Thesis for The Degree of Licentiate of Engineering

Machine Learning Enabled Functional Discovery in Yeast Systems Biology

Daniel Brunnsåker

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Machine Learning Enabled Functional Discovery in Yeast Systems Biology

Daniel Brunnsåker

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

To my friends and family

Machine Learning Enabled Functional Discovery in Yeast Systems Biology

Daniel Brunnsåker

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Saccharomyces cerevisiae is a well-studied organism, yet roughly 20 percent of its proteins remain poorly characterized. Recent studies also seem to indicate that the pace of functional discovery is slow. Previous work has implied that the most probable path forward is via not only regular automation but fully autonomous systems that can automatically guide and perform high-throughput experimentation.

This thesis explores various concepts to accelerate and perform functional discovery of gene and protein functions in *Saccharomyces cerevisiae*. It does so by combining ideas from artificial intelligence, such as active learning, with high-throughput analytical techniques like mass-spectrometry. The work performed as the basis for this thesis also served to aid in the further characterization of different aspects of yeast systems biology. Specifically, it delved into the diauxic shift and its regulators through the lens of untargeted metabolomics, as well as the regulatory patterns behind genome-wide intracellular proteomic abundances.

We find that it is essential not only to develop tools and techniques for facilitating high-throughput experimentation, but also to ensure their optimal utilization of already existing knowledge. It is also of paramount importance to ensure a holistic and encompassing view of systems biology by more fully integrating and using different levels of cellular organization and analytical techniques.

Keywords

Metabolomics, Proteomics, Mass Spectrometry, Machine Learning, Inductive Logic Programming, Metabolism, Systems Biology, Metabolic Modelling

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **D. Brunnsåker**, G.K. Reder, N.K. Soni, O.I. Savolainen, A.H. Gower, I.A. Tiukova & R.D. King, *High-throughput metabolomics for the design and validation of a diauxic shift model.*
npj Systems Biology and Applications, 9, Article number 11 (2023).

[**Paper II**] **D. Brunnsåker**, F. Kronström, I.A. Tiukova & R.D. King, *Interpreting protein abundance in Saccharomyces cerevisiae through relational learning.*
Submitted, under review.

Acknowledgment

First of all, I would like to thank my supervisors Ross and Ievgeniia. Thank you for giving me this opportunity. Thank you for your support, guidance and allowing me such a high degree of freedom. I also want to thank my examiner, Graham, for being extremely helpful in all situations, but also for tolerating my endless stream of questions.

I have been fortunate to work alongside exceptional colleagues in the King lab. I couldn't have imagined more pleasant people to collaborate with on a daily basis. Alexander, Gabriel, Filip, Beera and Erik, thanks for making (almost) every day of my PhD-journey amazing.

A big shout-out to the Data Science and AI division and all of its members for welcoming me with open arms. I felt at home right away. I especially want to thank all of the PhD-students at DSAI, thanks for answering my many questions and being a pleasure to hang out with!

Thanks to all of my friends (with some of you hitting almost every category of this acknowledgement!). I cannot emphasise enough how much you all mean to me. I am truly lucky to have gotten to know such amazing people.

To my family, thank you for your never-ending support, no matter the situation. I wouldn't even have thought of an undertaking of this magnitude if it wasn't for all of you.

Lastly, I would like to thank my partner Francine. Thank you for being my best friend and biggest supporter during these past few years. Thank you for all of the adventures, and hopefully the many more to come!

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Systems biology has been responsible for several recent significant developments in human health and environmental sustainability, paving the way for more complete understanding of complex biological systems. It makes use of computational and mathematical analysis to decipher biological systems. The field integrates many scientific disciplines—biology, computer science, physics and others—to predict how these systems change over time and under varying conditions[1].

Systems biology presents a challenge to the regular—human-based—scientific method[2, 3]. Commonly, scientific discovery in systems biology follows an iterative cycle (see Figure 1.1). However, the systems of interest are highly complex; even "simple" examples, such as *Escherichia coli* and *Saccharomyces cerevisiae* have thousands of genes, proteins and other small molecules interacting with each other in intricate spatial and temporal ways[4, 5]. This complexity implies the need for millions—if not billions—of guided experiments (and accompanying analysis). These experiments should rationally improve and build upon the knowledge we have already obtained but also fit within the frameworks of human understanding [2, 6]. In order to handle the complexity and sheer amount of data produced in biological studies, we need to be able to employ superhuman capabilities: Firstly, we can enhance throughput by utilizing increased degrees of lab-automation. Secondly, we can incorporate ideas and concepts from artificial intelligence (AI) to aid in analysis and interpretation. These two approaches can, in turn, help us integrate more types of data, experimental readouts and different levels of biological organization in a faster and more efficient manner.

The research conducted as the foundation for this thesis serves a dual purpose; we aim to develop methods that build upon our knowledge about biological systems in a high-throughput manner, while also maximizing the use of our already systematized knowledge. The biological system of choice is *Saccharomyces cerevisiae*—commonly known as baker's yeast—a single-celled eukaryotic model organism. It is commonly used to understand genetics, metabolism, and other fundamental cellular functions, and currently stands as the most extensively studied eukaryote[7].

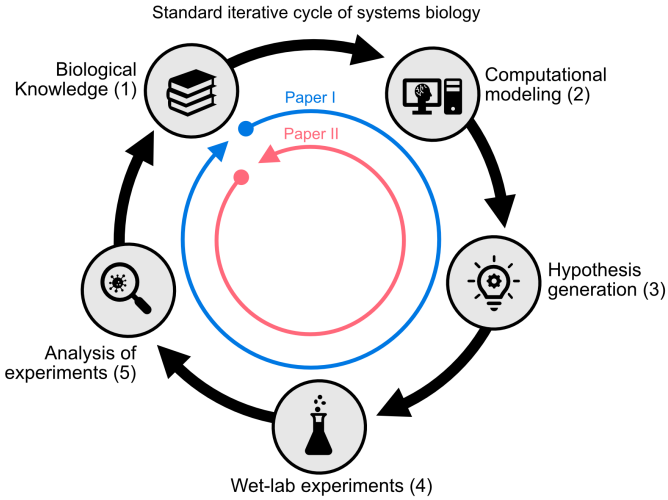

Figure 1.1: The classic iterative cycle of systems biology: (1) utilize existing knowledge, (2) mathematically represent and model it, (3) generate testable hypotheses, (4) test the hypotheses experimentally, (5) analyze the outcomes, and integrate new findings into the knowledge base. The blue arrow represents the methodology used in **Paper I**, closely resembling the classic cycle. The red arrow represents the methodology used in **Paper II**, instead evaluating systematized knowledge and generating hypotheses based on experimental data to uncover new regulatory rules

In the first paper of this thesis, the objective was to design and explore semi-automated high-throughput pipelines for functional genomics. This was achieved through computationally aided experimental design coupled with analytical methods and experimental setups highly amenable to automation. The goal was to enable future acceleration of the iterative cycle and further characterize the biological system itself. More specifically, we studied untargeted metabolomics (via mass spectrometry) for functional discovery of genes involved in the regulation of a complex dynamic transformation known as the diauxic shift.

In the second manuscript, we deviate from the classical methodology. We instead aggregate decades worth of highly structured knowledge on yeast systems biology, derived from millions of experiments. This knowledge is then combined with methodologies from relational learning and explainable AI. This approach enabled us to discover human-interpretable regulatory rules and genotype-phenotype relations in *S. cerevisiae*. These rules are subsequently evaluated through the use of supervised learning in conjunction with intracellular proteomic and metabolomic abundances. This enables us to infer patterns behind protein levels, and also establish connections between various levels of biological organization in a tangible and unified way, enabling the simultaneous use of highly heterogeneous data.

Chapter 2

Background

2.1 Systems Biology

Systems biology is a multidisciplinary approach meant to aid in the understanding of complex biological systems at the molecular, cellular, and organismal levels. This field has emerged as a competing concept to reductionist biology, driven in part by the necessity to integrate data from diverse sources and levels of cellular organization. The goal is to forge a comprehensive and holistic understanding of biological systems[8]. Systems biology aims to build models that can capture the behavior of biological systems and predict their responses to perturbations across a wide variety of conditions[9]. The approach has revolutionized our understanding of biology and accelerated the development of new biotechnologies[10]. Moreover, systems biology approaches are essential for addressing some of the most pressing challenges in biology today, such as understanding the mechanisms of aging and devising strategies to combat cancer.

2.1.1 *Saccharomyces cerevisiae*

Yeast systems biology is a subfield of systems biology that mainly focuses on the study of the baker's yeast—*Saccharomyces cerevisiae*—as a model organism for understanding complex biological systems. This yeast is a unicellular eukaryote which has been an essential part of human civilization for thousands of years through its use in food and beverage fermentation[11]. The ease of cultivation and overall resilience of *S. cerevisiae*, combined with the fact that it shares many fundamental biological processes with higher organisms has caused it to be an organism of high interest to the scientific community. Moreover, its biology makes it well-suited for genetic modification through a wide array of powerful genetic and molecular tools, such as homologous recombination[12, 13]. This has made it an ideal model system for systems biology research, and ultimately caused it to be the first eukaryote to have its genome sequenced in 1996[4, 14].

As a result, it has been the premier platform for the functional discovery

of genes in eukaryotes. Early efforts were focused on creating genome-wide yeast deletion mutant collections[12, 15]. These collections are comprised of large libraries of cells which had undergone processes to separately remove or alter most of the identified coding gene sequences in *S. cerevisiae*. This allowed researchers to thoroughly explore the genome through experimental means, one gene at a time.

2.1.2 Functional Discovery in Yeast Systems Biology

Functional discovery—or functional genomics, depending on the context—refers to the process of identifying and characterizing the function of specific biological molecules or subsystems, such as genes, proteins and metabolic pathways (a metabolic pathway is a series of interconnected biochemical reactions that converts molecules to other usable products, further explained in Section 2.1.3.1)[16]. This is a critical area of research, as it allows researchers the tools and know-how to better understand the fundamental processes that govern life, such as gene regulation and metabolism. These insights could, in turn, provide understanding in related domains, such as mechanisms of disease and their potential therapies.

Biological systems are extremely complex, and each molecule often interacts with many others in a multitude of different ways. Additionally, their behaviour can change dramatically due to temporal, environmental or conditional aspects. Even the slightest alteration can result in a wildly different phenotype (observable characteristics or traits). As such, functional discovery in this context is an inherently iterative process. Generally, it is done through a combination of experimental and computational processes, and tends to make use of large-scale experimental data at its core (see Figure 1.1). The types of data acquired may vary depending on the intent of the study, but could involve techniques such as transcriptomics, metabolomics and proteomics (see Section 2.3). Experiments tend to be carefully designed as to provide as much information on the target phenomena as possible[17]. These designs could include contrasting environmental conditions, changed genetic backgrounds or perturbations in the form of therapeutic treatments or nutrient limitations.

2.1.3 Knowledge Representation

As the research field seeks to understand the behaviour and function of biological components as part of a larger system, one of the key challenges is then how to properly represent and organize the large amounts of information that is available about these systems. These representations of reality may take many different forms, such as metabolic pathways and interaction networks, or even structured representations such as ontologies. Typically these are used in conjunction with mathematical modeling techniques, such as ODEs (ordinary differential equations), constraint-based approaches, graphical models or rule-based models. All of these are designed to capture different types of biological relationships. The work conducted in this thesis employs several different representations in conjunction with each other. These will be explained in

further detail below.

2.1.3.1 Models of Metabolism

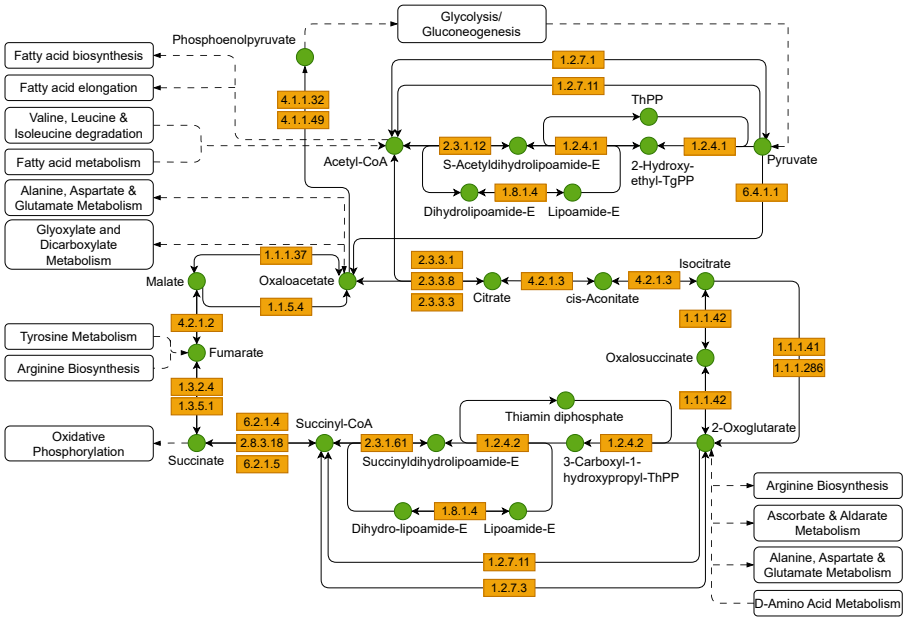

Figure 2.1: Pathway representation (as a directed graph) of the citric acid cycle in *Saccharomyces cerevisiae*. Green circles mark involved metabolites, orange squares represent reactions (via the involved enzyme) and the white squares denote interconnected pathways. Pathway information taken from KEGG (2023-08-02)[18].

Metabolism refers to a set of biochemical processes that occur within living organisms. It encompasses the reactions that involve the conversion of nutrients into energy, generating the building blocks required for growth, repair and maintenance.

Metabolic networks are a type representation that allow for insight into the molecular mechanisms of metabolism. The models attempt to acquire and represent all of the known metabolic information about a specific metabolic system, such as enzymes, metabolites and their associated reactions. These serve as valuable references for researchers studying metabolism, as these typically provide comprehensive maps and conditional descriptions of the reactions.

Examples of large-scale projects which aggregate different representations of metabolism would be KEGG, Reactome and Biocyc[18–22]. These models are typically subdivided into metabolic pathways—modules which perform some sort of localized task in the cell. Examples could encompass catabolic pathways like glycolysis or anabolic pathways such as amino acid biosynthesis.

This concept is amenable to other forms of representations (which could be used for simulation), as can be seen in Section 2.2.

2.1.3.2 Gene Regulatory Networks

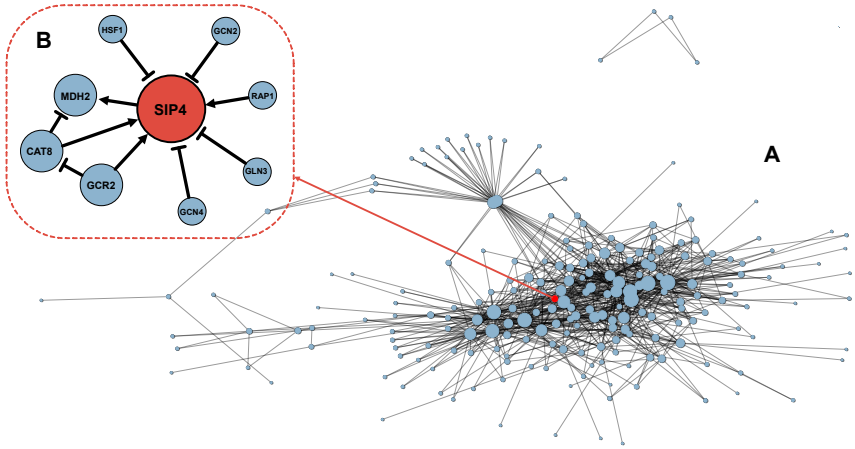

Figure 2.2: Gene regulatory network for the diauxic shift in *Saccharomyces cerevisiae*. **A.** Comprehensive gene regulatory network for the diauxic shift in *Saccharomyces cerevisiae*, derived from the work by Geistlinger *et al.*[23]. **B.** Subgraph of SIP4 and its regulatory interactions. Nodes represent the gene and/or gene products and the edges denote the interaction.

A gene regulatory network is typically depicted as a graphical model which describes the interactions and relationship among gene, gene products and their regulators (see Figure 2.2). It represents an intricate network that governs the expression and activity of genes and proteins within cells. It contains information about a multitude of different types of agents[24]. This usually includes DNA-binding proteins such as transcription factors (TFs) which either promote or inhibit gene expression. However, it is not uncommon to extend the concept to include other forms of activity regulation and signalling elements, such as protein kinases[2]. The interactions between genes and/or proteins can be direct or indirect in these representations, but they tend to reflect a regulatory role, such as repression or activation.

These models can be used to simulate biological systems by integrating various data sources, such as gene expression data (transcriptomics), and utilizing supervised machine learning techniques[2, 25]. This, in turn, could be used to predict system dynamics under different environmental conditions or perturbations

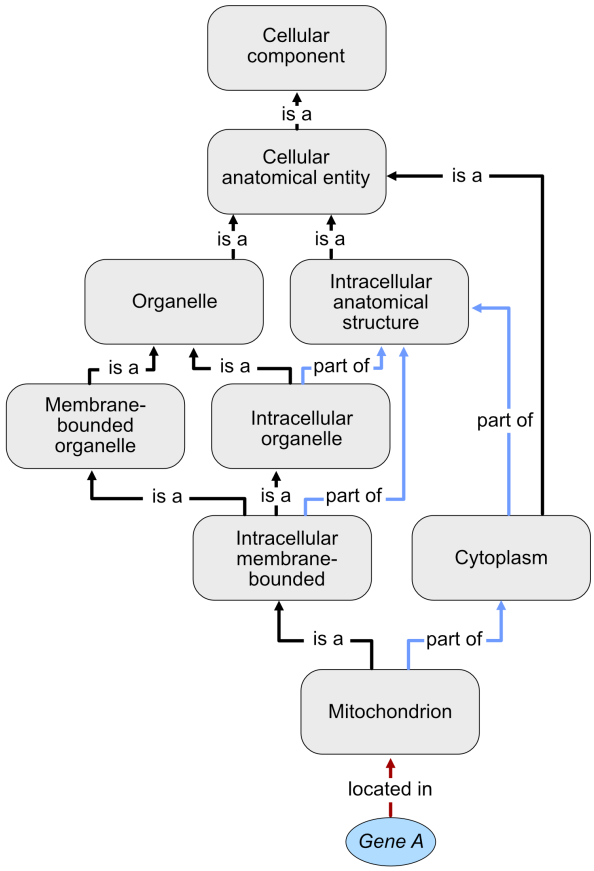

Figure 2.3: Example ancestor chart for the "mitochondrion" gene ontology term (in the cellular compartment category). Each arrow corresponds to a different semantically meaningful relation with the parent term. The hierarchical structure allows for classification of gene function at different levels of specificity.

2.1.3.3 Gene Ontology

Functional knowledge about genes need to be formalized, not only to simplify their study but also to enable higher degrees of semantic interoperability and data integration. There are several methodologies for ontologically representing functional data on genes, with the most commonly used one being the "Gene Ontology".

Gene Ontology (GO) is a widely used resource and provides a standardized vocabulary that enables a structured and controlled representation of biological knowledge related to genes and their functions. It generally classifies or annotates genes based on a few different categories[26]:

1. Molecular function, a category which describes the type of biochemical

activity or intrinsic property of the gene products. This could include concepts such as transport, transcription factor activity or phosphorylation activity.

2. Biological process, a category representing various molecular events and activities within living organisms. It encompasses terms such as "cellular metabolism" or "signal transduction" that describe the biological processes genes are involved in.

3. Cellular compartment, this category describes the locations or structures within a cell or organism where gene products are active or present. It includes terms such as "nucleus" or "mitochondrion," that indicate the sub-cellular locations or compartments associated with specific gene products.

Each term within the Gene Ontology is organized in a hierarchical manner, with more specific terms commonly being children of more general terms. This hierarchical structure allows for the organization and navigation of gene annotations at different levels of detail and specificity.

The Gene Onotology is not the only way of expressing gene function (and other biological components). Albeit semantically similar to the gene ontology, there exists other resources such as KEGG Brite (various biological objects), Panther (gene function), ChEBI (chemicals of biological interest) and many more[18, 19, 27].

2.2 Computational techniques

2.2.1 Machine Learning

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions. It involves "training" algorithms to recognize patterns and relationships within the data. Machine learning has applications in various fields and continues to advance rapidly, driving automation and data-driven decision-making. Algorithms used in machine learning can take many forms, such as decision trees or deep neural networks, as explained in Alzubaidi *et al.* or Ray *et al.*[28, 29].

2.2.1.1 Supervised Learning

Supervised learning is a fundamental concept in machine learning. It is a type of learning algorithm that uses labeled training data in order to learn. In supervised learning, a data-set typically consists of pairs of input samples (features) and their corresponding outputs (labels). The goal is then typically to train a model that can learn from the training data in order make accurate predictions or generalizations. The model learns from the labeled examples by identifying patterns, relationships, or statistical dependencies between the input and output variables.

Thus, given a set of training examples (X) and output labels (Y), the algorithm generally attempts to learn the function g ($g : X \to Y$).

2.2.1.2 Inductive Logic Programming

Inductive Logic Programming (ILP) is a subfield of artificial intelligence that aims to learn logic programs from examples. While this can take many forms, it is typically done by constructing hypotheses (h) to explain specific examples (E) with the aid of background knowledge (B). In essence, this means that the goal is to infer h, given B and E, such that $B \wedge h \models E$ (otherwise known as sufficiency)[30]. As such, it is mainly focused on inductive reasoning, which involves generalizing from observations or examples to formulate rules or hypotheses. The algorithms in use tend to attempt to induce logical rules or programs from examples by searching through a vast space of possible rules, and selecting the most accurate or descriptive ones. This search is usually performed using the prior knowledge (or background knowledge) as the search space—typically represented as logical facts or rules. These help the learning process, and grounds the search in the specified domain[31, 32].

A logic program (h) usually takes the following form:

$$h \leftarrow b_1, ..., b_n \tag{2.1}$$

Where h and b are atoms (a basic and indivisible proposition or statement, or rather a building block from where to build more complex logical statements).

In order to evaluate the correctness of the induced hypotheses or logic program, ILP typically makes use of the concept of inverse entailment (which is derived through the application of the deduction theorem to the condition of sufficiency). This is a way of turning an inductive problem into a deductive one[30]:

$$\begin{aligned} & B \wedge h \models E \Leftrightarrow B \models (h \rightarrow E) \\ & \Leftrightarrow B \models (\neg E \rightarrow \neg h) \\ & \Leftrightarrow B \wedge \neg E \models \neg h \end{aligned} \tag{2.2}$$

Thus, given a set of positive examples and negative examples, inverse entailment asks whether a given logical rule is capable of discriminating between the two sets of examples. That is, it attempts to verify that the rule entails the positive examples, while not entailing the negative ones[30, 33].

An example logic program in Prolog (using three of the biological concepts mentioned in Section 2.1.3 and used in **Paper II**) could take the following form:

$$\begin{aligned} & \mathsf{Gene(A)} := \\ & \quad \mathsf{Regulated\_By(A,\ B,\ Transcription\ factor)}, \\ & \quad \mathsf{Located\_In(B,\ Mitochondrion)}, \\ & \quad \mathsf{Enzyme\_Metabolite(A,\ Glutamine)} \end{aligned} \tag{2.3}$$

This could then be interpreted as: genes (A) that code for an enzyme catalyzing a reaction involving glutamine and are regulated by a transcription factor (B) located in the mitochondrion.

2.2.2 Genome Scale Modeling

Genome scale metabolic models (GEMs) are computational representations that aim to aid in the study and prediction of metabolic behavior at a systems level. It involves constructing a mathematical representation of the metabolic network of an organism using genomic, biochemical, and physiological data. It is a comprehensive representation of all of the metabolically related reactions occurring in an organism, such as enzymatic reactions and transport reactions. This allows researchers to simulate, analyze and interpret metabolic activity of a specific organism in various different conditions[34–36]. This is typically achieved through the use of methodologies such as flux balance analysis (FBA).

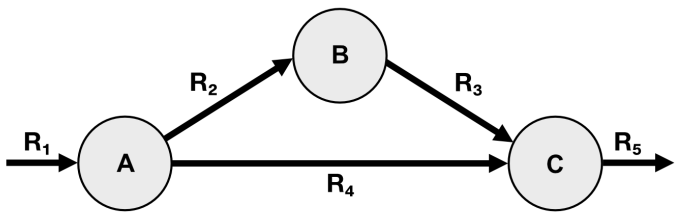

Figure 2.4: Toy metabolic network. Where A, B and C correspond to metabolites, and R_1, R_2, R_3, R_4 and R_5 denote metabolic reactions.

2.2.2.1 Flux Balance Analysis

Flux Balance Analysis (FBA) seeks to model the cell by simulating the flow of metabolites through a metabolic network. This can, for example, enable predictions of growth rates and specific metabolite production rates[34].

Typically, a GEM consists of mathematical representations of metabolic reactions, imposing constraints on the flow of metabolites through the network. Given the toy metabolic network in Figure 2.4 we could infer the following reactions:

$$\textit{Reaction } 1\ (R_1) : \rightarrow A \tag{2.4}$$
$$\textit{Reaction } 2\ (R_2) :\ A \rightarrow B \tag{2.5}$$
$$\textit{Reaction } 3\ (R_3) :\ B \rightarrow C \tag{2.6}$$
$$\textit{Reaction } 4\ (R_4) :\ A \rightarrow C \tag{2.7}$$
$$\textit{Reaction } 5\ (R_5) :\ C \rightarrow \tag{2.8}$$

This can be represented as a stoichiometric matrix (S), where the columns indicate the reactions, and the rows indicate their connected metabolites. FBA is built on the assumption of a steady state ($Sv = 0$). In practice, this implies that metabolites are immediately consumed after production, not allowing for any intracellular accumulation[34]. Note that v denotes the flux, representing the flow of mass through the reactions.

The toy network shown in Figure 2.4, with the underlying assumption of steady state, can then be represented as depicted below (Equation 2.9):

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}}_{\mathrm{S}} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}}_{\mathrm{v}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{2.9}$$

The goal is then to solve for v. However, this would typically result in an underdetermined problem, leading to an infinitely large solution space. By applying mass-balance constraints (introduced by the stoichiometric matrix S) and setting lower and upper bounds ($v_{i,max} < v_i < v_{i,min}$) on potential reaction fluxes, it becomes possible to define a constrained solution space.

By then optimizing for an objective (e.g. maximization of flux through R_5, as defined by the toy-example) we can identify optimal flux distributions by the use of linear programming (in applied cases, one would tend to optimize for maximum growth). This methodology enables the simulation of approximate flux distributions on a genome-wide scale, encompassing thousands of reactions and metabolites[34].

Note that the methodology is not without its limitations, as it typically does not account for regulatory effects (such as transcription or phosphorylation, see Figures 2.2, 2.5 and 2.6). Additionally, due to the steady-state assumption it is typically unable to accurately reflect intracellular concentrations of metabolites. Both of these limitations are revisited in **Paper I**.

An essential component of genome-scale metabolic reconstructions are Gene-Protein-Reaction (GPR) rules. These rules provide Boolean formalizations that connect genes to enzymes and reactions (see Figure 2.5) through boolean logic. It is possible to manipulate this layer, for example by knocking out a gene. In **Paper I** we utilize a combined signalling and gene-regulatory network to infer reaction bounds for the available reactions in a GEM. We then probe the regulatory network by deleting genes with regulation and/or signalling roles and observe the resulting effect on enzymes, predicted fluxes and growth phenotypes.

2.3 Multiomics & Integrative Analysis

"Omics" is a collective term that refers to set of interdisciplinary fields aimed at comprehensively studying certain kinds of biological molecules or processes. These disciplines generate specific types of data that can be used for functional discovery in biological systems. Together they represent a type of flow of information through biological systems, as illustrated in Figure 2.6A and Figure 2.6B. By integrating several different types of omics, one can achieve a much more holistic understanding of the biological system in question.

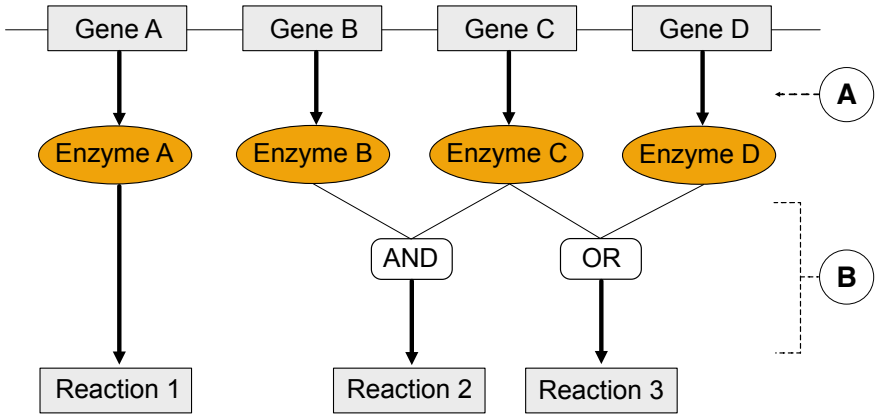

Figure 2.5: The Gene-Protein-Reaction (GPR) rules. Boolean logic is generally used to describe the relationships and conditions that need to be fulfilled for the reaction to carry flux. In this toy example, Gene A codes for Enzyme A, which in turn mediates Reaction 1 (note that the relationship could be one-to-many). Enzyme B and C are required for an active Reaction 2, and conversely, Enzyme C or D would be required for Reaction 3. **A**. Expression of genes into their corresponding proteins can be controlled by regulatory elements outside of the genome-scale model (e.g. a transcription factor). **B.** Enzyme activity can also be regulated by elements outside of the model-abstraction, e.g. kinases.

Genomics refers to the study of genes, transcriptomics the study of RNA (ribonucleic acid), proteomics the study of proteins, and metabolomics the study of metabolites. This thesis will mainly focus on the two latter types of data, namely proteomics and metabolomics.

2.3.1 Metabolomics

Metabolomics is the study of small molecules called metabolites in a biological system. It provides insights into the biochemical pathways and cellular processes that govern metabolism by studying the products and substrates of biochemical reactions. Metabolomics can aid in identifying biomarkers, elucidate metabolic pathways, and study responses to diseases and environmental factors. It is typically seen as the type of data most closely representing the phenotype (observable state) of the organism[37]. Metabolites are typically identified and quantified using advanced analytical techniques, such as mass spectrometry[38]. Mass spectrometry is an analytical technique used to measure the mass and chemical composition of various molecules. It works by ionizing molecules to generate charged particles, which are then separated based on their mass-to-charge ratios. This can provide valuable information about the structure and abundance of specific molecules[39].

Generally, metabolomics is divided into two separate classes of study, namely

extracellular and intracellular. These reflect the physiology of the cell in different ways. The extracellular metabolome describes the substrates and products that the cells input and output from and into the environment around them[40]. Intracellular metabolomics typically describes the internal concentrations of metabolites inside the cell, which are involved in various molecular processes governing the cells functions[41].

When studying metabolomics through mass spectrometry, it is generally approached in either a targeted or untargeted manner. Targeted metabolomics focuses on a predefined set of metabolites, often selected due to their relevance to the biological context of interest. The analysis itself is then usually optimized to allow for reliable detection and quantification of these metabolites. It is particularly useful when studying well-characterized metabolic pathways or systems. Untargeted metabolomics aims to comprehensively analyze the entire metabolome, with the goal of capturing a wide range of different metabolites. There is no reliance on prior knowledge in regards to the biological context, and can provide an unbiased view of the phenomenon of interest. However, it may not provide the same level of reliability and quantitative accuracy that a targeted approach might provide. Regardless of the used methodology, metabolite identification is not a trivial task, as explained in Monge *et al.*[42]. Two of the aforementioned concepts are explored in **Paper I**, namely intracellular and untargeted metabolomics.

2.3.2 Proteomics

Proteomics is a field that focuses on the comprehensive analysis of proteins within a biological system. Proteins are the functional units of the cell, enabling many different biological processes. They play vital roles in virtually all cellular processes, acting as enzymes, signaling molecules, structural components, and more. Understanding the intricate functions, interactions, and modifications of proteins is crucial for explaining and deciphering the complexity of biological systems. Proteomics employs a wide range of techniques and technologies to study proteins on a large scale. This also includes advanced analytical methods such as mass spectrometry[43].

High-throughput quantification of proteins has historically been time-consuming, difficult and expensive. However, during the last decade, mass spectometry-based proteomics has made considerable progress, and it is increasingly able to facilitate biological experiments at scale[43, 44]. This is enabling close to genome-wide coverage in a high-throughput and relatively inexpensive manner.

Paper II utilizes genome-wide proteomic abundances to train supervised machine learning models and evaluate systematized knowledge on *S. cerevisiae*.

2.3.3 Integrative Analysis

A key methodology in Systems biology is integrative analysis, that is, combining several different experimental readouts or levels of biological organization to

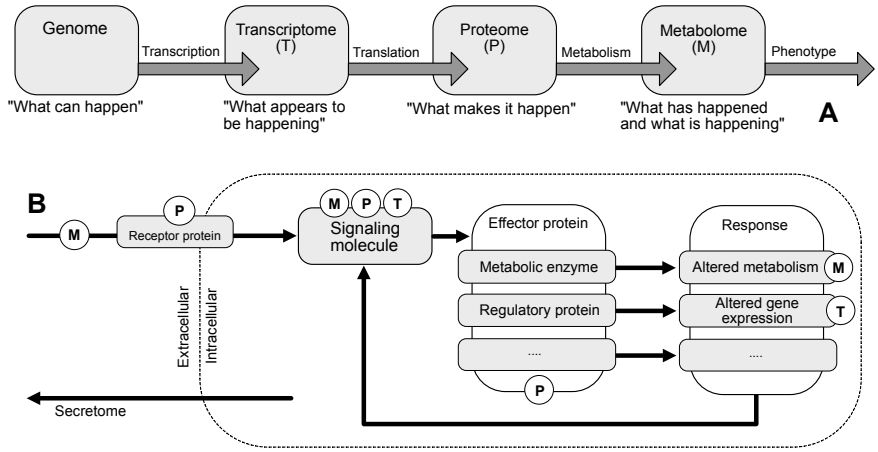

Figure 2.6: **A.** The "Omics-cascade" (Adapted from Dettmer et al.)[37]. Simplified description of the different types of data and levels of organization that could be used to describe the response of biological systems to perturbation (e.g. disease or environmental). **B.** Simplified schematic of mechanism of action in biological systems when exposed to a signalling event or perturbation. Signalling molecules (e.g. proteins, metabolites, RNA) causes an expression or activity change in effector proteins, in turn mediating a response. Response causes a change in internal state, which is communicated by signalling molecules. M, T and P denotes the omics-type that can feasibly represent the different states.

gain a more holistic understanding of the biological system in its entirety. There are a multitude of difficulties with this however, including (but not limited to):

1. Heterogeneity & quality: There are multiple sources and types of biological data, even from the same levels of organization. These could be due to instrumentation, experimental protocols and unit-scales. All of these are prone to different types and magnitudes of noise and fit within different modeling regimes.

2. Dimensionality: Increasing the amount and type of data used could also drastically increase the dimensionality of the problem, making analysis far more difficult.

3. Biological relevance: These systems are highly complex and dependent on conditions. As a result, their interpretation heavily relies on accurately decoupling time-scales and experimental conditions, or on attaining a comprehensive understanding of the conditions themselves.

The type of data integration heavily depends on the frameworks, algorithms, or structures in use. Different omics-types are also suited for investigations into different types of phenomena. As visualized in Figure 2.6B, it is crucial to

assess what type of data is needed for different biological investigations. When investigating metabolism, focusing on metabolomics and proteomics is likely the best approach. These provide robust readouts of the current metabolic state (metabolites) and the effectors of change (proteins).

Paper I is made possible by computational experiment selection, which makes use of a combined signalling- and regulatory network to simulate metabolic states through integration with a genome scale metabolic model. The experiments are analyzed with the help of untargeted metabolomics (in essence collecting a readout the current state of the cells metabolism). This is then contextualized (while also reducing dimensionality) using curated metabolic networks through topological enrichment. Thus, enabling biological interpretation and allowing us to infer the impact of gene deletion on metabolism.

Paper II aggregates several levels of organization (e.g. protein interactions, metabolite concentrations) and structures (e.g. the Gene Ontology) in one unified and flexible formalism (logic programs) to predict the state of the proteome. Thus, evaluating the predictive power of the different levels and their connection to each other, but also providing clues to protein function. Thus completely bypassing the difficulties of data heterogeneity.

Chapter 3

Summary of Included Papers

In this chapter, the two papers included in this thesis are summarized, including relevant research contributions and author contributions. Both papers investigate gene and protein functions, but in different ways. **Paper I** employs semi-automated experiment selection, high-throughput cultivation, and mass spectrometry to characterize several regulatory genes in the context of a biphasic complex biological phenomenon. **Paper II** uses a combination of structured biological priors, inductive logic programming, and supervised learning to learn predictive relationships between gene function, phenotype, and protein levels on a genome-wide scale.

Both papers also address the issue of integrative analysis in distinct ways. In **Paper I**, several modeling regimes and levels of organization are employed sequentially: metabolic flux simulation enabled by gene expression inference for experiment selection, and topological enrichment over a metabolic network for biological interpretation. In **Paper II**, a unified formalism is employed to handle all types of biological data, achieved by translating observed biological facts into logic programs and connecting them to quantified abundances of proteins and metabolites.

3.1 Paper I - High-throughput metabolomics for the design and validation of a diauxic shift model

Problem

In this work, we employ a combination of computer-aided experimental design, automated laboratory cells, and analytical tools to characterize the roles of several genes involved in a metabolic transformation known as the diauxic shift. When *S. cerevisiae* grows on glucose in an aerated batch culture—one can commonly observe a diauxic shift (or biphasic growth). During the initial growth phase, the yeast ferments glucose into ethanol; once glucose has been consumed, the yeast switches to an ethanol substrate through respiration[45][23]. This transition requires a substantial reconfiguration of the metabolic network and a similar phenomenon can be observed in cancer cells known as the Warburg effect (fermentation of glucose into lactate)[46].

Despite extensive research, the regulation of the diauxic shift remains poorly understood[45].

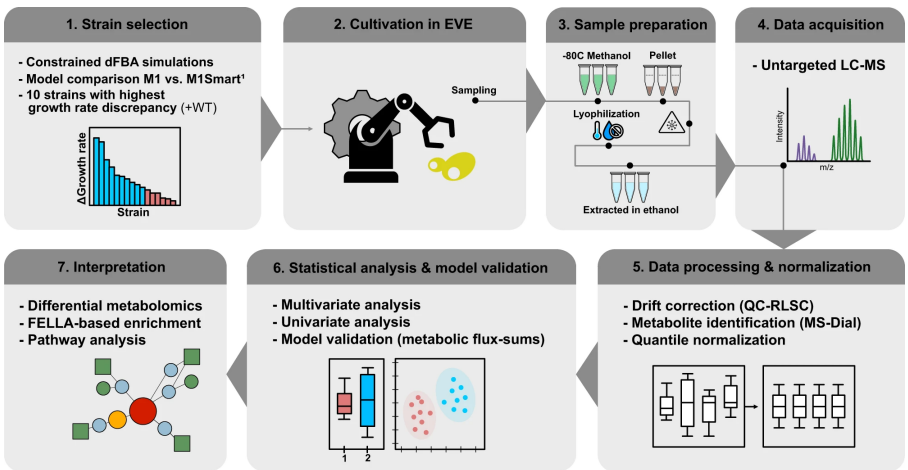

Figure 3.1: Workflow demonstrated in Paper I. dFBA (dynamic Flux Balance Analysis) simulation suggests deletant strains which are subsequently cultivated and analyzed using mass spectrometry and various bioinformatics tools. [1]Simulations using models proposed in Coutant *et al.*[2]

3.1.1 Methodology

Genes were selected based on the simulated impact of specific types of gene deletions on metabolism. This selection was performed using a combined signaling and regulatory network, along with flux balance analysis, within a framework established by previous iterations of the robot scientist concept.

The selection criteria were based on differences in growth phenotype given the structural changes caused by the semi-autonomous model improvements suggested in Coutant *et al.*[2]. These were then investigated through the use of deletant strains (strains of *S. cerevisiae* where the selected gene has been deleted), automated cultivation techniques and untargeted metabolomics. A complete summary of the methodology can be seen in Figure 3.1.

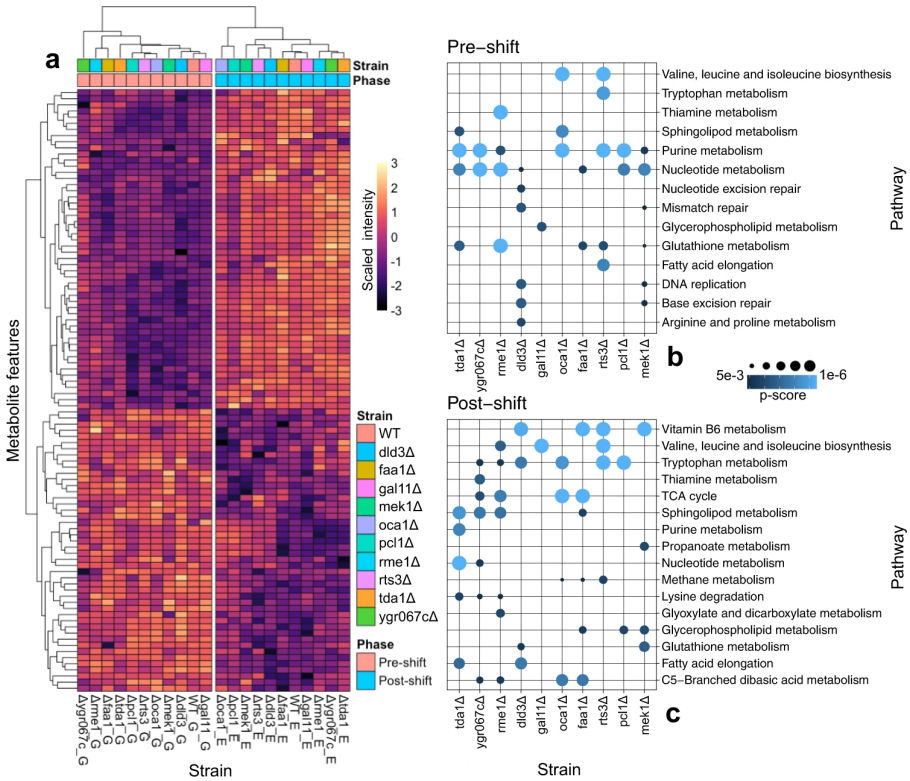

Figure 3.2: Overview of the effects of gene deletion on metabolic profiles. **a.** Metabolic profiles (levels of observable metabolites) for the deletant mutants pre and post diauxic shift. **b**. Pathway enrichment (with the KEGG-derived yeast metabolic network as the background) for the deletant strains pre-shift. **c**. Pathway enrichment (with the KEGG-derived *S. cerevisiae* metabolic network as the background) for the deletant strains post-shift. Pathway enrichment (overrepresentation) aids in inferring impact of the deletions on metabolism (and, in turn, the role of the gene).

3.1.2 Contributions

We demonstrate the suggested workflow by successfully characterizing several genes involved in the diauxic shift, as seen in Figure 3.2. Three of these are (to this date) of largely unknown function (*TDA1*, YGR067C and *RTS3*), and two

have corresponding homologues (a gene that shares a common evolutionary ancestry with another gene) in humans (*DLD3* and *FAA1*).

The study also further characterized the diauxic shift, leveraging the strength of untargeted metabolomics to find subtle, and previously unexplored, changes in metabolism triggered by the metabolic transformation itself. A secondary objective of the study was also to demonstrate the effectiveness of the aforementioned tools for the purposes of future automation and model improvement studies.

3.1.3 Author contributions

Daniel Brunnsåker, Ievgeniia A. Tiukova and Ross D. King concieved and designed the experiments. **Daniel Brunnsåker** performed the wet-lab experiments. Nikul K. Soni, **Daniel Brunnsåker** and Gabriel K. Reder performed the LC/MS sample processing and analysis. **Daniel Brunnsåker**, Gabriel K. Reder and Otto I. Savolainen analyzed the data. Otto I. Savolainen performed the compound identification. **Daniel Brunnsåker**, Alexander H. Gower and Ievgeniia A. Tiukova designed the automated cultivation protocols. **Daniel Brunnsåker**, Ievgeniia A. Tiukova and Ross D. King wrote the manuscript.

3.2 Paper II - Interpreting protein abundance in *Saccharomyces cerevisiae* through relational learning

3.2.1 Problem

Exploring the impact of gene deletions on biological readouts is a fundamental problem in systems biology. Despite having functional annotations for the majority of genes in extensively studied organisms like *Saccharomyces cerevisiae*, achieving a comprehensive understanding of regulatory rules at a systems level remains challenging. In this study, we investigate proteomic and metabolomic profiles derived from a collection of *S. cerevisiae* deletants, utilizing structured priors, relational learning, and supervised machine learning.

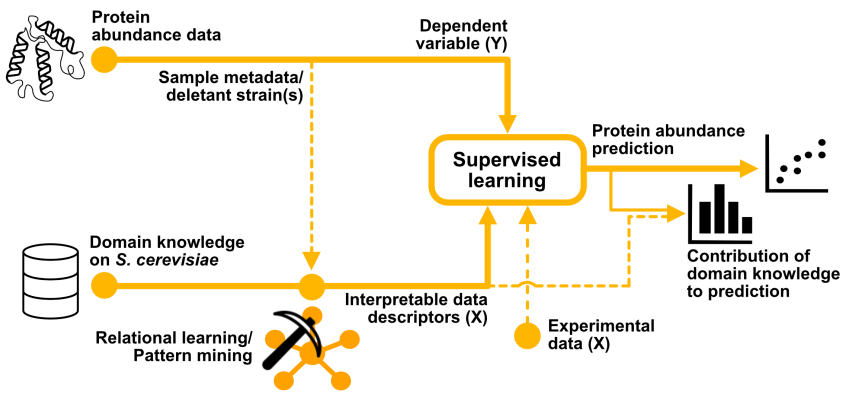

Figure 3.3: Methodology applied in Paper II. Metadata (strain) from data-sets on proteomic abundances are used to identify frequent patterns in a relational database. The frequent patterns are propositionalized and used to predict protein levels in a highly explainable manner.

3.2.2 Methodology

Saccharomyces cerevisiae is a very well studied organism, as such the community has systematized a substantial amount of highly structured and expressive knowledge on its biology. This work subsequently makes use of this prior to learn predictive relationships between proteomic profiles (generated by Messner et al. [44]) and the functional characterization of the yeast genome. This is done by translating this knowledge into an expressive Datalog database, and using frequent pattern mining to generate logic programs—representing biologically relevant regulatory rules. These were then evaluated using supervised learning and feature analysis. See Figure 3.3 for a visual summary.

Some examples of the relations present in the pattern-search can be seen below. Note that this includes concepts from gene regulation, proteomics,

metabolism and phenomics.

ORF_metabolite(+Gene, #Metabolite)
ORF_pathway(+Gene, #Pathway)
ORF_nullphenotype_chemical(+Gene, #Phenotype, #Chemical)
ORF_has_protein_domain(+Gene, #Domain)
regulates(+Gene, −Gene, #Type)

For example, the mode "regulates", consists of an input (+Gene), output (-Gene), and a constant (#Type). This would mean that an allowed clause could include a relation in which gene A (+Gene) regulates gene B (-Gene) by regulating expression or activity (which is discerned from #Type).

The end result would be logic programs consisting of several atoms, such as the example seen in Section 2.2.1.2. These hypotheses/relational features are then assessed by evaluating their predictive power (in terms of proteomic abundances) as seen in Figure 3.4. They can then be easily combined with other types quantitative data if needed.

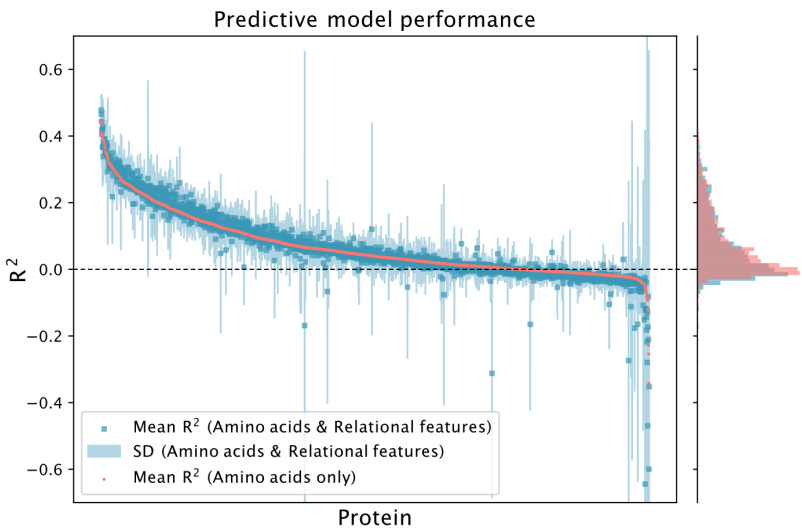

Figure 3.4: Predictability of protein abundance given relational features and metabolite concentrations. R^2 denotes the coefficient of determination (proportion of variance explained).

3.2.3 Contribution

The methodology enabled us to predict protein abundances in an explainable manner. We also learnt several predictive relationships between protein abundances, function and phenotype; such as α-amino acid accumulations and

deviations in chronological lifespan. This was also extended to investigate some specific proteins more closely, namely His4 and Ilv2 (see Figure 3.5); the methodology successfully validated existing literature, but also inferred their roles as regulatory elements for neighboring processes.

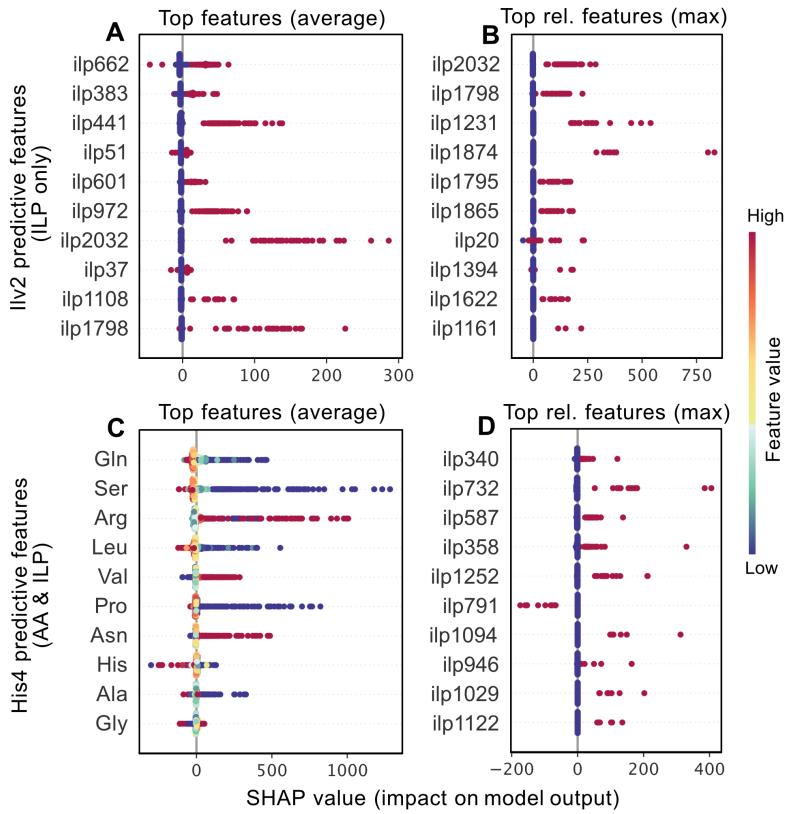

Figure 3.5: **A.** Top features for the prediction of Ilv2, given only relational features. Sorted by average contribution in descending order. **B.** Top relational features for Ilv2, according to maximum change in model output, given only relational features. **C.** Top features for the prediction of His4, given relational features and metabolite concentrations. Sorted by average contribution in descending order. **D.** Top relational features according to maximum change in model output for His4, given relational features and metabolite concentrations. Each dot corresponds to one sample. ilp- denotes that the feature is a generated relational feature. Complete explanations for these descriptors can be seen in the appended manuscript. The x-axis denotes the change in predicted protein abundance caused by a feature (for each sample).

3.2.4 Author contributions

Daniel Brunnsåker and Ross D. King conceived the idea. **Daniel Brunnsåker** and Filip Kronström designed the Datalog database. **Daniel Brunnsåker** performed the analysis, generated the features and trained the machine learning algorithms. **Daniel Brunnsåker** and Ievgeniia A. Tiukova performed the biological interpretation of the results. **Daniel Brunnsåker**, Ievgeniia A. Tiukova and Ross D. King wrote the manuscript.

Chapter 4

Concluding Remarks and Future Directions

This thesis explored methods to enable or accelerate functional discovery in yeast systems biology—either by integrating automation and employing high-throughput analytical techniques or by utilizing various machine learning approaches.

Paper I demonstrates a proof-of-concept for semi-automated experiment selection and untargeted metabolomics as a platform for functional genomics. The automation of this process is likely to offer numerous advantages beyond the increased throughput. Automated experiment selection is likely to reduce research bias and contribute to a more comprehensive and holistic exploration of systems biology. Enabling the integration of robotics and automated sampling processes is anticipated to mitigate human errors. This bears significant importance since the field of biology is currently grappling with a reproducibility crisis, wherein a considerable portion of experiments cannot be reliably replicated[47].

Paper II bridges several levels of biological organization and millions of previously performed experiments to more efficiently use already existing data. To some extent validating some previous findings, but also connecting qualitative statements to quantified experimental readouts. While the predictions may exhibit imperfections and cases of poor accuracy, these issues could (among others) stem from factors such as inconsistencies in the underlying data. Despite this, the predictions offer a high degree of interpretability and draw meaningful conclusions about gene product function and previously uninvestigated regulatory patterns.

However, more concretely, the work conducted as part of the thesis has contributed to the further characterization of the systems biology of *S. cerevisiae*. This contribution primarily involved characterizing several genes, many of which have previously unknown functions. Additionally, the works in this thesis also aided in evaluating existing literature and structured knowledge on *S. cerevisiae* systems biology.

Work in **Paper II** aided in finding connections between metabolite con-

centrations (more specifically, α-amino acid concentrations) and proteomic abundances. Likely due to (among others) the priors involvement in important nutrient signalling pathways and protein translation. This seemed especially true for amino acids like glutamine, glycine and proline. These connections also highlighted the roles of several genes in the context of these abundances.

Paper I also further elucidated the cellular transformation that is the diauxic shift. We showcased subtle metabolic changes in lipid metabolism, amino acid metabolism and oxidative stress responses before and after the transition. The diauxic shift has been an important target of study since the initial discovery by Jacques Monod almost 80 years ago[48]. This due to being an exemplary cellular transformation, but also due to its relevance for industrial and bioengineering purposes. It is also of biomedical interest due to its analogies with the Warburg effect and its potential implications in cancer.

4.1 Future directions

Functional discovery in biological systems needs to be made faster and more efficient. While **Paper I** explored mass-spectrometry as a characterization tool in systems biology, the process needs to be accelerated and automated to a much higher degree. This to remove human bias in processing, but also to accommodate the throughput needed for biological experimentation on a massive scale.

Integrating more experimental modalities bring difficulties in analysis, but they also provide a significant boon for holistic assessments of systems biology. **Paper II** makes use of several ideas from logic programming to aid in this, but these should be integrated more fully with the methodologies from **Paper I** or the concepts introduced in Gower *et al.*[49]. This could be in the form of informed experiment selection, or automated reasoning regarding gene function.

Interpretability of these methods is also paramount. While the proposed methods should be made faster, we need to make sure that the gain in knowledge fit within human understanding and usability. This also means that metadata about experiments (e.g. conditions, experimental techniques and genetic backgrounds) need to be much more stringently recorded than what is typically done today. Enabling computational representations, such as ontologies, to more fully integrate this metadata will allow for more robust experiments and reasoning.

All of these approaches could provide the capability to do automated experiments, analysis and reasoning at scale. Future research will contribute to this by further extending and combining the concepts introduced in this thesis—especially in regards to data integration, experiment selection and automated model improvement.

Bibliography

[1] A. Aderem, "Systems Biology: Its Practice and Challenges," *Cell*, vol. 121, no. 4, pp. 511–513, May 2005, ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2005.04.020.

[2] A. Coutant, K. Roper, D. Trejo-Banos, D. Bouthinon, M. Carpenter, J. Grzebyta, G. Santini, H. Soldano, M. Elati, J. Ramon, C. Rouveirol, L. N. Soldatova and R. D. King, "Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast," *Proceedings of the National Academy of Sciences*, vol. 116, no. 36, pp. 18 142–18 147, Sep. 2019, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1900548116.

[3] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan and A. Clare, "The Automation of Science," *Science*, vol. 324, no. 5923, pp. 85–89, Apr. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1165620.

[4] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin and S. G. Oliver, "Life with 6000 Genes," *Science*, vol. 274, no. 5287, pp. 546–567, Oct. 1996. DOI: 10.1126/science.274.5287.546.

[5] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, Sep. 1997, ISSN: 0036-8075. DOI: 10.1126/science.277.5331.1453.

[6] L. N. Soldatova and R. D. King, "An ontology of scientific experiments," *Journal of The Royal Society Interface*, vol. 3, no. 11, pp. 795–803, Jun. 2006. DOI: 10.1098/rsif.2006.0134.

[7] M. Parapouli, A. Vasileiadis, A.-S. Afendra and E. Hatziloukas, "Saccharomyces cerevisiae and its industrial applications," *AIMS Microbiology*, vol. 6, no. 1, pp. 1–31, Feb. 2020, ISSN: 2471-1888. DOI: 10.3934/microbiol.2020001.

[8] M. H. V. Regenmortel, "Reductionism and complexity in molecular biology," *EMBO Reports*, vol. 5, no. 11, pp. 1016–1020, Nov. 2004, ISSN: 1469-221X. DOI: 10.1038/sj.embor.7400284.

[9] I. Tavassoly, J. Goldfarb and R. Iyengar, "Systems biology primer: The basic methods and approaches," *Essays in Biochemistry*, vol. 62, no. 4, pp. 487–500, Oct. 2018, ISSN: 0071-1365. DOI: 10.1042/EBC20180003.

[10] J. Nielsen and M. C. Jewett, "Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae," *FEMS Yeast Research*, vol. 8, no. 1, pp. 122–131, Feb. 2008, ISSN: 1567-1356. DOI: 10.1111/j.1567-1364.2007.00302.x.

[11] S.-F. Duan, P.-J. Han, Q.-M. Wang, W.-Q. Liu, J.-Y. Shi, K. Li, X.-L. Zhang and F.-Y. Bai, "The origin and adaptive evolution of domesticated populations of yeast from Far East Asia," *Nature Communications*, vol. 9, no. 1, p. 2690, Jul. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-05106-7.

[12] G. Giaever *et al.*, "Functional profiling of the Saccharomyces cerevisiae genome," *Nature*, vol. 418, no. 6896, pp. 387–391, Jul. 2002, ISSN: 1476-4687. DOI: 10.1038/nature00935.

[13] Z. Yang and M. Blenner, "Genome editing systems across yeast species," *Current Opinion in Biotechnology*, Tissue, Cell and Pathway Engineering, vol. 66, pp. 255–266, Dec. 2020, ISSN: 0958-1669. DOI: 10.1016/j.copbio.2020.08.011.

[14] D. Botstein and G. R. Fink, "Yeast: An Experimental Organism for 21st Century Biology," *Genetics*, vol. 189, no. 3, pp. 695–704, Nov. 2011, ISSN: 0016-6731. DOI: 10.1534/genetics.111.130765.

[15] W. Ea *et al.*, "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, Aug. 1999, ISSN: 0036-8075. DOI: 10.1126/science.285.5429.901.

[16] E. M. Bunnik and K. G. Le Roch, "An Introduction to Functional Genomics and Systems Biology," *Advances in Wound Care*, vol. 2, no. 9, pp. 490–498, Nov. 2013, ISSN: 2162-1918. DOI: 10.1089/wound.2012.0379.

[17] R. Haas, A. Zelezniak, J. Iacovacci, S. Kamrad, S. Townsend and M. Ralser, "Designing and interpreting 'multi-omic' experiments that may change our understanding of biology," *Current Opinion in Systems Biology*, Systems biology of model organisms, vol. 6, pp. 37–45, Dec. 2017, ISSN: 2452-3100. DOI: 10.1016/j.coisb.2017.08.009.

[18] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27.

[19] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima and M. Ishiguro-Watanabe, "KEGG for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Research*, vol. 51, no. 1, pp. 587–592, Jan. 2023, ISSN: 1362-4962. DOI: 10.1093/nar/gkac963.

[20] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Science: A Publication of the Protein Society*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019, ISSN: 1469-896X. DOI: 10.1002/pro.3715.

[21] M. Gillespie *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Research*, vol. 50, no. D1, pp. 687–692, Jan. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkab1028.

[22] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, A. Kothari, I. M. Keseler, M. Krummenacker, P. E. Midford, Q. Ong, W. K. Ong, S. M. Paley and P. Subhraveti, "The BioCyc collection of microbial genomes and metabolic pathways," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1085–1093, Jul. 2019, ISSN: 1477-4054. DOI: 10.1093/bib/bbx085.

[23] L. Geistlinger, G. Csaba, S. Dirmeier, R. Küffner and R. Zimmer, "A comprehensive gene regulatory network for the diauxic shift in Saccharomyces cerevisiae," *Nucleic Acids Research*, vol. 41, no. 18, pp. 8452–8463, Oct. 2013, ISSN: 1362-4962. DOI: 10.1093/nar/gkt631.

[24] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, Oct. 2002, ISSN: 1095-9203. DOI: 10.1126/science.1075090.

[25] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei and L. Chen, "Inference of Gene Regulatory Network Based on Local Bayesian Networks," *PLOS Computational Biology*, vol. 12, no. 8, e1005024, Aug. 2016, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005024.

[26] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene Ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000, ISSN: 1061-4036. DOI: 10.1038/75556.

[27] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic acids research*, vol. 44, no. 1, pp. 1214–9, Jan. 2016, ISSN: 1362-4962. DOI: 10.1093/nar/gkv1031.

[28] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, ISSN: 2196-1115. DOI: 10.1186/s40537-021-00444-8.

[29] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Feb. 2019, pp. 35–39. DOI: 10.1109/COMITCon.2019.8862451.

[30] S. Muggleton, "Inductive Logic Programming: Issues, results and the challenge of Learning Language in Logic," *Artificial Intelligence*, vol. 114, no. 1, pp. 283–296, Oct. 1999, ISSN: 0004-3702. DOI: 10.1016/S0004-3702(99)00067-3.

[31] L. D. Raedt, *Logical and Relational Learning*. Springer Science & Business Media, Sep. 2008, ISBN: 978-3-540-68856-3.

[32] A. Srinivasan, *The Aleph Manual*. [Online]. Available: https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html.

[33] S. Muggleton and L. De Raedt, "Inductive logic programming: Theory and methods," *The Journal of Logic Programming*, vol. 19, pp. 629–679, 1994.

[34] J. D. Orth, I. Thiele and B. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, Mar. 2010, ISSN: 1546-1696. DOI: 10.1038/nbt.1614.

[35] H. Lu, F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven and J. Nielsen, "A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism," *Nature Communications*, vol. 10, no. 1, p. 3586, Aug. 2019, ISSN: 2041-1723. DOI: 10.1038/s41467-019-11581-3.

[36] M. L. Mo, B. Palsson and M. J. Herrgård, "Connecting extracellular metabolomic measurements to intracellular flux states in yeast," *BMC Systems Biology*, vol. 3, no. 1, p. 37, Mar. 2009, ISSN: 1752-0509. DOI: 10.1186/1752-0509-3-37.

[37] K. Dettmer, P. A. Aronov and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007, ISSN: 0277-7037. DOI: 10.1002/mas.20108.

[38] S. Alseekh *et al.*, "Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices," *Nature Methods*, vol. 18, no. 7, pp. 747–756, Jul. 2021, ISSN: 1548-7105. DOI: 10.1038/s41592-021-01197-1.

[39] G. L. Glish and R. W. Vachet, "The basics of mass spectrometry in the twenty-first century," *Nature Reviews Drug Discovery*, vol. 2, no. 2, pp. 140–150, Feb. 2003, ISSN: 1474-1784. DOI: 10.1038/nrd1011.

[40] F. R. Pinu and S. G. Villas-Boas, "Extracellular Microbial Metabolomics: The State of the Art," *Metabolites*, vol. 7, no. 3, p. 43, Aug. 2017, ISSN: 2218-1989. DOI: 10.3390/metabo7030043.

[41] A. Zhang, H. Sun, H. Xu, S. Qiu and X. Wang, "Cell Metabolomics," *OMICS : a Journal of Integrative Biology*, vol. 17, no. 10, pp. 495–501, Oct. 2013, ISSN: 1536-2310. DOI: 10.1089/omi.2012.0090.

[42] M. E. Monge, J. N. Dodds, E. S. Baker, A. S. Edison and F. M. Fernández, "Challenges in Identifying the Dark Molecules of Life," *Annual review of analytical chemistry (Palo Alto, Calif.)*, vol. 12, no. 1, pp. 177–199, Jun. 2019, ISSN: 1936-1327. DOI: 10.1146/annurev-anchem-061318-114959.

[43] C. B. Messner, V. Demichev, Z. Wang, J. Hartl, G. Kustatscher, M. Mülleder and M. Ralser, "Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology," *Proteomics*, vol. 23, Nov. 2022, ISSN: 1615-9861. DOI: 10.1002/pmic.202200013.

[44] C. B. Messner, V. Demichev, J. Muenzner, S. K. Aulakh, N. Barthel, A. Röhl, L. Herrera-Domínguez, A.-S. Egger, S. Kamrad, J. Hou, G. Tan, O. Lemke, E. Calvani, L. Szyrwiel, M. Mülleder, K. S. Lilley, C. Boone, G. Kustatscher and M. Ralser, "The proteomic landscape of genome-wide genetic perturbations," *Cell*, vol. 186, no. 9, 2018–2034.e21, Apr. 2023, ISSN: 1097-4172. DOI: 10.1016/j.cell.2023.03.026.

[45] D. Brunnsåker, G. K. Reder, N. K. Soni, O. I. Savolainen, A. H. Gower, I. A. Tiukova and R. D. King, "High-throughput metabolomics for the design and validation of a diauxic shift model," *npj Systems Biology and Applications*, vol. 9, no. 1, pp. 1–9, Apr. 2023, ISSN: 2056-7189. DOI: 10.1038/s41540-023-00274-9.

[46] M. V. Liberti and J. W. Locasale, "The Warburg Effect: How Does it Benefit Cancer Cells?" *Trends in biochemical sciences*, vol. 41, no. 3, pp. 211–218, Mar. 2016, ISSN: 0968-0004. DOI: 10.1016/j.tibs.2015.12.001.

[47] K. Roper, A. Abdel-Rehim, S. Hubbard, M. Carpenter, A. Rzhetsky, L. Soldatova and R. D. King, "Testing the reproducibility and robustness of the cancer biology literature by robot," *Journal of The Royal Society Interface*, vol. 19, no. 189, p. 20 210 821, Apr. 2022. DOI: 10.1098/rsif.2021.0821.

[48] J. Monod, "The Growth of Bacterial Cultures," *Annual Review of Microbiology*, vol. 3, no. 1, pp. 371–394, 1949. DOI: 10.1146/annurev.mi.03.100149.002103.

[49] A. H. Gower, K. Korovin, D. Brunnsåker, I. A. Tiukova and R. D. King, *LGEM+: A first-order logic framework for automated improvement of metabolic network models through abduction*, arXiv preprint, Jun. 2023. DOI: 10.48550/arXiv.2306.06065.