pp. 267–271, Mar. 2014, doi: 10.1016/j.cell.2014.03.004.

[19] S. Huang, "Back to the biology in systems biology: What can we learn from biomolecular networks?," *Brief Funct Genomics*, vol. 2, no. 4, pp. 279–297, Feb. 2004, doi: 10.1093/BFGP/2.4.279.

[20] J. S. Yu and N. Bagheri, "Multi-class and multi-scale models of complex biological phenomena," *Curr Opin Biotechnol*, vol. 39, pp. 167–173, Jun. 2016, doi: 10.1016/J.COPBIO.2016.04.002.

[21] A. Millar-Wilson, Ó. Ward, E. Duffy, and G. Hardiman, "Multiscale modeling in the framework of biological systems and its potential for spaceflight biology studies," *iScience*, vol. 25, no. 11, p. 105421, Nov. 2022, doi: 10.1016/J.ISCI.2022.105421.

[22] F. Castiglione, F. Pappalardo, C. Bianca, G. Russo, and S. Motta, "Modeling biology spanning different scales: An open challenge," *Biomed Res Int*, vol. 2014, 2014, doi: 10.1155/2014/902545.

[23] M. M. Rahman, Y. Feng, T. E. Yankeelov, and J. T. Oden, "A fully coupled space–time multiscale modeling framework for predicting tumor growth," *Comput Methods Appl Mech Eng*, vol. 320, pp. 261–286, Jun. 2017, doi: 10.1016/J.CMA.2017.03.021.

[24] M. Garrido-Rodriguez, K. Zirngibl, O. Ivanova, S. Lobentanzer, and J. Saez-Rodriguez, "Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks," *Mol Syst Biol*, vol. 18, no. 7, pp. 1–15, 2022, doi: 10.15252/msb.202211036.

[25] G. A. Pavlopoulos *et al.*, "Using graph theory to analyze biological networks," *BioData Min*, vol. 4, no. 1, pp. 1–27, Apr. 2011, doi: 10.1186/1756-0381-4-10/FIGURES/11.

[26] M. Tantardini, F. Ieva, L. Tajoli, and C. Piccardi, "Comparing methods for comparing networks," *Scientific Reports 2019 9:1*, vol. 9, no. 1, pp. 1–19, Nov. 2019, doi: 10.1038/s41598-019-53708-y.

[27] G. Iñiguez, F. Battiston, and M. Karsai, "Bridging the gap between graphs and networks," *Communications Physics 2020 3:1*, vol. 3, no. 1, pp. 1–5, May 2020, doi: 10.1038/s42005-020-0359-6.

[28] Y. Yang, Y. Lu, and W. Yan, "A comprehensive review on knowledge graphs for complex diseases," *Brief Bioinform*, vol. 24, no. 1, Jan. 2023, doi: 10.1093/BIB/BBAC543.

[29] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data 2023 10:1*, vol. 10, no. 1, pp. 1–16, Feb. 2023, doi: 10.1038/s41597-023-01960-3.

[30] J. Gunawardena, "Models in biology: 'Accurate descriptions of our pathetic thinking,'" *BMC Biol*, vol. 12, no. 1, pp. 1–11, Apr. 2014, doi: 10.1186/1741-7007-12-29/FIGURES/3.

[31] M. M. Saint-Antoine and A. Singh, "Network inference in systems biology: recent developments, challenges, and applications," *Curr Opin Biotechnol*, vol. 63, pp. 89–98, Jun. 2020, doi: 10.1016/J.COPBIO.2019.12.002.

[32] P. Badia-i-Mompel *et al.*, "Gene regulatory network inference in the era of single-cell multi-omics," *Nature Reviews Genetics 2023 24:11*, vol. 24, no. 11, pp. 739–754, Jun. 2023, doi: 10.1038/s41576-023-00618-5.

[33] A. Dugourd *et al.*, "Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses," *Mol Syst Biol*, vol. 17, no. 1, pp. 1–17, 2021, doi: 10.15252/msb.20209730.

[34] M. H. Schaefer, L. Serrano, and M. A. Andrade-Navarro, "Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types," *Front Genet*, vol. 6, no. Aug, p. 137790, Aug. 2015, doi: 10.3389/FGENE.2015.00260/BIBTEX.

[35] S. A. Alcalá-Corona, S. Sandoval-Motta, J. Espinal-Enríquez, and E. Hernández-Lemus, "Modularity in Biological Networks," *Front Genet*, vol. 12, p. 701331, Sep. 2021, doi: 10.3389/FGENE.2021.701331/BIBTEX.

[36] E. Gonçalves *et al.*, "Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models," *Mol Biosyst*, vol. 9, no. 7, pp. 1576–1583, Jun. 2013, doi: 10.1039/C3MB25489E.

[37] S. Vaulont, M. Vasseur-Cognet, and A. Kahn, "Glucose regulation of gene transcription," *Journal of Biological Chemistry*, vol. 275, no. 41, pp. 31555–31558, Oct. 2000, doi: 10.1074/jbc.R000016200.

[38] A. Raue *et al.*, "Lessons Learned from Quantitative Dynamical Modeling in Systems Biology," *PLoS One*, vol. 8, no. 9, p. e74335, Sep. 2013, doi: 10.1371/JOURNAL.PONE.0074335.

[39] X. Liu *et al.*, "Robustness and lethality in multilayer biological molecular networks," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–12, Nov. 2020, doi: 10.1038/s41467-020-19841-3.

[40] S. Chaudhuri and A. Srivastava, "Network approach to understand biological systems: From single to multilayer networks," *Journal of Biosciences 2022 47:4*, vol. 47, no. 4, pp. 1–18, Sep. 2022, doi: 10.1007/S12038-022-00285-4.

[41] A. Dugourd and J. Saez-Rodriguez, "Footprint-based functional analysis of multiomic data," *Curr Opin Syst Biol*, vol. 15, pp. 82–90, Jun. 2019, doi: 10.1016/J.COISB.2019.04.002.

[42] D. Szklarczyk *et al.*, "The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/NAR/GKAC1000.

[43] H. Han *et al.*, "TRRUST: a reference database of human transcriptional regulatory interactions.," *Sci Rep*, vol. 5, no. 1, p. 11432, Jun. 2015, doi: 10.1038/srep11432.
[44] H. Y. Huang *et al.*, "miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions," *Nucleic Acids Res*, vol. 50, no. D1, p. D222, Jan. 2022, doi: 10.1093/NAR/GKAB1079.
[45] Y. Chen and X. Wang, "miRDB: an online database for prediction of functional microRNA targets," *Nucleic Acids Res*, vol. 48, no. D1, pp. D127–D131, Jan. 2020, doi: 10.1093/NAR/GKZ757.
[46] L. Cheng *et al.*, "LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res*, vol. 47, no. Database issue, p. D140, Jan. 2019, doi: 10.1093/NAR/GKY1051.
[47] H. Zhao *et al.*, "LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases," *Nucleic Acids Res*, vol. 51, no. D1, pp. D199–D207, Jan. 2023, doi: 10.1093/NAR/GKAC984.
[48] A. Elefsinioti, M. Ackermann, and A. Beyer, "Accounting for Redundancy when Integrating Gene Interaction Databases," *PLoS One*, vol. 4, no. 10, p. 7492, Oct. 2009, doi: 10.1371/JOURNAL.PONE.0007492.
[49] T. Chiang, D. Scholtens, D. Sarkar, R. Gentleman, and W. Huber, "Coverage and error models of protein-protein interaction data by directed graph analysis," *Genome Biol*, vol. 8, no. 9, pp. 1–14, Sep. 2007, doi: 10.1186/GB-2007-8-9-R186/FIGURES/7.
[50] M. Melkonian, C. Juigne, O. Dameron, G. Rabut, and E. Becker, "Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases," *Bioinformatics*, vol. 38, no. 6, pp. 1685–1691, Mar. 2022, doi: 10.1093/BIOINFORMATICS/BTAC013.
[51] A. L. Turinsky, S. Razick, B. Turner, I. M. Donaldson, and S. J. Wodak, "Literature curation of protein interactions: measuring agreement across major public databases," *Database (Oxford)*, vol. 2010, 2010, doi: 10.1093/DATABASE/BAQ026.
[52] T. G. O. Consortium *et al.*, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, May 2023, doi: 10.1093/GENETICS/IYAD031.
[53] S. Köhler *et al.*, "The Human Phenotype Ontology in 2021," *Nucleic Acids Res*, vol. 49, no. D1, p. D1207, Jan. 2021, doi: 10.1093/NAR/GKAA1043.
[54] D. Türei *et al.*, "Integrated intra- and intercellular signaling knowledge for multicellular omics analysis," *Mol Syst Biol*, vol. 17, no. 3, pp. 1–16, 2021, doi: 10.15252/msb.20209923.
[55] N. Le Novere, "Quantitative and logic modelling of molecular and gene networks," *Nature Reviews Genetics 2015 16:3*, vol. 16, no. 3, pp. 146–158, Feb. 2015, doi: 10.1038/nrg3885.
[56] N. Le Novère *et al.*, "The Systems Biology Graphical Notation," *Nat Biotechnol*, vol. 27, no. 8, pp. 735–741, Aug. 2009, doi: 10.1038/NBT.1558.
[57] H. Mi *et al.*, "Systems Biology Graphical Notation: Activity Flow language Level 1 Version 1.2," *J Integr Bioinform*, vol. 12, no. 2, p. 265, 2015, doi: 10.2390/BIECOLL-JIB-2015-265.

[58] A. Rougny *et al.*, "Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0," *J Integr Bioinform*, vol. 16, no. 2, Jun. 2019, doi: 10.1515/JIB-2019-0022.

[59] T. Vogt, T. Czauderna, and F. Schreiber, "Translation of SBGN maps: Process Description to Activity Flow," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–19, Oct. 2013, doi: 10.1186/1752-0509-7-115/FIGURES/22.

[60] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data 2016 3:1*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.

[61] K. Hanspers *et al.*, "Ten simple rules for creating reusable pathway models for computational analysis and visualization," *PLoS Comput Biol*, vol. 17, no. 8, Aug. 2021, doi: 10.1371/JOURNAL.PCBI.1009226.

[62] S. M. Keating *et al.*, " SBML Level 3: an extensible format for the exchange and reuse of biological models ," *Mol Syst Biol*, vol. 16, no. 8, p. 9110, Aug. 2020, doi: 10.15252/MSB.20199110/ASSET/89EBD24B-CBA2-49DB-943E-A26E16A76525/ASSETS/GRAPHIC/MSB199110-GRA-0001.PNG.

[63] A. Rougny *et al.*, "SBGN Bricks Ontology as a tool to describe recurring concepts in molecular networks," *Brief Bioinform*, vol. 22, no. 5, Sep. 2021, doi: 10.1093/BIB/BBAB049.

[64] A. Funahashi, M. Morohashi, Y. Matsuoka, A. Jouraku, and H. Kitano, "CellDesigner: A Graphical Biological Network Editor and Workbench Interfacing Simulator," in *Introduction to Systems Biology*, Totowa, NJ: Humana Press, 2007, pp. 422–434. doi: 10.1007/978-1-59745-531-2_21.

[65] C. Wrzodek, F. Büchel, M. Ruff, A. Dräger, and A. Zell, "Precise generation of systems biology models from KEGG pathways," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–12, Feb. 2013, doi: 10.1186/1752-0509-7-15/TABLES/3.

[66] M. Kutmon *et al.*, "PathVisio 3: An Extendable Pathway Analysis Toolbox," *PLoS Comput Biol*, vol. 11, no. 2, 2015, doi: 10.1371/JOURNAL.PCBI.1004085.

[67] E. Demir *et al.*, "The BioPAX community standard for pathway data sharing," *Nature Biotechnology 2010 28:9*, vol. 28, no. 9, pp. 935–942, Sep. 2010, doi: 10.1038/nbt.1666.

[68] L. Beltrame *et al.*, "The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways," *Bioinformatics*, vol. 27, no. 15, pp. 2127–2133, Aug. 2011, doi: 10.1093/BIOINFORMATICS/BTR339.

[69] K. Moutselos, I. Kanaris, A. Chatziioannou, I. Maglogiannis, and F. N. Kolisis, "KEGGconverter: A tool for the in-silico modelling of metabolic networks of the KEGG Pathways database," *BMC Bioinformatics*, vol. 10, no. 1, p. 324, Oct. 2009, doi: 10.1186/1471-2105-10-324/FIGURES/8.

[70] K.-E. Lee, M.-H. Jang, A.-R. Rhie, C. T. Thong, S.-D. Yang, and H.-S. Park, "Java DOM Parsers to Convert KGML into SBML and BioPAX Common Exchange Formats.," *Genomics Inform*, vol. 8, no. 2, pp. 94–96, Jun. 2010, doi: 10.5808/GI.2010.8.2.094.

[71] M. Martens *et al.*, "WikiPathways: connecting communities," *Nucleic Acids Res*, vol. 49, no. D1, pp. D613–D621, Jan. 2021, doi: 10.1093/nar/gkaa1024.

[72] M. Sari *et al.*, "SBGNViz: A Tool for Visualization and Complexity Management of SBGN Process Description Maps," *PLoS One*, vol. 10, no. 6, p. e0128985, Jun. 2015, doi: 10.1371/JOURNAL.PONE.0128985.
[73] H. Balci *et al.*, "Newt: a comprehensive web-based tool for viewing, constructing and analyzing biological maps," *Bioinformatics*, vol. 37, no. 10, pp. 1475–1477, Jun. 2021, doi: 10.1093/BIOINFORMATICS/BTAA850.
[74] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res*, 2017, doi: 10.1093/nar/gkw1092.
[75] M. Milacic *et al.*, "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Res*, vol. 52, no. D1, pp. D672–D678, Jan. 2024, doi: 10.1093/NAR/GKAD1025.
[76] A. Mazein *et al.*, "Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms," *NPJ Syst Biol Appl*, vol. 4, no. 1, p. 21, Dec. 2018, doi: 10.1038/s41540-018-0059-y.
[77] M. Ostaszewski *et al.*, "Community-driven roadmap for integrated disease maps.," *Brief Bioinform*, vol. 20, no. 2, pp. 659–670, Mar. 2019, doi: 10.1093/bib/bby024.
[78] K. A. Fujita *et al.*, "Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map," *Mol Neurobiol*, vol. 49, no. 1, pp. 88–102, Feb. 2014, doi: 10.1007/s12035-013-8489-4.
[79] V. Singh *et al.*, "Computational Systems Biology Approach for the Study of Rheumatoid Arthritis: From a Molecular Map to a Dynamical Model.," *Genom Comput Biol*, vol. 4, no. 1, p. 100050, Dec. 2018, doi: 10.18547/gcb.2018.vol4.iss1.e100050.
[80] A. Mazein *et al.*, "AsthmaMap: An expert-driven computational representation of disease mechanisms," *Clinical & Experimental Allergy*, vol. 48, no. 8, pp. 916–918, Aug. 2018, doi: 10.1111/cea.13211.
[81] A. Parton, V. McGilligan, M. Chemaly, M. O'Kane, and S. Watterson, "New models of atherosclerosis and multi-drug therapeutic interventions.," *Bioinformatics*, vol. 35, no. 14, pp. 2449–2457, Jul. 2019, doi: 10.1093/bioinformatics/bty980.
[82] M. Ostaszewski *et al.*, "COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms," *Mol Syst Biol*, vol. 17, no. 10, p. e10387, Oct. 2021, doi: 10.15252/msb.202110387.
[83] P. Gawron *et al.*, "MINERVA—a platform for visualization and curation of molecular interaction networks," *NPJ Syst Biol Appl*, vol. 2, no. 1, p. 16020, Dec. 2016, doi: 10.1038/npjsba.2016.20.
[84] D. Hoksza, P. Gawron, M. Ostaszewski, J. Hasenauer, and R. Schneider, "Closing the gap between formats for storing layout information in systems biology," *Brief Bioinform*, vol. 21, no. 4, pp. 1249–1260, Jul. 2020, doi: 10.1093/BIB/BBZ067.
[85] D. Hoksza, P. Gawron, M. Ostaszewski, and R. Schneider, "MolArt: a molecular structure annotation and visualization tool," *Bioinformatics*, vol. 34, no. 23, pp. 4127–4128, Dec. 2018, doi: 10.1093/BIOINFORMATICS/BTY489.

[86] X. Watkins, L. J. Garcia, S. Pundir, M. J. Martin, and U. Consortium, "ProtVista: visualization of protein sequence annotations," *Bioinformatics*, vol. 33, no. 13, pp. 2040–2041, Jul. 2017, doi: 10.1093/BIOINFORMATICS/BTX120.

[87] Y. Wu and T. Jiang, "Developments in FRET- and BRET-Based Biosensors," *Micromachines (Basel)*, vol. 13, no. 10, Oct. 2022, doi: 10.3390/MI13101789.

[88] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nat Rev Genet*, vol. 13, no. 4, pp. 227–232, Apr. 2012, doi: 10.1038/NRG3185.

[89] R. J. Nelson *et al.*, "Time of day as a critical variable in biology," *BMC Biology 2022 20:1*, vol. 20, no. 1, pp. 1–16, Jun. 2022, doi: 10.1186/S12915-022-01333-Z.

[90] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork, "Drug-induced regulation of target expression," *PLoS Comput Biol*, vol. 6, no. 9, 2010, doi: 10.1371/JOURNAL.PCBI.1000925.

[91] F. Iorio, J. Saez-Rodriguez, and D. di Bernardo, "Network based elucidation of drug response: From modulators to targets," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–9, Dec. 2013, doi: 10.1186/1752-0509-7-139/FIGURES/1.

[92] L. Toker, G. S. Nido, and C. Tzoulis, "Not every estimate counts – evaluation of cell composition estimation approaches in brain bulk tissue data," *Genome Med*, vol. 15, no. 1, pp. 1–14, Dec. 2023, doi: 10.1186/S13073-023-01195-2/FIGURES/6.

[93] K. Lu *et al.*, "Integrated network analysis of symptom clusters across disease conditions," *J Biomed Inform*, p. 103482, Jun. 2020, doi: 10.1016/j.jbi.2020.103482.

[94] C. M. Eckhardt *et al.*, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, no. 2, pp. 376–381, Feb. 2023, doi: 10.1007/S00167-022-07233-7/METRICS.

[95] K. P. F.R.S., "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[96] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *J Open Source Softw*, vol. 3, no. 29, p. 861, Sep. 2018, doi: 10.21105/JOSS.00861.

[97] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A Guide to Conquer the Biological Network Era Using Graph Theory," *Front Bioeng Biotechnol*, vol. 8, p. 504360, Jan. 2020, doi: 10.3389/FBIOE.2020.00034/BIBTEX.

[98] G. A. Pavlopoulos *et al.*, "Using graph theory to analyze biological networks," *BioData Min*, vol. 4, no. 1, pp. 1–27, Apr. 2011, doi: 10.1186/1756-0381-4-10/FIGURES/11.

[99] S. Klamt and A. von Kamp, "Computing paths and cycles in biological interaction graphs," *BMC Bioinformatics*, vol. 10, p. 181, Jun. 2009, doi: 10.1186/1471-2105-10-181.

[100] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A Guide to Conquer the Biological Network Era Using Graph Theory," *Front Bioeng Biotechnol*, vol. 8, p. 34, Jan. 2020, doi: 10.3389/FBIOE.2020.00034/BIBTEX.

[101] V. Janjić and N. Pržulj, "Biological function through network topology: a survey of the human diseasome," *Brief Funct Genomics*, vol. 11, no. 6, pp. 522–532, Nov. 2012, doi: 10.1093/BFGP/ELS037.

[102] P. Shannon *et al.*, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res*, vol. 13, no. 11, p. 2498, Nov. 2003, doi: 10.1101/GR.1239303.

[103] G. Scardoni, M. Petterlini, and C. Laudanna, "Analyzing biological network parameters with CentiScaPe.," *Bioinformatics*, vol. 25, no. 21, pp. 2857–2859, Nov. 2009, doi: 10.1093/bioinformatics/btp517.

[104] N. Alcaraz *et al.*, "KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape.," *BMC Syst Biol*, vol. 8, no. 1, p. 99, 2014, doi: 10.1186/s12918-014-0099-x.

[105] M. König, A. Dräger, and H.-G. Holzhütter, "CySBML: a Cytoscape plugin for SBML.," *Bioinformatics*, vol. 28, no. 18, pp. 2402–3, Sep. 2012, doi: 10.1093/bioinformatics/bts432.

[106] A. Hagberg, P. Swart, and D. Chult, *Exploring Network Structure, Dynamics, and Function Using NetworkX*. 2008.

[107] M. Garrido-Rodriguez, K. Zirngibl, O. Ivanova, S. Lobentanzer, and J. Saez-Rodriguez, "Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks," *Mol Syst Biol*, vol. 18, no. 7, pp. 1–15, 2022, doi: 10.15252/msb.202211036.

[108] M. R. Hidalgo *et al.*, "High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes," *Oncotarget*, vol. 8, no. 3, pp. 5160–5178, Dec. 2016, doi: 10.18632/ONCOTARGET.14107.

[109] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, and J. M. Stuart, "Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)," *Bioinformatics*, vol. 29, no. 21, pp. 2757–2764, Nov. 2013, doi: 10.1093/BIOINFORMATICS/BTT471.

[110] T. M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, "Identifying significantly impacted pathways: A comprehensive review and assessment," *Genome Biol*, vol. 20, no. 1, pp. 1–15, Oct. 2019, doi: 10.1186/S13059-019-1790-4/FIGURES/7.

[111] T. M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, "Identifying significantly impacted pathways: A comprehensive review and assessment," *Genome Biol*, vol. 20, no. 1, pp. 1–15, Oct. 2019, doi: 10.1186/S13059-019-1790-4/FIGURES/7.

[112] P. Khatri, M. Sirota, and A. J. Butte, "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges," *PLoS Comput Biol*, vol. 8, no. 2, p. e1002375, 2012, doi: 10.1371/JOURNAL.PCBI.1002375.

[113] "Data formats - GeneSetEnrichmentAnalysisWiki." Accessed: Apr. 09, 2024. [Online]. Available: https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

[114] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–50, Oct. 2005, doi: 10.1073/pnas.0506580102.

[115] P. D. Karp, P. E. Midford, R. Caspi, and A. Khodursky, "Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics," *BMC Genomics*, vol. 22, no. 1, pp. 1–11, Dec. 2021, doi: 10.1186/S12864-021-07502-8/FIGURES/1.

[116] N. Gerstner *et al.*, "GeneTrail 3: advanced high-throughput enrichment analysis," *Nucleic Acids Res*, vol. 48, no. W1, pp. W515–W520, Jul. 2020, doi: 10.1093/NAR/GKAA306.

[117] G. Hong, W. Zhang, H. Li, X. Shen, and Z. Guo, "Separate enrichment analysis of pathways for up- and downregulated genes," *J R Soc Interface*, vol. 11, no. 92, p. 20130950, Mar. 2014, doi: 10.1098/rsif.2013.0950.

[118] C. D. Warden, N. Kanaya, S. Chen, and Y.-C. Yuan, "BD-Func: a streamlined algorithm for predicting activation and inhibition of pathways," *PeerJ*, vol. 1, no. 1, p. e159, Sep. 2013, doi: 10.7717/peerj.159.

[119] A. Zito *et al.*, "Gene Set Enrichment Analysis of Interaction Networks Weighted by Node Centrality," *Front Genet*, vol. 12, p. 577623, Feb. 2021, doi: 10.3389/fgene.2021.577623.

[120] N. L. Catlett *et al.*, "Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data," *BMC Bioinformatics*, vol. 14, no. 1, Nov. 2013, doi: 10.1186/1471-2105-14-340.

[121] H. Mi and P. Thomas, "PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools," *Methods Mol Biol*, vol. 563, p. 123, 2009, doi: 10.1007/978-1-60761-175-2_7.

[122] B. T. Sherman *et al.*, "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)," *Nucleic Acids Res*, vol. 50, no. W1, pp. W216–W221, Jul. 2022, doi: 10.1093/NAR/GKAC194.

[123] E. Y. Chen *et al.*, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, no. 1, p. 128, Apr. 2013, doi: 10.1186/1471-2105-14-128.

[124] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.," *Bioinformatics*, vol. 21, no. 16, pp. 3448–3449, Aug. 2005, doi: 10.1093/bioinformatics/bti551.

[125] G. Bindea *et al.*, "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, Apr. 2009, doi: 10.1093/bioinformatics/btp101.

[126] J. Griss, G. Viteri, K. Sidiropoulos, V. Nguyen, A. Fabregat, and H. Hermjakob, "ReactomeGSA - Efficient Multi-Omics Comparative Pathway

Analysis," *Molecular and Cellular Proteomics*, vol. 19, no. 12, pp. 2115–2124, Dec. 2020, doi: 10.1074/mcp.TIR120.002155.

[127] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, "Causal analysis approaches in Ingenuity Pathway Analysis," *Bioinformatics*, vol. 30, no. 4, pp. 523–530, Feb. 2014, doi: 10.1093/bioinformatics/btt703.

[128] H. Nguyen *et al.*, "CPA: a web-based platform for consensus pathway analysis and interactive visualization," *Nucleic Acids Res*, vol. 49, no. W1, pp. W114–W124, Jul. 2021, doi: 10.1093/NAR/GKAB421.

[129] C. W. T. Koh, J. S. G. Ooi, E. Z. Ong, and K. R. Chan, "STAGEs: A web-based tool that integrates data visualization and pathway enrichment analysis for gene expression studies," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–12, May 2023, doi: 10.1038/s41598-023-34163-2.

[130] T. Helikar *et al.*, "The Cell Collective: Toward an open and collaborative approach to systems biology," *BMC Syst Biol*, vol. 6, no. 1, pp. 1–14, Aug. 2012, doi: 10.1186/1752-0509-6-96/FIGURES/8.

[131] T. Helikar and J. A. Rogers, "ChemChains: a platform for simulation and analysis of biochemical networks aimed to laboratory scientists," *BMC Syst Biol*, vol. 3, no. 1, p. 58, Dec. 2009, doi: 10.1186/1752-0509-3-58.

[132] G. Stoll, E. Viara, E. Barillot, and L. Calzone, "Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm," *BMC Syst Biol*, vol. 6, Aug. 2012, doi: 10.1186/1752-0509-6-116.

[133] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles, "Structural and functional analysis of cellular networks with CellNetAnalyzer," *BMC Syst Biol*, vol. 1, Jan. 2007, doi: 10.1186/1752-0509-1-2.

[134] C. Müssel, M. Hopfensitz, and H. A. Kestler, "BoolNet--an R package for generation, reconstruction and analysis of Boolean networks," *Bioinformatics*, vol. 26, no. 10, pp. 1378–1380, Apr. 2010, doi: 10.1093/BIOINFORMATICS/BTQ124.

[135] H. Klarner, A. Streck, and H. Siebert, "PyBoolNet: a python package for the generation, analysis and visualization of boolean networks," *Bioinformatics*, vol. 33, no. 5, pp. 770–772, Mar. 2017, doi: 10.1093/BIOINFORMATICS/BTW682.

[136] C. Chaouiya *et al.*, "SBML qualitative models: A model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–15, Dec. 2013, doi: 10.1186/1752-0509-7-135/FIGURES/7.

[137] S. S. Aghamiri, V. Singh, A. Naldi, T. Helikar, S. Soliman, and A. Niarakis, "Automated inference of Boolean models from molecular interaction maps using CaSQ," *Bioinformatics*, vol. 36, no. 16, pp. 4473–4482, Aug. 2020, doi: 10.1093/bioinformatics/btaa484.

[138] V. Singh *et al.*, "RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis," *Database (Oxford)*, vol. 2020, 2020, doi: 10.1093/DATABASE/BAAA017.

[139] V. Singh, A. Naldi, S. Soliman, and A. Niarakis, "A large-scale Boolean model of the rheumatoid arthritis fibroblast-like synoviocytes predicts drug

synergies in the arthritic joint," *NPJ Syst Biol Appl*, vol. 9, no. 1, Dec. 2023, doi: 10.1038/S41540-023-00294-5.

[140] A. Schäbitz *et al.*, "Spatial transcriptomics landscape of lesions from non-communicable inflammatory skin diseases," *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–13, Dec. 2022, doi: 10.1038/s41467-022-35319-w.

[141] A. Bayani, J. L. Dunster, J. J. Crofts, and M. R. Nelson, "Spatial considerations in the resolution of inflammation: Elucidating leukocyte interactions via an experimentally-calibrated agent-based model," *PLoS Comput Biol*, vol. 16, no. 11, p. e1008413, Nov. 2020, doi: 10.1371/JOURNAL.PCBI.1008413.

[142] J. Dutta-Moscato *et al.*, "A Multiscale Agent-Based in silico Model of Liver Fibrosis Progression," *Front Bioeng Biotechnol*, vol. 2, p. 18, May 2014, doi: 10.3389/FBIOE.2014.00018.

[143] M. Ponce-de-Leon *et al.*, "PhysiBoSS 2.0: a sustainable integration of stochastic Boolean and agent-based modelling frameworks," *npj Systems Biology and Applications 2023 9:1*, vol. 9, no. 1, pp. 1–12, Oct. 2023, doi: 10.1038/s41540-023-00314-4.

[144] A. Montagud, M. Ponce-de-Leon, and A. Valencia, "Systems biology at the giga-scale: Large multiscale models of complex, heterogeneous multicellular systems," *Curr Opin Syst Biol*, vol. 28, p. 100385, Dec. 2021, doi: 10.1016/J.COISB.2021.100385.

[145] F. Hausburg *et al.*, "(Re-)programming of subtype specific cardiomyocytes," Oct. 2017. doi: 10.1016/j.addr.2017.09.005.

[146] K. Wolstencroft *et al.*, "FAIRDOMHub: a repository and collaboration environment for sharing systems biology research," *Nucleic Acids Res*, vol. 45, no. D1, pp. D404–D407, Jan. 2017, doi: 10.1093/NAR/GKW1032.

[147] E. C. Bowers, J. Stephenson, M. Furlong, and K. S. Ramos, "Scope and financial impact of unpublished data and unused samples among U.S. academic and government researchers," *iScience*, vol. 26, no. 7, p. 107166, Jul. 2023, doi: 10.1016/J.ISCI.2023.107166.

[148] J. Scheel, M. Hoch, M. Wolfien, and S. Gupta, "NaviCenta – The disease map for placental research," *Placenta*, vol. 143, pp. 12–15, Nov. 2023, doi: 10.1016/J.PLACENTA.2023.09.007.

[149] M. Ostaszewski *et al.*, "COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms," *Mol Syst Biol*, vol. 17, no. 10, p. e10387, Oct. 2021, doi: 10.15252/msb.202110387.

[150] C. N. Serhan, "Pro-resolving lipid mediators are leads for resolution physiology," *Nature*, vol. 510, no. 7503, pp. 92–101, 2014, doi: 10.1038/NATURE13479.

[151] C. N. Serhan and J. Savill, "Resolution of inflammation: the beginning programs the end," *Nat Immunol*, vol. 6, no. 12, pp. 1191–1197, Dec. 2005, doi: 10.1038/ni1276.

[152] O. Bara, J. Day, and S. M. Djouadi, "Nonlinear state estimation for complex immune responses," *Proceedings of the IEEE Conference on Decision and Control*, pp. 3373–3378, 2013, doi: 10.1109/CDC.2013.6760399.

[153] V. Chiurchiù, A. Leuti, and M. Maccarrone, "Bioactive Lipids and Chronic Inflammation: Managing the Fire Within," *Front Immunol*, vol. 9, no. JAN, Jan. 2018, doi: 10.3389/FIMMU.2018.00038.
[154] M. Perretti, D. Cooper, J. Dalli, and L. V. Norling, "Immune resolution mechanisms in inflammatory arthritis," *Nature Reviews Rheumatology 2017 13:2*, vol. 13, no. 2, pp. 87–99, Jan. 2017, doi: 10.1038/nrrheum.2016.193.
[155] J. Viola and O. Soehnlein, "Atherosclerosis – A matter of unresolved inflammation," *Semin Immunol*, vol. 27, no. 3, pp. 184–193, May 2015, doi: 10.1016/J.SMIM.2015.03.013.
[156] P. Libby, I. Tabas, G. Fredman, and E. A. Fisher, "Inflammation and its Resolution as Determinants of Acute Coronary Syndromes," *Circ Res*, vol. 114, no. 12, pp. 1867–1879, Jun. 2014, doi: 10.1161/CIRCRESAHA.114.302699.
[157] S. I. Grivennikov, F. R. Greten, and M. Karin, "Immunity, Inflammation, and Cancer," *Cell*, vol. 140, no. 6, pp. 883–899, Mar. 2010, doi: 10.1016/J.CELL.2010.01.025.
[158] E. Brennan, P. Kantharidis, M. E. Cooper, and C. Godson, "Pro-resolving lipid mediators: regulators of inflammation, metabolism and kidney function," *Nature Reviews Nephrology 2021 17:11*, vol. 17, no. 11, pp. 725–739, Jul. 2021, doi: 10.1038/s41581-021-00454-y.
[159] E. A. Dennis and P. C. Norris, "Eicosanoid storm in infection and inflammation," *Nat Rev Immunol*, vol. 15, no. 8, pp. 511–523, Aug. 2015, doi: 10.1038/NRI3859.
[160] C. N. Serhan and B. D. Levy, "Resolvins in inflammation: Emergence of the pro-resolving superfamily of mediators," *Journal of Clinical Investigation*, vol. 128, no. 7, pp. 2657–2669, 2018, doi: 10.1172/JCI97943.
[161] C. N. Serhan, "Treating inflammation and infection in the 21st century: New hints from decoding resolution mediators and mechanisms," *FASEB Journal*, vol. 31, no. 4, pp. 1273–1288, 2017, doi: 10.1096/fj.201601222R.
[162] A. C. Doran, "Inflammation Resolution: Implications for Atherosclerosis," 2021, doi: 10.1161/CIRCRESAHA.121.319822.
[163] J. Rodriguez-Vita and T. Lawrence, "The resolution of inflammation and cancer," *Cytokine Growth Factor Rev*, vol. 21, no. 1, pp. 61–65, Feb. 2010, doi: 10.1016/j.cytogfr.2009.11.006.
[164] C. T. Robb, K. H. Regan, D. A. Dorward, and A. G. Rossi, "Key mechanisms governing resolution of lung inflammation," *Semin Immunopathol*, vol. 38, no. 4, pp. 425–448, Jul. 2016, doi: 10.1007/s00281-016-0560-6.
[165] B. D. Levy and C. N. Serhan, "Resolution of Acute Inflammation in the Lung," *Annu Rev Physiol*, vol. 76, no. 1, pp. 467–492, Feb. 2014, doi: 10.1146/annurev-physiol-021113-170408.
[166] B. E. Sansbury and M. Spite, "Resolution of acute inflammation and the role of resolvins in immunity, thrombosis, and vascular biology," *Circ Res*, vol. 119, no. 1, pp. 113–130, 2016, doi: 10.1161/CIRCRESAHA.116.307308.
[167] J. Pirault and M. Bäck, "Lipoxin and Resolvin Receptors Transducing the Resolution of Inflammation in Cardiovascular Disease.," *Front Pharmacol*, vol. 9, p. 1273, 2018, doi: 10.3389/fphar.2018.01273.

[168] J. Park, C. J. Langmead, and D. M. Riddy, "New Advances in Targeting the Resolution of Inflammation: Implications for Specialized Pro-Resolving Mediator GPCR Drug Discovery," *ACS Pharmacol Transl Sci*, vol. 3, no. 1, pp. 88–106, Feb. 2020, doi: 10.1021/acsptsci.9b00075.

[169] J. N. Fullerton and D. W. Gilroy, "Resolution of inflammation: a new therapeutic frontier," *Nat Rev Drug Discov*, vol. 15, no. 8, pp. 551–567, Aug. 2016, doi: 10.1038/nrd.2016.39.

[170] T. J. Ahmed, M. K. Kaneva, C. Pitzalis, D. Cooper, and M. Perretti, "Resolution of inflammation: examples of peptidergic players and pathways," *Drug Discov Today*, vol. 19, no. 8, pp. 1166–1171, Aug. 2014, doi: 10.1016/j.drudis.2014.05.020.

[171] C. K. Glass and G. Natoli, "Molecular control of activation and priming in macrophages," *Nature Immunology 2015 17:1*, vol. 17, no. 1, pp. 26–33, Dec. 2015, doi: 10.1038/ni.3306.

[172] H. Du *et al.*, "Tuning immunity through tissue mechanotransduction," *Nature Reviews Immunology 2022 23:3*, vol. 23, no. 3, pp. 174–188, Aug. 2022, doi: 10.1038/s41577-022-00761-w.

[173] A. Bayani, J. L. Dunster, J. J. Crofts, and M. R. Nelson, "Spatial considerations in the resolution of inflammation: Elucidating leukocyte interactions via an experimentally-calibrated agent-based model," *PLoS Comput Biol*, vol. 16, no. 11, Nov. 2020, doi: 10.1371/journal.pcbi.1008413.

[174] M. A. Sugimoto, L. P. Sousa, V. Pinho, M. Perretti, and M. M. Teixeira, "Resolution of Inflammation: What Controls Its Onset?," *Front Immunol*, vol. 7, no. April, p. 160, Apr. 2016, doi: 10.3389/fimmu.2016.00160.

[175] M. Peiseler and P. Kubes, "More friend than foe: The emerging role of neutrophils in tissue repair," *Journal of Clinical Investigation*, vol. 129, no. 7, pp. 2629–2639, Jul. 2019, doi: 10.1172/JCI124616.

[176] N. Azhar *et al.*, "A putative 'chemokine switch' that regulates systemic acute inflammation in humans," *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–14, May 2021, doi: 10.1038/s41598-021-88936-8.

[177] S. Minucci, R. L. Heise, M. S. Valentine, F. J. K. Gninzeko, and A. M. Reynolds, "Mathematical modeling of ventilator-induced lung inflammation," *bioRxiv*, p. 2020.06.03.132258, Nov. 2020, doi: 10.1101/2020.06.03.132258.

[178] A. Bensussen, E. R. Álvarez-Buylla, and J. Díaz, "SARS-CoV-2 Nsp5 Protein Causes Acute Lung Inflammation, A Dynamical Mathematical Model," *Frontiers in Systems Biology*, vol. 1, Dec. 2021, doi: 10.3389/FSYSB.2021.764155.

[179] P. Diaz *et al.*, "A Mathematical Model of the Immune System's Role in Obesity-Related Chronic Inflammation," *SIAM Undergrad Res Online*, vol. 2, no. 2, pp. 26–45, 2009, doi: 10.1137/08S010323.

[180] S. Maiti, W. Dai, R. C. Alaniz, J. Hahn, and A. Jayaraman, "Mathematical Modeling of Pro- and Anti-Inflammatory Signaling in Macrophages," *Processes 2015, Vol. 3, Pages 1-18*, vol. 3, no. 1, pp. 1–18, Dec. 2014, doi: 10.3390/PR3010001.

[181] S. E. Calvano *et al.*, "A network-based analysis of systemic inflammation in humans," *Nature 2005 437:7061*, vol. 437, no. 7061, pp. 1032–1037, Aug. 2005, doi: 10.1038/nature03985.
[182] A. Obaid *et al.*, "Model of the adaptive immune response system against HCV infection reveals potential immunomodulatory agents for combination therapy," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–19, Jun. 2018, doi: 10.1038/s41598-018-27163-0.
[183] K. Talaei *et al.*, "A Mathematical Model of the Dynamics of Cytokine Expression and Human Immune Cell Activation in Response to the Pathogen Staphylococcus aureus," *Front Cell Infect Microbiol*, vol. 11, Nov. 2021, doi: 10.3389/FCIMB.2021.711153.
[184] O. F. Voropaeva and T. V. Bayadilov, "A Mathematical Model of Aseptic Inflammation Dynamics," *Journal of Applied and Industrial Mathematics*, vol. 14, no. 4, pp. 779–791, Nov. 2020, doi: 10.1134/S1990478920040158/METRICS.
[185] M. M. Davis, C. M. Tato, and D. Furman, "Systems immunology: just getting started," *Nature Immunology 2017 18:7*, vol. 18, no. 7, pp. 725–732, Jun. 2017, doi: 10.1038/ni.3768.
[186] M. M. Davis, "Systems immunology," *Curr Opin Immunol*, vol. 65, pp. 79–82, Aug. 2020, doi: 10.1016/j.coi.2020.06.006.
[187] C. N. Serhan *et al.*, "The Atlas of Inflammation Resolution (AIR)," *Mol Aspects Med*, vol. 74, p. 100894, Aug. 2020, doi: 10.1016/j.mam.2020.100894.
[188] A. Martin, M. E. Ochagavia, L. C. Rabasa, J. Miranda, J. Fernandez-de-Cossio, and R. Bringas, "BisoGenet: a new tool for gene network building, visualization and analysis," *BMC Bioinformatics*, vol. 11, Feb. 2010, doi: 10.1186/1471-2105-11-91.
[189] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res*, vol. 32, no. Database issue, Jan. 2004, doi: 10.1093/NAR/GKH086.
[190] R. Oughtred *et al.*, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Sci*, vol. 30, no. 1, pp. 187–200, Jan. 2021, doi: 10.1002/PRO.3978.
[191] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database--2009 update," *Nucleic Acids Res*, vol. 37, no. Database issue, 2009, doi: 10.1093/NAR/GKN892.
[192] N. del Toro *et al.*, "The IntAct database: efficient access to fine-grained molecular interaction data," *Nucleic Acids Res*, vol. 50, no. D1, pp. D648–D653, Jan. 2022, doi: 10.1093/NAR/GKAB1006.
[193] L. Licata *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Res*, vol. 40, no. D1, pp. D857–D861, Jan. 2012, doi: 10.1093/NAR/GKR930.
[194] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic Acids Res*, vol. 47, no. D1, pp. D155–D162, Jan. 2019, doi: 10.1093/NAR/GKY1141.

[195] H. Y. Huang *et al.*, "miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions," *Nucleic Acids Res*, vol. 50, no. D1, p. D222, Jan. 2022, doi: 10.1093/NAR/GKAB1079.
[196] U. Schmitz, X. Lai, F. Winter, O. Wolkenhauer, J. Vera, and S. K. Gupta, "Cooperative gene regulation by microRNA pairs and their identification using a computational workflow," *Nucleic Acids Res*, vol. 42, no. 12, pp. 7539–7552, Jul. 2014, doi: 10.1093/NAR/GKU465.
[197] V. Matys *et al.*, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res*, vol. 34, no. Database issue, 2006, doi: 10.1093/NAR/GKJ143.
[198] H. Han *et al.*, "TRRUST: a reference database of human transcriptional regulatory interactions.," *Sci Rep*, vol. 5, no. 1, p. 11432, Jun. 2015, doi: 10.1038/srep11432.
[199] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden, and J. B. Demoulin, "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data," *Nucleic Acids Res*, vol. 38, no. 11, Mar. 2010, doi: 10.1093/NAR/GKQ149.
[200] L. A. Bovolenta, M. L. Acencio, and N. Lemke, "HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions," *BMC Genomics*, vol. 13, no. 1, pp. 1–10, Aug. 2012, doi: 10.1186/1471-2164-13-405/FIGURES/8.
[201] B. Zhou *et al.*, "EVLncRNAs 3.0: an updated comprehensive database for manually curated functional long non-coding RNAs validated by low-throughput experiments," *Nucleic Acids Res*, vol. 52, no. D1, pp. D98–D106, Jan. 2024, doi: 10.1093/NAR/GKAD1057.
[202] X. Lin *et al.*, "LncRNADisease v3.0: an updated database of long non-coding RNA-associated diseases," *Nucleic Acids Res*, vol. 52, no. D1, pp. D1365–D1369, Jan. 2024, doi: 10.1093/NAR/GKAD828.
[203] A. Mazein *et al.*, "Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms," *NPJ Syst Biol Appl*, vol. 4, no. 1, p. 21, Dec. 2018, doi: 10.1038/s41540-018-0059-y.
[204] M. Ostaszewski *et al.*, "Community-driven roadmap for integrated disease maps.," *Brief Bioinform*, vol. 20, no. 2, pp. 659–670, Mar. 2019, doi: 10.1093/bib/bby024.
[205] D. Hoksza, P. Gawron, M. Ostaszewski, E. Smula, and R. Schneider, "MINERVA API and plugins: opening molecular network analysis and visualization to the community," *Bioinformatics*, vol. 35, no. 21, pp. 4496–4498, Nov. 2019, doi: 10.1093/BIOINFORMATICS/BTZ286.
[206] P. Gawron *et al.*, "MINERVA—a platform for visualization and curation of molecular interaction networks," *NPJ Syst Biol Appl*, vol. 2, no. 1, p. 16020, Dec. 2016, doi: 10.1038/npjsba.2016.20.
[207] L. Chen *et al.*, "Inflammatory responses and inflammation-associated diseases in organs," *Oncotarget*, vol. 9, no. 6, p. 7204, Jan. 2018, doi: 10.18632/ONCOTARGET.23208.

[208] R. Laubenbacher *et al.*, "Building digital twins of the human immune system: toward a roadmap," *NPJ Digit Med*, vol. 5, no. 1, Dec. 2022, doi: 10.1038/S41746-022-00610-Z.
[209] M. Hoch *et al.*, "Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps," *NPJ Syst Biol Appl*, vol. 8, no. 1, p. 13, Apr. 2022, doi: 10.1038/s41540-022-00222-z.
[210] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–50, Oct. 2005, doi: 10.1073/pnas.0506580102.
[211] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[212] E. I. Boyle *et al.*, "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, Dec. 2004, doi: 10.1093/BIOINFORMATICS/BTH456.
[213] L. Martignetti, L. Calzone, E. Bonnet, E. Barillot, and A. Zinovyev, "ROMA: Representation and Quantification of Module Activity from Target Expression Data," *Front Genet*, vol. 7, no. FEB, p. 18, 2016, doi: 10.3389/FGENE.2016.00018.
[214] A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero, "Down-weighting overlapping genes improves gene set analysis.," *BMC Bioinformatics*, vol. 13, no. 1, p. 136, Jun. 2012, doi: 10.1186/1471-2105-13-136/TABLES/8.
[215] P. Czarnewski *et al.*, "Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification," *Nat Commun*, vol. 10, no. 1, p. 2892, Dec. 2019, doi: 10.1038/s41467-019-10769-x.
[216] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*. O'Reilly, 2018. [Online]. Available: https://books.google.de/books?id=cV34swEACAAJ
[217] M. Hornschuh, E. Wirthgen, M. Wolfien, K. P. Singh, O. Wolkenhauer, and J. Däbritz, "The role of epigenetic modifications for the pathogenesis of Crohn's disease," *Clin Epigenetics*, vol. 13, no. 1, pp. 1–14, Dec. 2021, doi: 10.1186/S13148-021-01089-3/TABLES/2.
[218] F. Genovese *et al.*, "Computational molecular interaction maps of signaling events within the olfactory epithelium," in *Chemical Senses*, 2022. doi: 10.1093/chemse/bjac031.
[219] H. K. Brittain, R. Scott, and E. Thomas, "The rise of the genome and personalised medicine," *Clinical Medicine*, vol. 17, no. 6, pp. 545–551, Dec. 2017, doi: 10.7861/CLINMEDICINE.17-6-545.
[220] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, p. 2156, Aug. 2011, doi: 10.1093/BIOINFORMATICS/BTR330.
[221] P. Tamayo, G. Steinhardt, A. Liberzon, and J. P. Mesirov, "The Limitations of Simple Gene Set Enrichment Analysis Assuming Gene Independence,"

Stat Methods Med Res, vol. 25, no. 1, p. 472, Feb. 2016, doi: 10.1177/0962280212460441.

[222] L. Chen *et al.*, "Inflammatory responses and inflammation-associated diseases in organs," *Oncotarget*, vol. 9, no. 6, p. 7204, Jan. 2018, doi: 10.18632/ONCOTARGET.23208.

[223] C. Nathan, "Points of control in inflammation," *Nature*, vol. 420, no. 6917, pp. 846–852, Dec. 2002, doi: 10.1038/NATURE01320.

[224] M. A. Sugimoto, J. P. Vago, M. Perretti, and M. M. Teixeira, "Mediators of the Resolution of the Inflammatory Response," *Trends Immunol*, vol. 40, no. 3, pp. 212–227, Mar. 2019, doi: 10.1016/J.IT.2019.01.007.

[225] V. Chiurchiù, A. Leuti, and M. Maccarrone, "Bioactive Lipids and Chronic Inflammation: Managing the Fire Within," *Front Immunol*, vol. 9, no. JAN, Jan. 2018, doi: 10.3389/FIMMU.2018.00038.

[226] P. C. Norris and C. N. Serhan, "Metabololipidomic Profiling of Functional Immunoresolvent Clusters and Eicosanoids in Mammalian Tissues," *Biochem Biophys Res Commun*, vol. 504, no. 3, p. 553, Oct. 2018, doi: 10.1016/J.BBRC.2018.03.037.

[227] A. Chatterjee *et al.*, "Biosynthesis of proresolving lipid mediators by vascular cells and tissues", doi: 10.1096/fj.201700082R.

[228] S. Hong *et al.*, "Maresin-like lipid mediators are produced by leukocytes and platelets and rescue reparative function of diabetes-impaired macrophages," *Chem Biol*, vol. 21, no. 10, pp. 1318–1329, Oct. 2014, doi: 10.1016/j.chembiol.2014.06.010.

[229] H. Zeng, "What is a cell type and how to define it?," *Cell*, vol. 185, no. 15, pp. 2739–2755, Jul. 2022, doi: 10.1016/J.CELL.2022.06.031.

[230] Z. Ye and C. A. Sarkar, "Towards a quantitative understanding of cell identity," *Trends Cell Biol*, vol. 28, no. 12, p. 1030, Dec. 2018, doi: 10.1016/J.TCB.2018.09.002.

[231] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell RNA-seq data," *Nature Reviews Genetics 2018 20:5*, vol. 20, no. 5, pp. 273–282, Jan. 2019, doi: 10.1038/s41576-018-0088-9.

[232] J. Liu, Z. Fan, W. Zhao, and X. Zhou, "Machine Intelligence in Single-Cell Data Analysis: Advances and New Challenges," *Front Genet*, vol. 12, p. 807, May 2021, doi: 10.3389/FGENE.2021.655536/BIBTEX.

[233] J. Li, W. Jiang, H. Han, J. Liu, B. Liu, and Y. Wang, "ScGSLC: An unsupervised graph similarity learning framework for single-cell RNA-seq data clustering," *Comput Biol Chem*, vol. 90, p. 107415, Feb. 2021, doi: 10.1016/J.COMPBIOLCHEM.2020.107415.

[234] S. T. Gal-Oz *et al.*, "ImmGen report: sexual dimorphism in the immune system transcriptome," *Nat Commun*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-12348-6.

[235] H. Yoshida *et al.*, "The cis-Regulatory Atlas of the Mouse Immune System," *Cell*, vol. 176, no. 4, pp. 897-912.e20, 2019, doi: 10.1016/j.cell.2018.12.036.

[236] D. Lee and K. H. Cho, "Topological estimation of signal flow in complex signaling networks," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–11, Mar. 2018, doi: 10.1038/s41598-018-23643-5.

[237] M. L. Waskom, "seaborn: statistical data visualization," *J Open Source Softw*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/JOSS.03021.
[238] H. Li, F. Wang, X. Guo, and Y. Jiang, "Decreased MEF2A Expression Regulated by Its Enhancer Methylation Inhibits Autophagy and May Play an Important Role in the Progression of Alzheimer's Disease," *Front Neurosci*, vol. 15, p. 669, Jun. 2021, doi: 10.3389/FNINS.2021.682247/BIBTEX.
[239] V. Kumar, S. Kaur, L. Kapil, C. Singh, and A. Singh, "HDAC11: A novel inflammatory biomarker in Huntington's disease," *EXCLI J*, vol. 21, p. 647, Jan. 2022, doi: 10.17179/EXCLI2022-4741.
[240] L. Sun *et al.*, "Loss of HDAC11 ameliorates clinical symptoms in a multiple sclerosis mouse model," *Life Sci Alliance*, vol. 1, no. 5, Oct. 2018, doi: 10.26508/LSA.201800039.
[241] Y. Liu, L. Yu, Y. Xu, X. Tang, and X. Wang, "Substantia nigra Smad3 signaling deficiency: relevance to aging and Parkinson's disease and roles of microglia, proinflammatory factors, and MAPK," *J Neuroinflammation*, vol. 17, no. 1, Dec. 2020, doi: 10.1186/S12974-020-02023-9.
[242] A. Deczkowska *et al.*, "Mef2C restrains microglial inflammatory response and is lost in brain ageing in an IFN-I-dependent manner," *Nat Commun*, vol. 8, no. 1, Dec. 2017, doi: 10.1038/S41467-017-00769-0.
[243] M. Gong *et al.*, "Abnormal microglial polarization induced by Arid1a deletion leads to neuronal differentiation deficits," *Cell Prolif*, vol. 55, no. 11, Nov. 2022, doi: 10.1111/CPR.13314.
[244] L. Su *et al.*, "Microglia homeostasis mediated by epigenetic ARID1A regulates neural progenitor cells response and leads to autism-like behaviors," *Molecular Psychiatry 2022*, pp. 1–15, Jul. 2022, doi: 10.1038/s41380-022-01703-7.
[245] C. Sousa *et al.*, "Single-cell transcriptomics reveals distinct inflammation-induced microglia signatures," *EMBO Rep*, vol. 19, no. 11, Nov. 2018, doi: 10.15252/EMBR.201846171.
[246] S. Xiang *et al.*, "Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients," *BMC Med Genomics*, vol. 11, no. 6, pp. 39–51, Dec. 2018, doi: 10.1186/S12920-018-0431-1/FIGURES/4.
[247] N. Jantaratnotai, A. Ling, J. Cheng, C. Schwab, P. L. Mcgeer, and J. G. Mclarnon, "Upregulation and expression patterns of the angiogenic transcription factor ets-1 in Alzheimer's disease brain," *J Alzheimers Dis*, vol. 37, no. 2, pp. 367–377, 2013, doi: 10.3233/JAD-122191.
[248] V. Waetzig *et al.*, "c-Jun N-terminal kinases (JNKs) mediate pro-inflammatory actions of microglia," *Glia*, vol. 50, no. 3, pp. 235–246, May 2005, doi: 10.1002/GLIA.20173.
[249] I. Diaz-Aparicio *et al.*, "Microglia Actively Remodel Adult Hippocampal Neurogenesis through the Phagocytosis Secretome," *The Journal of Neuroscience*, vol. 40, no. 7, p. 1453, Feb. 2020, doi: 10.1523/JNEUROSCI.0993-19.2019.

[250] J. Gruchot *et al.*, "Siponimod Modulates the Reaction of Microglial Cells to Pro-Inflammatory Stimulation," *Int J Mol Sci*, vol. 23, no. 21, p. 13278, Oct. 2022, doi: 10.3390/IJMS232113278.
[251] J. D. Aaker *et al.*, "Transcriptional Fingerprint of Hypomyelination in Zfp191null and Shiverer (Mbpshi) Mice," *ASN Neuro*, vol. 8, no. 5, Sep. 2016, doi: 10.1177/1759091416670749.
[252] D. C. Shippy, J. J. Watters, and T. K. Ulland, "Transcriptional response of murine microglia in Alzheimer's disease and inflammation," *BMC Genomics*, vol. 23, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S12864-022-08417-8/FIGURES/6.
[253] S. Prasad and K. Singh, "Interaction of USF1/USF2 and α-Pal/Nrf1 to Fmr-1 promoter increases in mouse brain during aging," *Biochem Biophys Res Commun*, vol. 376, no. 2, pp. 347–351, Nov. 2008, doi: 10.1016/J.BBRC.2008.08.155.
[254] J. H. Kim *et al.*, "DCLK1 promotes colorectal cancer stemness and aggressiveness via the XRCC5/COX2 axis," *Theranostics*, vol. 12, no. 12, pp. 5258–5271, 2022, doi: 10.7150/THNO.72037.
[255] M. Joo *et al.*, "Lipopolysaccharide-dependent interaction between PU.1 and c-Jun determines production of lipocalin-type prostaglandin D synthase and prostaglandin D2 in macrophages," *Am J Physiol Lung Cell Mol Physiol*, vol. 296, no. 5, May 2009, doi: 10.1152/AJPLUNG.90320.2008.
[256] M. Lappas, "Identification of SMAD3 as a Novel Mediator of Inflammation in Human Myometrium In Vitro," *Mediators Inflamm*, vol. 2018, 2018, doi: 10.1155/2018/3140420.
[257] F. Gao, M. I. Zafar, S. Jüttner, M. Höcker, and B. Wiedenmann, "Expression and Molecular Regulation of the Cox2 Gene in Gastroenteropancreatic Neuroendocrine Tumors and Antiproliferation of Nonsteroidal Anti-Inflammatory Drugs (NSAIDs)," *Med Sci Monit*, vol. 24, pp. 8125–8140, Nov. 2018, doi: 10.12659/MSM.912419.
[258] F. Cilenti *et al.*, "A PGE2-MEF2A axis enables context-dependent control of inflammatory gene expression," *Immunity*, vol. 54, no. 8, pp. 1665-1682.e14, Aug. 2021, doi: 10.1016/J.IMMUNI.2021.05.016.
[259] F. Hausburg, J. J. Jung, M. Hoch, M. Wolfien, C. Rimmbach, and R. David, "(Re-)Programming of Subtype specific Cardiomyocytes," p. in Revision, 2017.
[260] M. Wahlestedt *et al.*, "Critical Modulation of Hematopoietic Lineage Fate by Hepatic Leukemia Factor," *Cell Rep*, vol. 21, no. 8, p. 2251, Nov. 2017, doi: 10.1016/J.CELREP.2017.10.112.
[261] L. Jiang *et al.*, "Blockade of Notch signaling promotes acetaminophen-induced liver injury," *Immunol Res*, vol. 65, no. 3, pp. 739–749, Jun. 2017, doi: 10.1007/S12026-017-8913-3.
[262] H. Z. Imtiyaz *et al.*, "Hypoxia-inducible factor 2α regulates macrophage function in mouse models of acute and tumor inflammation," *J Clin Invest*, vol. 120, no. 8, p. 2699, Aug. 2010, doi: 10.1172/JCI39506.
[263] T. Veremeyko, A. W. Y. Yung, D. C. Anthony, T. Strekalova, and E. D. Ponomarev, "Early Growth Response Gene-2 Is Essential for M1 and M2

Macrophage Activation and Plasticity by Modulation of the Transcription Factor CEBPβ," *Front Immunol*, vol. 9, no. NOV, Nov. 2018, doi: 10.3389/FIMMU.2018.02515.

[264] Y. Oishi *et al.*, "SREBP1 Contributes to Resolution of Pro-inflammatory TLR4 Signaling by Reprogramming Fatty Acid Metabolism," *Cell Metab*, vol. 25, no. 2, pp. 412–427, Feb. 2017, doi: 10.1016/J.CMET.2016.11.009.

[265] L. T. Huang *et al.*, "Klf10 deficiency in mice exacerbates pulmonary inflammation by increasing expression of the proinflammatory molecule NPRA," *Int J Biochem Cell Biol*, vol. 79, pp. 231–238, Oct. 2016, doi: 10.1016/J.BIOCEL.2016.08.027.

[266] W. Zhang, Q. Li, D. Li, J. Li, D. Aki, and Y. C. Liu, "The E3 ligase VHL controls alveolar macrophage function via metabolic–epigenetic regulation," *J Exp Med*, vol. 215, no. 12, p. 3180, Dec. 2018, doi: 10.1084/JEM.20181211.

[267] S. Li *et al.*, "Mir-204 Regulates LPS-Induced A549 Cell Damage by Targeting FOXK2," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/7404671.

[268] N. Nejati Moharrami, E. B. Tande, L. Ryan, T. Espevik, and V. Boyartchuk, "RORα controls inflammatory state of human macrophages," *PLoS One*, vol. 13, no. 11, Nov. 2018, doi: 10.1371/JOURNAL.PONE.0207374.

[269] Y. Li *et al.*, "FIP200 is Involved in Murine Pseudomonas Infection by Regulating HMGB1 Intracellular Translocation," *Cellular Physiology and Biochemistry*, vol. 33, no. 6, pp. 1733–1744, 2014, doi: 10.1159/000362954.

[270] C. R. Cochrane *et al.*, "Trp53 and Rb1 regulate autophagy and ligand-dependent Hedgehog signaling," *J Clin Invest*, vol. 130, no. 8, pp. 4006–4018, Aug. 2020, doi: 10.1172/JCI132513.

[271] V. Moresi *et al.*, "Histone deacetylases 1 and 2 regulate autophagy flux and skeletal muscle homeostasis in mice," *Proc Natl Acad Sci U S A*, vol. 109, no. 5, pp. 1649–1654, Jan. 2012, doi: 10.1073/PNAS.1121159109/-/DCSUPPLEMENTAL.

[272] T. A. McCaffrey *et al.*, "High-level expression of Egr-1 and Egr-1–inducible genes in mouse and human atherosclerosis," *Journal of Clinical Investigation*, vol. 105, no. 5, p. 653, Mar. 2000, doi: 10.1172/JCI8592.

[273] W. D. Patino, J. G. Kang, S. Matoba, O. Y. Mian, B. R. Gochuico, and P. M. Hwang, "Atherosclerotic plaque macrophage transcriptional regulators are expressed in blood and modulated by tristetraprolin," *Circ Res*, vol. 98, no. 10, pp. 1282–1289, May 2006, doi: 10.1161/01.RES.0000222284.48288.28.

[274] J. Wang *et al.*, "Inhibition of c-Jun N-Terminal Kinase Attenuates Low Shear Stress–Induced Atherogenesis in Apolipoprotein E–Deficient Mice," *Molecular Medicine*, vol. 17, no. 9–10, p. 990, 2011, doi: 10.2119/MOLMED.2011.00073.

[275] E. Harja *et al.*, "Early growth response-1 promotes atherogenesis: mice deficient in early growth response-1 and apolipoprotein E display decreased atherosclerosis and vascular inflammation," *Circ Res*, vol. 94, no. 3, pp. 333–339, Feb. 2004, doi: 10.1161/01.RES.0000112405.61577.95.

[276] M. Stula *et al.*, "Influence of sustained mechanical stress on Egr-1 mRNA expression in cultured human endothelial cells," *Mol Cell Biochem*, vol. 210, no. 1–2, pp. 101–108, 2000, doi: 10.1023/A:1007126218740.

[277] C. T. Keith, A. A. Borisy, and B. R. Stockwell, "Multicomponent therapeutics for networked systems," *Nat Rev Drug Discov*, vol. 4, no. 1, pp. 71–78, 2005, doi: 10.1038/nrd1609.

[278] A. Moya-García *et al.*, "Structural and Functional View of Polypharmacology," *Sci Rep*, vol. 7, no. 1, pp. 1–14, Dec. 2017, doi: 10.1038/s41598-017-10012-x.

[279] F. S. Dreyer *et al.*, "A web platform for the network analysis of high-throughput data in melanoma and its use to investigate mechanisms of resistance to anti-PD1 immunotherapy," *Biochim Biophys Acta Mol Basis Dis*, vol. 1864, no. 6, pp. 2315–2328, 2018, doi: 10.1016/j.bbadis.2018.01.020.

[280] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," *Drug Discov Today*, vol. 12, no. 1–2, pp. 34–42, 2007, doi: 10.1016/j.drudis.2006.11.008.

[281] M. Kibble, N. Saarinen, J. Tang, K. Wennerberg, S. Mäkelä, and T. Aittokallio, "Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products," *Nat Prod Rep*, vol. 32, no. 8, pp. 1249–1266, 2015, doi: 10.1039/c5np00005j.

[282] P. Li, J. Chen, W. Zhang, B. Fu, and W. Wang, "Transcriptome inference and systems approaches to polypharmacology and drug discovery in herbal medicine," *J Ethnopharmacol*, vol. 195, pp. 127–136, Jan. 2017, doi: 10.1016/j.jep.2016.10.020.

[283] A. Moya-García *et al.*, "Structural and Functional View of Polypharmacology," *Sci Rep*, vol. 7, no. 1, pp. 1–14, Dec. 2017, doi: 10.1038/s41598-017-10012-x.

[284] M. J. Parnham and G. Geisslinger, "Pharmacological plasticity—How do you hit a moving target?," *Pharmacol Res Perspect*, vol. 7, no. 6, p. e00532, Dec. 2019, doi: 10.1002/prp2.532.

[285] R. R. Ramsay, M. R. Popovic-Nikolic, K. Nikolic, E. Uliassi, and M. L. Bolognesi, "A perspective on multi-target drug discovery and design for complex diseases," *Clin Transl Med*, vol. 7, no. 1, p. 3, Dec. 2018, doi: 10.1186/s40169-017-0181-2.

[286] J. M. Bennett, G. Reeves, G. E. Billman, and J. P. Sturmberg, "Inflammation-nature's way to efficiently respond to all types of challenges: Implications for understanding and managing 'the epidemic' of chronic diseases," *Front Med (Lausanne)*, vol. 5, no. NOV, p. 316, 2018, doi: 10.3389/fmed.2018.00316.

[287] T. W. Corson and C. M. Crews, "Molecular Understanding and Modern Application of Traditional Medicines: Triumphs and Trials," *Cell*, vol. 130, no. 5, pp. 769–774, Sep. 2007, doi: 10.1016/j.cell.2007.08.021.

[288] N. E. Thomford *et al.*, "Natural Products for Drug Discovery in the 21st Century: Innovations for Novel Drug Discovery," *International Journal of Molecular Sciences 2018, Vol. 19, Page 1578*, vol. 19, no. 6, p. 1578, May 2018, doi: 10.3390/IJMS19061578.

[289] W. Zhang, Y. Huai, Z. Miao, A. Qian, and Y. Wang, "Systems Pharmacology for Investigation of the Mechanisms of Action of Traditional Chinese Medicine in Drug Discovery," *Front Pharmacol*, vol. 10, p. 743, Jul. 2019, doi: 10.3389/fphar.2019.00743.

[290] F. Cheng, I. A. Kovács, and A. L. Barabási, "Network-based prediction of drug combinations," *Nat Commun*, vol. 10, no. 1, pp. 1–11, Mar. 2019, doi: 10.1038/s41467-019-09186-x.
[291] K. A. Ryall and A. C. Tan, "Systems biology approaches for advancing the discovery of effective drug combinations Rajarshi Guha," *J Cheminform*, vol. 7, no. 1, p. 7, Dec. 2015, doi: 10.1186/s13321-015-0055-9.
[292] F. Iorio, J. Saez-Rodriguez, and D. di Bernardo, "Network based elucidation of drug response: from modulators to targets," *BMC Syst Biol*, vol. 7, no. 1, p. 139, Dec. 2013, doi: 10.1186/1752-0509-7-139.
[293] E. L. Leung, Z. W. Cao, Z. H. Jiang, H. Zhou, and L. Liu, "Network-based drug discovery by integrating systems biology and computational technologies," *Brief Bioinform*, vol. 14, no. 4, pp. 491–505, 2013, doi: 10.1093/bib/bbs043.
[294] P. Li, J. Chen, W. Zhang, B. Fu, and W. Wang, "Transcriptome inference and systems approaches to polypharmacology and drug discovery in herbal medicine," *J Ethnopharmacol*, vol. 195, pp. 127–136, Jan. 2017, doi: 10.1016/j.jep.2016.10.020.
[295] G. S. Laurent *et al.*, "Deep sequencing transcriptome analysis of murine wound healing: Effects of a multicomponent, multitarget natural product Therapy-Tr14," *Front Mol Biosci*, vol. 4, no. AUG, p. 57, Aug. 2017, doi: 10.3389/fmolb.2017.00057.
[296] G. St. Laurent *et al.*, "RNAseq analysis of treatment-dependent signaling changes during inflammation in a mouse cutaneous wound healing model," *BMC Genomics*, vol. 22, no. 1, p. 854, Dec. 2021, doi: 10.1186/s12864-021-08083-2.
[297] T. J. Gan, "Diclofenac: An update on its mechanism of action and safety profile," *Curr Med Res Opin*, vol. 26, no. 7, pp. 1715–1731, 2010, doi: 10.1185/03007995.2010.486301.
[298] N. K. Evanson, "Diclofenac," in *xPharm: The Comprehensive Pharmacology Reference*, Elsevier, 2007, pp. 1–7. doi: 10.1016/B978-008055232-3.61588-0.
[299] K. Muders *et al.*, "Effects of Traumeel (Tr14) on recovery and inflammatory immune response after repeated bouts of exercise: a double-blind RCT," *Eur J Appl Physiol*, vol. 117, no. 3, pp. 591–605, Mar. 2017, doi: 10.1007/s00421-017-3554-8.
[300] K. Muders *et al.*, "Effects of Traumeel (Tr14) on Exercise-Induced Muscle Damage Response in Healthy Subjects: A Double-Blind RCT," *Mediators Inflamm*, vol. 2016, pp. 1–9, 2016, doi: 10.1155/2016/1693918.
[301] P. M. Jordan *et al.*, "The natural combination medicine traumeel (Tr14) improves resolution of inflammation by promoting the biosynthesis of specialized pro-resolving mediators," *Pharmaceuticals*, vol. 14, no. 11, p. 1123, Nov. 2021, doi: 10.3390/ph14111123.
[302] F. M. Khan *et al.*, "Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures," *Nat Commun*, vol. 8, no. 1, pp. 1–15, Aug. 2017, doi: 10.1038/s41467-017-00268-2.

[303] A. Cebrián-Prats, A. Pinto, À. González-Lafont, P. A. Fernandes, and J. M. Lluch, "The role of acetylated cyclooxygenase-2 in the biosynthesis of resolvin precursors derived from eicosapentaenoic acid," *Org Biomol Chem*, vol. 20, no. 6, pp. 1260–1274, Feb. 2022, doi: 10.1039/d1ob01932e.

[304] A. M. Hidalgo-Estévez, K. Stamatakis, M. Jiménez-Martínez, R. López-Pérez, and M. Fresno, "Cyclooxygenase 2-Regulated Genes an Alternative Avenue to the Development of New Therapeutic Drugs for Colorectal Cancer," *Front Pharmacol*, vol. 11, p. 533, Apr. 2020, doi: 10.3389/fphar.2020.00533.

[305] P. Blomgran, M. Hammerman, and P. Aspenberg, "Systemic corticosteroids improve tendon healing when given after the early inflammatory phase," *Sci Rep*, vol. 7, no. 1, Dec. 2017, doi: 10.1038/S41598-017-12657-0.

[306] M. Parisien *et al.*, "Acute inflammatory response via neutrophil activation protects against the development of chronic pain," *Sci Transl Med*, vol. 14, no. 644, May 2022, doi: 10.1126/SCITRANSLMED.ABJ9954.

[307] T. Schmid and B. Brüne, "Prostanoids and Resolution of Inflammation – Beyond the Lipid-Mediator Class Switch," *Front Immunol*, vol. 12, p. 2838, Jul. 2021, doi: 10.3389/fimmu.2021.714042.

[308] G. Kaur and O. Silakari, "Multiple target-centric strategy to tame inflammation," *Future Med Chem*, vol. 9, no. 12, pp. 1361–1376, Aug. 2017, doi: 10.4155/fmc-2017-0050.

[309] A. Koeberle and O. Werz, "Multi-target approach for natural products in inflammation," *Drug Discov Today*, vol. 19, no. 12, pp. 1871–1882, 2014, doi: 10.1016/j.drudis.2014.08.006.

[310] S. Jude and S. Gopi, "Multitarget approach for natural products in inflammation," *Inflammation and Natural Products*, pp. 83–111, Jan. 2021, doi: 10.1016/b978-0-12-819218-4.00004-3.

[311] K. Thobe, C. Kuznia, C. Sers, and H. Siebert, "Evaluating uncertainty in signaling networks using logical modeling," *Front Physiol*, vol. 9, no. OCT, p. 1335, Oct. 2018, doi: 10.3389/FPHYS.2018.01335/FULL.

[312] D. Lee and K. H. Cho, "Topological estimation of signal flow in complex signaling networks," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–11, Mar. 2018, doi: 10.1038/s41598-018-23643-5.

[313] H. Sanhedrai, J. Gao, A. Bashan, M. Schwartz, S. Havlin, and B. Barzel, "Reviving a failed network through microscopic interventions," *Nature Physics 2022 18:3*, vol. 18, no. 3, pp. 338–349, Jan. 2022, doi: 10.1038/s41567-021-01474-y.

[314] M. Ney *et al.*, "Systematic review with meta-analysis: Nutritional screening and assessment tools in cirrhosis," *Liver International*, vol. 40, no. 3, pp. 664–673, Mar. 2020, doi: 10.1111/liv.14269.

[315] X. Theodoridis, M. G. Grammatikopoulou, A. Petalidou, S.-M. Kontonika, S. P. Potamianos, and D. P. Bogdanos, "A Systematic Review of Medical Nutrition Therapy Guidelines for Liver Cirrhosis: Do We Agree?," *Nutrition in Clinical Practice*, vol. 35, no. 1, pp. 98–107, Feb. 2020, doi: 10.1002/ncp.10393.

[316] M. T. Siddiqui, W. Al-Yaman, A. Singh, and D. F. Kirby, "Short-Bowel Syndrome: Epidemiology, Hospitalization Trends, In-Hospital Mortality,

and Healthcare Utilization," *Journal of Parenteral and Enteral Nutrition*, vol. 45, no. 7, pp. 1441–1455, Sep. 2021, doi: 10.1002/jpen.2051.

[317] F. Meyer and L. Valentini, "Disease-Related Malnutrition and Sarcopenia as Determinants of Clinical Outcome," *Visc Med*, vol. 35, no. 5, pp. 282–291, Oct. 2019, doi: 10.1159/000502867.

[318] L. Ehlers *et al.*, "Preclinical insights into the gut-skeletal muscle axis in chronic gastrointestinal diseases," *J Cell Mol Med*, vol. 24, no. 15, pp. 8304–8314, Aug. 2020, doi: 10.1111/jcmm.15554.

[319] A. Ganapathy and J. W. Nieves, "Nutrition and Sarcopenia—What Do We Know?," *Nutrients*, vol. 12, no. 6, pp. 1–25, Jun. 2020, doi: 10.3390/NU12061755.

[320] M. Bojko, "Causes of Sarcopenia in Liver Cirrhosis," *Clin Liver Dis (Hoboken)*, vol. 14, no. 5, pp. 167–170, Nov. 2019, doi: 10.1002/CLD.851.

[321] O. M. Nardone *et al.*, "Inflammatory Bowel Diseases and Sarcopenia: The Role of Inflammation and Gut Microbiota in the Development of Muscle Failure," *Front Immunol*, vol. 12, p. 2783, Jul. 2021, doi: 10.3389/FIMMU.2021.694217.

[322] K. Norman *et al.*, "Increased intestinal permeability in malnourished patients with liver cirrhosis," *Eur J Clin Nutr*, vol. 66, no. 10, pp. 1116–1119, Oct. 2012, doi: 10.1038/EJCN.2012.104.

[323] T. Kurosawa *et al.*, "Liver fibrosis-induced muscle atrophy is mediated by elevated levels of circulating TNFα," *Cell Death & Disease 2021 12:1*, vol. 12, no. 1, pp. 1–16, Jan. 2021, doi: 10.1038/s41419-020-03353-5.

[324] B. Sharma and R. Dabur, "Role of Pro-inflammatory Cytokines in Regulation of Skeletal Muscle Metabolism: A Systematic Review," *Curr Med Chem*, vol. 27, no. 13, pp. 2161–2188, Nov. 2020, doi: 10.2174/0929867326666181129095309.

[325] A. L. Barabási, G. Menichetti, and J. Loscalzo, "The unmapped chemical complexity of our diet," *Nature Food 2019 1:1*, vol. 1, no. 1, pp. 33–37, Dec. 2019, doi: 10.1038/s43016-019-0005-1.

[326] A. Tripathi *et al.*, "The gut–liver axis and the intersection with the microbiome," *Nature Reviews Gastroenterology & Hepatology 2018 15:7*, vol. 15, no. 7, pp. 397–411, May 2018, doi: 10.1038/s41575-018-0011-z.

[327] B. Egan and J. R. Zierath, "Exercise Metabolism and the Molecular Regulation of Skeletal Muscle Adaptation," *Cell Metab*, vol. 17, no. 2, pp. 162–184, Feb. 2013, doi: 10.1016/J.CMET.2012.12.012.

[328] M. Chudtong, A. De Gaetano, M. Chudtong, and A. De Gaetano, "A mathematical model of food intake," *Mathematical Biosciences and Engineering 2021 2:1238*, vol. 18, no. 2, pp. 1238–1279, 2021, doi: 10.3934/MBE.2021067.

[329] O. Röhrle, J. B. Davidson, and A. J. Pullan, "A physiologically based, multi-scale model of skeletal muscle structure and function," *Front Physiol*, vol. 3 SEP, p. 358, 2012, doi: 10.3389/FPHYS.2012.00358/BIBTEX.

[330] L. R. Smith, G. Meyer, and R. L. Lieber, "Systems analysis of biological networks in skeletal muscle function," *Wiley Interdiscip Rev Syst Biol Med*, vol. 5, no. 1, p. 55, Jan. 2013, doi: 10.1002/WSBM.1197.

[331] E. M. Maldonado *et al.*, "Multi-scale, whole-system models of liver metabolic adaptation to fat and sugar in non-alcoholic fatty liver disease," *npj Systems Biology and Applications 2018 4:1*, vol. 4, no. 1, pp. 1–10, Aug. 2018, doi: 10.1038/s41540-018-0070-3.

[332] Y. Zhao, R. E. Barrere-Cain, and X. Yang, "Nutritional systems biology of type 2 diabetes," *Genes & Nutrition 2015 10:5*, vol. 10, no. 5, pp. 1–18, Jul. 2015, doi: 10.1007/S12263-015-0481-3.

[333] U. J. F. Tietge *et al.*, "Alterations in glucose metabolism associated with liver cirrhosis persist in the clinically stable long-term course after liver transplantation," *Liver Transplantation*, vol. 10, no. 8, pp. 1030–1040, Aug. 2004, doi: 10.1002/LT.20147.

[334] X. P. Bai, Y. M. Fan, L. Zhang, G. H. Yang, and X. Li, "Influence of Liver Cirrhosis on Blood Glucose, Insulin Sensitivity and Islet Function in Mice," *Am J Med Sci*, vol. 362, no. 4, pp. 403–417, Oct. 2021, doi: 10.1016/J.AMJMS.2021.07.005.

[335] A. Dhaliwal and M. J. Armstrong, "Sarcopenia in cirrhosis: A practical overview," *Clinical Medicine*, vol. 20, no. 5, pp. 489–492, Sep. 2020, doi: 10.7861/clinmed.2020-0089.

[336] M. Ebadi, R. A. Bhanji, V. C. Mazurak, and A. J. Montano-Loza, "Sarcopenia in cirrhosis: from pathogenesis to interventions," *J Gastroenterol*, vol. 54, no. 10, pp. 845–859, Oct. 2019, doi: 10.1007/S00535-019-01605-6/FIGURES/3.

[337] L. Laffel, "Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes," *Diabetes Metab Res Rev*, vol. 15, no. 6, pp. 412–426, 1999, doi: 10.1002/(sici)1520-7560(199911/12)15:6<412::aid-dmrr72>3.0.co;2-8.

[338] M. E. Rinella *et al.*, "A multisociety Delphi consensus statement on new fatty liver disease nomenclature.," *J Hepatol*, vol. 79, no. 6, pp. 1542–1556, Dec. 2023, doi: 10.1016/j.jhep.2023.06.003.

[339] S. A. Harrison, A. M. Allen, J. Dubourg, M. Noureddin, and N. Alkhouri, "Challenges and opportunities in NASH drug development," *Nature Medicine 2023 29:3*, vol. 29, no. 3, pp. 562–573, Mar. 2023, doi: 10.1038/s41591-023-02242-6.

[340] E. E. Powell, V. W. S. Wong, and M. Rinella, "Non-alcoholic fatty liver disease," *Lancet*, vol. 397, no. 10290, pp. 2212–2224, Jun. 2021, doi: 10.1016/S0140-6736(20)32511-3.

[341] L. F. Laurindo *et al.*, "GLP-1a: Going beyond Traditional Use," *International Journal of Molecular Sciences 2022, Vol. 23, Page 739*, vol. 23, no. 2, p. 739, Jan. 2022, doi: 10.3390/IJMS23020739.

[342] S. A. Harrison *et al.*, "A Phase 3, Randomized, Controlled Trial of Resmetirom in NASH with Liver Fibrosis," *N Engl J Med*, vol. 390, no. 6, pp. 497–509, Feb. 2024, doi: 10.1056/NEJMOA2309000.

[343] H. G. Holzhütter and N. Berndt, "Computational Hypothesis: How Intra-Hepatic Functional Heterogeneity May Influence the Cascading Progression of Free Fatty Acid-Induced Non-Alcoholic Fatty Liver Disease (NAFLD)," *Cells*, vol. 10, no. 3, pp. 1–17, Mar. 2021, doi: 10.3390/CELLS10030578.

[344] L. Niu *et al.*, "Defining NASH from a Multi-Omics Systems Biology Perspective," *J Clin Med*, vol. 10, no. 20, Oct. 2021, doi: 10.3390/JCM10204673.
[345] N. Berndt *et al.*, "HEPATOKIN1 is a biochemistry-based model of liver metabolism for applications in medicine and pharmacology," *Nat Commun*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/S41467-018-04720-9.
[346] H. G. Holzhütter and N. Berndt, "Computational Hypothesis: How Intra-Hepatic Functional Heterogeneity May Influence the Cascading Progression of Free Fatty Acid-Induced Non-Alcoholic Fatty Liver Disease (NAFLD)," *Cells*, vol. 10, no. 3, pp. 1–17, Mar. 2021, doi: 10.3390/CELLS10030578.
[347] W. Dai, Y. Sun, Z. Jiang, K. Du, N. Xia, and G. Zhong, "Key Genes Associated with Non-Alcoholic Fatty Liver Disease and Acute Myocardial Infarction," *Med Sci Monit*, vol. 26, pp. e922492-1, Jun. 2020, doi: 10.12659/MSM.922492.
[348] A. S. Meijnikman *et al.*, "A systems biology approach to study non-alcoholic fatty liver (NAFL) in women with obesity," *iScience*, vol. 25, no. 8, p. 104828, Aug. 2022, doi: 10.1016/J.ISCI.2022.104828.
[349] E. M. Maldonado *et al.*, "Multi-scale, whole-system models of liver metabolic adaptation to fat and sugar in non-alcoholic fatty liver disease," *npj Systems Biology and Applications 2018 4:1*, vol. 4, no. 1, pp. 1–10, Aug. 2018, doi: 10.1038/s41540-018-0070-3.
[350] L. A. D'Alessandro, Klingmüller, and M. Schilling, "Deciphering signal transduction networks in the liver by mechanistic mathematical modelling," *Biochemical Journal*, vol. 479, no. 12, pp. 1361–1374, Jun. 2022, doi: 10.1042/BCJ20210548.
[351] S. M. Keating *et al.*, "SBML Level 3: an extensible format for the exchange and reuse of biological models," *Mol Syst Biol*, vol. 16, no. 8, p. e9110, Aug. 2020, doi: 10.15252/msb.20199110.
[352] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, "OmniPath: guidelines and gateway for literature-curated signaling pathway resources," *Nat Methods*, vol. 13, no. 12, pp. 966–967, Dec. 2016, doi: 10.1038/NMETH.4077.
[353] L. Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez, "Benchmark and integration of resources for the estimation of human transcription factor activities," *Genome Res*, vol. 29, no. 8, pp. 1363–1375, Aug. 2019, doi: 10.1101/GR.240663.118.
[354] S. Müller-Dott *et al.*, "Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities," *Nucleic Acids Res*, vol. 51, no. 20, pp. 10934–10949, Nov. 2023, doi: 10.1093/NAR/GKAD841.
[355] T. Xu *et al.*, "NAFLDkb: A Knowledge Base and Platform for Drug Development against Nonalcoholic Fatty Liver Disease," *J Chem Inf Model*, 2023, doi: 10.1021/ACS.JCIM.3C00395.
[356] S. A. Hoang *et al.*, "Gene Expression Predicts Histological Severity and Reveals Distinct Molecular Profiles of Nonalcoholic Fatty Liver Disease," *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/S41598-019-48746-5.

[357] J. Ren *et al.*, "Targeting CCN2 protects against progressive non-alcoholic steatohepatitis in a preclinical model induced by high-fat feeding and type 2 diabetes," *J Cell Commun Signal*, vol. 16, no. 3, pp. 447–460, Sep. 2022, doi: 10.1007/S12079-022-00667-1/FIGURES/9.

[358] H. Karimkhanloo *et al.*, "Mouse strain-dependent variation in metabolic associated fatty liver disease (MAFLD): a comprehensive resource tool for pre-clinical studies," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–14, Mar. 2023, doi: 10.1038/s41598-023-32037-1.

[359] X. Li, R. Chen, S. Kemper, and D. R. Brigstock, "Production, Exacerbating Effect, and EV-Mediated Transcription of Hepatic CCN2 in NASH: Implications for Diagnosis and Therapy of NASH Fibrosis," *Int J Mol Sci*, vol. 24, no. 16, Aug. 2023, doi: 10.3390/IJMS241612823/S1.

[360] M. Sørensen, K. S. Mikkelsen, K. Frisch, G. E. Villadsen, and S. Keiding, "Regional metabolic liver function measured in patients with cirrhosis by 2-[^{18}F]fluoro-2-deoxy-D-galactose PET/CT," *J Hepatol*, vol. 58, no. 6, pp. 1119–1124, Jun. 2013, doi: 10.1016/J.JHEP.2013.01.012.

[361] H. Wang, M. Feng, K. A. Frey, R. K. Ten Haken, T. S. Lawrence, and Y. Cao, "Predictive models for regional hepatic function based on 99mTc-IDA SPECT and local radiation dose for physiologic adaptive radiation therapy," *Int J Radiat Oncol Biol Phys*, vol. 86, no. 5, pp. 1000–1006, Aug. 2013, doi: 10.1016/J.IJROBP.2013.04.007.

[362] S. Bonekamp *et al.*, "Spatial Distribution of MRI-Determined Hepatic Proton Density Fat Fraction in Adults with Nonalcoholic Fatty Liver Disease," *J Magn Reson Imaging*, vol. 39, no. 6, p. 1525, 2014, doi: 10.1002/JMRI.24321.

[363] K. Ščupáková *et al.*, "Spatial Systems Lipidomics Reveals Nonalcoholic Fatty Liver Disease Heterogeneity," *Anal Chem*, vol. 90, no. 8, pp. 5130–5138, Apr. 2018, doi: 10.1021/ACS.ANALCHEM.7B05215/SUPPL_FILE/AC7B05215_SI_001.XLSX.

[364] H. G. Holzhütter and N. Berndt, "Computational Hypothesis: How Intra-Hepatic Functional Heterogeneity May Influence the Cascading Progression of Free Fatty Acid-Induced Non-Alcoholic Fatty Liver Disease (NAFLD)," *Cells 2021, Vol. 10, Page 578*, vol. 10, no. 3, p. 578, Mar. 2021, doi: 10.3390/CELLS10030578.

[365] T. J. Velenosi *et al.*, "Postprandial Plasma Lipidomics Reveal Specific Alteration of Hepatic-derived Diacylglycerols in Nonalcoholic Fatty Liver Disease," *Gastroenterology*, vol. 162, no. 7, pp. 1990–2003, Jun. 2022, doi: 10.1053/J.GASTRO.2022.03.004.

[366] A. Bayat, "Bioinformatics," *BMJ*, vol. 324, no. 7344, pp. 1018–1022, Apr. 2002, doi: 10.1136/BMJ.324.7344.1018.

[367] R. Yue and A. Dutta, "Computational systems biology in disease modeling and control, review and perspectives," *npj Systems Biology and Applications 2022 8:1*, vol. 8, no. 1, pp. 1–16, Oct. 2022, doi: 10.1038/s41540-022-00247-4.

[368] F. Meng and T. Ellis, "The second decade of synthetic biology: 2010–2020," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–4, Oct. 2020, doi: 10.1038/s41467-020-19092-2.

Appendix A

Publications

Articles published/accepted in peer-reviewed journals with first-authorship

Hoch M, Smita S, Cesnulevicius K, *et al*. Network- and enrichment-based inference of phenotypes and targets from large-scale Disease Maps. *npj Syst Biol Appl.* 2022;8(1):13. doi:10.1038/s41540-022-00222-z

> **Contributions**: I developed the methodology, performed data processing and analysis for the case studies, and developed web-based Disease Map tools that employ the described approach. I prepared the first draft of the manuscript.

Hoch M*, Ehlers L*, Bannert K, *et al*. In silico investigation of molecular networks linking gastrointestinal diseases, malnutrition, and sarcopenia. *Front Nutr.* 2022;9. doi:10.3389/fnut.2022.989453

> **Contributions:** I reviewed the Submaps, published the Sarcopenia Map, designed the model, developed the Disease Map tools, and performed the analyses. I wrote significant portions of the manuscript.

* shared first authorship

Hoch M, Rauthe J, Cesnulevicius K, *et al*. Cell-Type-Specific Gene Regulatory Networks of Pro-Inflammatory and Pro-Resolving Lipid Mediator Biosynthesis in the Immune System. *Int J Mol Sci.* 2023;24(5):4342. doi:10.3390/IJMS24054342/S1

> **Contributions:** The concept of the study was developed by me. I processed knowledge graphs extracted from the AIR and developed the methodology. I integrated and processed single-cell RNA-Seq data, performed the data analysis, and interpreted the results. I prepared the first draft of the manuscript.

Hoch M, Smita S, Cesnulevicius K, *et al*. Network analyses reveal new insights into the effect of multi-component Tr14 compared to single-component diclofenac in an acute inflammation model. *J Inflamm.* 2023;20(1):1-15. doi:10.1186/S12950-023-00335-0

Contributions: The concept of the study was developed by Shailendra Gupta and myself. I processed the RNA-Seq data and performed the data analysis employing the AIR and Disease Map tools developed in the previous studies. I prepared the first draft of the manuscript.

Hoch M, Olaf Wolkenhauer, and Shailendra Gupta. Large-Scale Knowledge Graph Representations of Disease Processes. *Curr Opin Syst Biol*. 2024;38:100517. doi:10.1016/J.COISB.2024.100517

Contributions: I collected the reviewed literature, provided the structure of the manuscript, and prepared the first draft of the manuscript.

Articles published/accepted in peer-reviewed journals with co-authorship.

Hausburg F, …, **Hoch M**, *et al*. (Re-)programming of subtype specific cardiomyocytes. *Adv Drug Deliv Rev.* 2017;120:142-167. doi:10.1016/j.addr.2017.09.005

Contributions: The concept of the study was developed by Robert David and Markus Wolfien. I supported the study by analyzing RNA-Seq data using knowledge graphs from public databases and enrichment tools in CytoScape.

Serhan CN, Gupta SK, …, **Hoch M**, *et al*. The Atlas of Inflammation Resolution (AIR). *Mol Aspects Med.* 2020;74:100894. doi:10.1016/j.mam.2020.100894

Contributions: The concept of the study was developed by Charles Serhan and Shailendra Gupta. I curated several knowledge graphs on molecular and cellular mechanisms in the immune response and supported the creation of graphical images. I was responsible for compiling the information and publishing the Disease Map as a web-based resource.

Ostaszewski M, …, **Hoch M**, *et al.* COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms. *Mol Syst Biol*. 2021;17(10):e10387. doi:10.15252/msb.202110387

Contributions: The concept of the study was developed by Marek Ostaszewski and the Disease Maps Community. I supported the study by curating knowledge graphs on SARS-CoV-2 interactions in the innate immune response.

Van Welzen A, **Hoch M**, *et al.* The Response and Tolerability of a Novel Cold Atmospheric Plasma Wound Dressing for the Healing of Split Skin Graft Donor Sites: A Controlled Pilot Study. *Skin Pharmacol. Physiol*. 2021;34(6):328. doi:10.1159/000517524

Contributions: The concept of the study was developed by Annika van Welzen and Alexander Thiem. I supported the study with scripts for data management and data processing. I also supported the statistical analysis.

Scheel J, **Hoch M**, *et al.* NaviCenta – The Disease Map for placental research. *Placenta*. 2023;143:12-15. doi:10.1016/j.placenta.2023.09.007

Contributions: The concept of the study was developed by Julia Scheel. I supported the study through the development of Disease Map tools and their integration into the NaviCenta Disease Map presented in the study.

Book Chapters

Müller R, **Matti H**, *et al*. Inflammation Resolution Mediators: Future Prospects. To be published in: Inflammation Resolution and Chronic Diseases by Tripathi A, Dwivedi A, Gupta S, and Poojan S (ed.) *Springer Nature Singapore*. 2024

Cavallo C, **Matti H**, *et al*. Are Multi-Component Drugs Better in Resolving the Inflammation Compared to the Single-Component Drugs? To be published in: Inflammation Resolution and Chronic Diseases by Tripathi A, Dwivedi A, Gupta S, and Poojan S (ed.) *Springer Nature Singapore*. 2024

Müller R, ..., **Matti H**, *et al*. Mechanistic Understanding of Inflammation Resolution Using the Atlas of Inflammation Resolution (AIR). To be published in: Inflammation Resolution and Chronic Diseases by Tripathi A, Dwivedi A, Gupta S, and Poojan S (ed.) *Springer Nature Singapore*. 2024

Articles published in non-peer-reviewed journals with co-authorship.

Rasche H, **Hoch M**, *et al.* Reproducible Exploration of Disease Maps with Galaxy Workflows and the MINERVA Platform. *Preprints*. 2024. doi:10.20944/preprints202403.1211.v1

Published articles in conference proceedings

Schopohl P, …, **Hoch M**, *et al.* A systems approach to investigate inflammation resolution by multi-component medicinal product TR14. *Ann Rheum Dis*. 2019;78:1496. doi:10.1136/annrheumdis-2019-eular.5084

Genovese F, …, **Hoch M**, *et al*. Computational molecular interaction maps of signaling events within the olfactory epithelium. *Chemical Senses*. 2022;47. doi:10.1093/chemse/bjac031

Appendix B

Curriculum Vitae

Education

Ph.D. Research Program
Department of Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Germany — 2019 – 2024

Master of Science, Medical Biotechnology
Rostock University Medical Center, Rostock, Germany — 2017 - 2019
Thesis
Title: "Curation of an immune cell interactome and its analysis"
Department of Systems Biology and Bioinformatics, University of Rostock
Supervisors: Prof. Olaf Wolkenhauer and Dr. Suchi Smita
- Creating immune cell type-specific interactomes by integrating single-cell RNA-Seq Data into a reference interactome and evaluation of inter- and intra-cellular signaling

Thesis Grade: 1.3; Final Grade: 1.4

Bachelor of Science, Medical Biotechnology
Rostock University Medical Center, Rostock, Germany — 2014 - 2017
Thesis:
Title: "Identification of Cardiac Stem Cell Types using Network Analysis Approaches"
Department of Systems Biology and Bioinformatics, University of Rostock
Supervisors: Prof. Olaf Wolkenhauer and Dr. Markus Wolfien
- Integration of RNA-Seq data from cardiomyocytes in a molecular interaction network curated from public databases to identify key pathways and enriched biological processes

Thesis Grade: 1.3; Final Grade: 1.5

Bachelor of Science, Biomedical Technology
University of Rostock, Rostock, Germany — 2013 – 2014

Experience

Research Associate
Department of Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Germany — 2019 – To date
Topics:
- Curating and publishing the "Atlas of Inflammation Resolution" (PMID:32893032, >110 citations), "Sarcopenia Map" (PMID:36407505), and MASDL Map
- Developing tools to integrate and analyze heterogenous biomolecular data on Disease Maps
- Analyzing RNA-Seq data and drug interactome data through network approaches

Research Assistant
Department of Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Germany 2017 – 2019
Topics:

- Curation of molecular interaction networks
- Designing the "Atlas of Inflammation Resolution"

Internship
Reference und Translation Center for Cardiac Stem Cell Therapy (RTC), Rostock University Medical Center Juli - September 2016
Topics:

- Characterization and microRNA-based Cardiac Reprogramming of Adipose-derived Stem Cells
- Culturing and miRNA Transfection of ADSCs and histology of derived cardiac pacemaker cells

Collaborations

Initiation, realization & successful completion of interdisciplinary collaboration with clinical & industry partners:

- Traumeel Inflammation Resolution in-silico Research Program (TIRIP), 2020-2022, in collaboration with Dr. Myron Schultz, Dr. Konstantin and Dr. David Lescheid from Heel GmbH, Baden-Baden, Germany
  [PMID: 32893032, PMID: 35473910, PMID: 36901771, PMID: 36973809]
    - Creating the "Atlas of Inflammation Resolution", an interactive online resource on molecular and cellular processes in inflammatory processes and developing tools for data integration, visualization and analysis
    - Supervising two student assistant researchers and the curation efforts of the AIR
    - Supervising the associated bachelor thesis of Jannik Rauthe "Kinetic Modeling of the Lipid Mediator Class Switch in the Innate Immune Response"
    - 6 Oral Presentations and 9 Poster Presentations at 10 Conferences
- Collaboration with Prof. Robert Jaster from the Division of Gastroenterology at the Rostock University Medical Center, Rostock, Germany, 2021 - 2022
  [PMID: 36407505]
    - Developing of the Sarcopenia Map, a state-of-the art online resource for in silico simulations of molecular processes connecting nutrition, gastrointestinal diseases and sarcopenia
    - Supervising two student assistant researchers
    - 3 Oral Presentations and 2 Poster Presentations at 4 Conferences
- Collaboration with Prof. Matthias Löhr from Karolinska Institute, Stockholm, Sweden 2022 - to date
    - Employing network approaches to connect the endocrine and exocrine pancreas with brain function

    - Supervising the associated bachelor thesis of Ronja Müller "A Network-Based Model to Investigate the Pancreas-Brain-Axis"
- Non-alcoholic fatty liver disease In-silico Research Project (NIRP), 2023 - to date, in collaboration with Dr. Myron Schultz, Dr. Konstantin and Dr. David Lescheid from Heel GmbH, Baden-Baden, Germany
    - Developing a large-scale, multi-component model of MASLD to investigate the disease pathology and predict effects of diet and drug interventions
    - Designing and communicating the project with the clinical partners Prof. Andreas Geier, Prof. Jörn Schattenberg, and Prof. Ali Canbay
    - Supervising the associated master thesis of Jannik Rauthe "Investigating the pathogenesis of non-alcoholic fatty liver disease using a multi-compartmental network model"
    - 3 Oral Presentations and 2 Poster Presentations at 3 Conferences

Teaching Experience

Lecture "BioSystems Modelling and Simulation"
University of Rostock 2019 – To date
Teaching Biochemistry, Medical Biotechnology, and Computer Science Students on creation of Disease Maps and ordinary differential equation models of elementary and enzymatic reactions

Course "in silico modelling of molecular networks"
Rostock University Medical Center 2020 - To date
Teaching Medical Biotechnology students the basics of curating and analysis molecular pathways in standardized network representations

Course "Neueste Entwicklungen in der Informatik"
University of Rostock 2023 - To date
Practical course teaching Computer Science students to develop a webserver application in Python, JavaScript and HTML

Leadership

Supervised Bachelor Theses

Jannik Rauthe, 2021
Topic: "Kinetic Modeling of the Lipid Mediator Class Switch in the Innate Immune Response"

Ronja Müller, 2023
Topic "A Network-Based Model to Investigate the Pancreas-Brain-Axis"

Supervised Master Theses

Mahla Haghzad, 2023
Topic: "Predicting effective drug combinations using machine learning"

Jannik Rauthe, 2023
Topic: "Investigating the pathogenesis of non-alcoholic fatty liver disease using a multi-compartmental network model"

Supervised Student Asistant Researchers

Dragana Gjorgevikj
Juliane Proksch
Christina Stanke
Vanessa Caton
Jannik Rauthe
Wilhelm Sponholz
Ronja Müller

Extracurricular Activities

Member of the Council of the Rostock Medical Faculty 2017 – 2019
Member of the Student Council of the University of Rostock 2017 – 2019
Member of the Council of the University of Rostock 2017 – 2018
Member of the Student Council of the Rostock Medical Faculty 2015 – 2019

Conferences Attendance

Event Name	Date	Location	Presentation
Disease Maps Community Meeting 2019	2-4 October 2019	Sevilla, Spain	Poster
INCOME2019	25-29 November 2019	Berlin, Germany	Poster
e:Med Meeting 2020	8-10 March 2020	Online	
EUvsVirus Hackathon	24-26 April 2020	Online	
Biohackathon EU 2020	9-13 November 2020	Online	
Disease Maps Community Meeting 2020	12-14 November 2020	Online	Talk
INCOME 2021	1-4 March 2021	Online	
Systems biology of Human Diseases 2021	5-7 July 2021	Berlin, Germany	Poster
e:Med Meeting 2021	20-22 September 2021	Online	
European Congress of Immunology 2021	1-4 September 2021	Online	Talk & Poster
Disease Maps Community Meeting 2021	29-30 November 2021	Online	Talk & 2 Posters
ASPEN2022	25-27 March 2022	Online	Talk
World Congress on Inflammation 2022	05-08 June 2022	Rome, Italy	Talk & Poster
Viszeralmedizin 2022	17-19 September 2022	Hamburg, Germany	Talk
e:Med Meeting 2022	28-30 November 2022	Heidelberg, Germany	Talk & poster

Event Name	Date	Location	Presentation
Spring School From Omics to Systems Biology	6-10 March 2023	Potsdam, Germany	Talk & 2 Posters
Disease Maps Community Meeting 2023	3-5 April, 2023	Maastricht, Netherlands	Talk & poster
GMDS 2023	17-21 September 2023	Heilbronn, Germany	2 Workshops
e:Med Meeting 2023	09-11 October 2023	Berlin, Germany	Talk
Biohackathon Europe 2023	30 October-02 November 2023	Barcelona, Spain	-
Keystone Symposium "MASH & Fibrosis"	03-07 March 2024	Banff, Canada	Talk & Poster
Disease Maps Community Meeting 2024	25-27 March 2024	Luxembourg	Talk & Poster

Organization of Scientific Meetings:
6th Disease Maps Community Meeting (DMCM 2021) November 29-30, 2021 (https://disease-maps.org/DMCM2021)

Appendix C

Pseudocode

Algorithm 1: Calculation of topological weightings between nodes and phenotypes for the two-dimensional enrichment analysis (2DEA).

```
Result: Calculate and return topological weightings for each node in a graph
Function CalculateWeightings(self, graphAdjList):
    nodeWeightings ← {}
    setOfAllPaths ← {}
    nodesOnEachPath ← defaultdict(set)
    countOfPathsViaNode ← defaultdict(int)
    isNodeVisited ← defaultdict(bool)

    Function DepthFirstSearch(currentNode, currentPath):
        nodesOnEachPath[currentNode].update(currentPath)
        if isNodeVisited[currentNode] then
            return countOfPathsViaNode[currentNode]
        end
        if currentNode not in graphAdjList or not graphAdjList[currentNode] then
            countOfPathsViaNode[currentNode] ← countOfPathsViaNode[currentNode] + 1
            return countOfPathsViaNode[currentNode]
        end
        if currentNode ∈ currentPath then
            // Detect cycle
        end
        return countOfPathsViaNode[currentNode]
        isNodeVisited[currentNode] ← true
        for (neighborNode, skippedNodes) ∈ graphAdjList[currentNode] do
            if skippedNodes then
                for skippedNode ∈ skippedNodes do
                    nodesOnEachPath[skippedNode].update(currentPath)
                    pathsSkippedViaNeighbor ← DepthFirstSearch(neighborNode, currentPath +
                     [skippedNode, currentNode])
                    countOfPathsViaNode[currentNode] ← countOfPathsViaNode[currentNode] +
                     pathsSkippedViaNeighbor
                    countOfPathsViaNode[skippedNode] ← countOfPathsViaNode[skippedNode] +
                     pathsSkippedViaNeighbor
                end
            end
            else
                countOfPathsViaNode[currentNode] ← countOfPathsViaNode[currentNode] +
                 DepthFirstSearch(neighborNode, currentPath + [currentNode])
            end
        end
        return countOfPathsViaNode[currentNode]
    return
    totalPathsInGraph ← DepthFirstSearch(self, [])
    if totalPathsInGraph = 0 then
        return {}
    end
    nodesOnEachPath ← {node: len(nodes) for node, nodes in nodesOnEachPath.items() if
     nodes and node ≠ self}
    totalUniqueNodesOnPaths ← len(nodesOnEachPath)
    for (node, countOnPaths) in nodesOnEachPath.items() do
        nodeWeightings[node] ← (countOfPathsViaNode[node] / totalPathsInGraph) +
         (countOnPaths / totalUniqueNodesOnPaths)
    end
    if nodeWeightings then
        maxScore ← max(nodeWeightings.values())
        if maxScore then
            return {node: (score × optNodeWeights.get(node, 1) / maxScore) for node, score
             in nodeWeightings.items()}
        end
        else
            return {}
        end
    end
    else
        return nodeWeightings
    end
return
```

Algorithm 2: Pseudocode of an adapted form of Dijkstra's algorithm to identify the widest path in a directed graph with weighted edges.

```
Result: Find the widest path between a source and a target in a graph and reconstruct the path
Input : Graph, a list of lists where each list contains tuples (weight, node) of outgoing edges, source node 'src',
        target node 'target'
Function WidestPathProblem(Graph, src, target):
    n ← length(Graph)
    widest ← [ −∞, −∞, ..., −∞ ] (length n)
    parent ← [0, 0, ..., 0] (length n)
    container ← []
    append(container, (0, src))
    widest[src] ← ∞
    Sort(container)
    while container is not empty do
        temp ← last element of container
        current_src ← temp[1]
        Remove last element from container
        for vertex in Graph[current_src] do
            distance ← max(widest[vertex[1]], min(widest[current_src], vertex[0]))
            if distance > widest[vertex[1]] then
                widest[vertex[1]] ← distance
                parent[vertex[1]] ← current_src
                append(container, (distance, vertex[1]))
                Sort(container)
            end
        end
    end
    path ← GetPath(parent, target, target)
    widest_distance ← widest[target]
    return [path, widest_distance]
return
Function GetPath(parent, vertex, target):
    if vertex = 0 then
        return []
        // Return an empty path if vertex is zero
    end
    path ← [vertex] + GetPath(parent, parent[vertex], target)
    // Recursively call GetPath
    return path
    // Return the constructed path
return
```

Algorithm 3: Pseudocode for multi-compartmental Boolean modeling, updating only the states of the nodes whose input nodes have changed in the previous step, and storing only the values of the compartments in which a change occurred in sparse arrays.

```
Result: Update the state of nodes in a simulation
Function ActivityStep(first_step = False):
    if first_step then
        nodes_to_eval ← nodes
    else
        nodes_to_eval ← {node ∈ nodes | node.always_update}
        for node ∈ self.nodes_with_changes ∪ {node ∈ nodes | node.delay ∨ node.refill ≠
         None} do
            nodes_to_eval.update(node.outgoing_targets)
        end
    end

    nodes_to_clear_cache ← nodes_with_changes ∪ {node ∈ nodes | node.always_update}
    foreach node ∈ nodes_to_clear_cache do
        node.active.cache_clear()
    end

    nodes_with_changes ← ∅
    sparse_row_indices ← [ ]
    sparse_col_indices ← [ ]
    sparse_data_indices ← [ ]
    current_idx ← 0

    foreach node ∈ nodes_to_eval do
        previous_activity, new_activity ← node.update_activity()
        diff ← new_activity - previous_activity
        non_zero_positions ← np.where(diff ≠ 0)[0]

        if len(non_zero_positions) > 0 then
            nodes_with_changes.add(node)
            current_activities[node.index] ← new_activity
            non_zero_values ← diff[non_zero_positions]
            num_updates ← len(non_zero_positions)

            sparse_row_indices[current_idx :  current_idx + num_updates] ← node.index
            sparse_col_indices[current_idx :  current_idx + num_updates] ←
             non_zero_positions
            sparse_data[current_idx :  current_idx + num_updates] ← non_zero_values
            current_idx ← current_idx + num_updates
        end
    end

    row_indices ← sparse_row_indices[:current_idx]
    col_indices ← sparse_col_indices[:current_idx]
    data ← self.sparse_data[:current_idx]

    sparse_activity_matrix ← coo_matrix((data, (row_indices, col_indices)),
     shape=(len(self.nodes), self.grid_size), dtype=np.int32)
    stored_activities.append(sparse_activity_matrix)
```

Declaration of Authorship

I hereby declare that this thesis was independently composed and authored by myself.

All content and ideas drawn directly or indirectly from external sources are indicated as such. All sources and materials that have been used are referred to in this thesis.

The thesis has not been submitted to any other examining body and has not been published.

Signed: Matti van Welzen
Date: 31.10.2024
Place: Rostock