Thesis for The Degree of Doctor of Philosophy

Automating Hypothesis Generation and Testing: Towards Self-driving Biology

Enabling high-throughput scientific discovery in Saccharomyces cerevisiae

Daniel Brunnsåker

Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2025

Automating Hypothesis Generation and Testing:
Towards Self-driving Biology
Enabling high-throughput scientific discovery in Saccharomyces cerevisiae

Daniel Brunnsåker

ISBN 978-91-8103-297-0
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5755.
ISSN 0346-718X

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

Cover illustration by Daniel Brunnsåker (2025). The image depicts a robot conducting scientific research, surrounded by icons representing the stages of the scientific process (hypothesis formation, planning, experimentation, analysis, and reporting).

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

"To make interesting scientific discoveries, you should acquire as many good friends as possible who are energetic, intelligent and knowledgeable as they can be. You will find all the programs you need are stored in your friends, and will execute productively and creatively as long as you don't interfere too much."

- Herbert A. Simon

Automating Hypothesis Generation and Testing: Towards Self-driving Biology

Enabling high-throughput scientific discovery in Saccharomyces cerevisiae

Daniel Brunnsåker

Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Biological systems remain only partially understood, and the relative pace of functional discovery has been slowing down despite advances in measurement technologies. A growing consensus suggests that the most promising way forward is not only via conventional laboratory automation, but through the development of fully autonomous systems that can generate, prioritize, draw insight from, and execute high-throughput experimentation.

This thesis explores how such automation can accelerate scientific discovery by combining methods from artificial intelligence—such as inductive logic programming, explainable AI, and large language models—with physical instrumentation, including laboratory robotics and high-throughput analytical platforms like mass spectrometry. The work spans the entire discovery cycle, from hypothesis generation to experimental evaluation.

As a case study, the methods are applied to *Saccharomyces cerevisiae* (baker's yeast), an extensively studied eukaryote and a powerful model organism for systems biology. In doing so, the thesis contributes to further characterization of key aspects of yeast biology, including the diauxic shift and its regulators (via untargeted metabolomics), genome-wide proteomic regulation, phenotypic determinants of fitness, and metabolic interactions involving amino acids.

The findings emphasize that automation in biology requires more than throughput alone. Automated systems must also leverage existing knowledge, provide interpretable reasoning processes, and preferably capture enough metadata for auditability. These studies also highlight how automation, when combined with structured knowledge and high-throughput experimentation, can refine existing approaches and move biology toward more integrative and transparent modes of discovery.

Keywords

Automation of science, laboratory automation, machine learning, inductive logic programming, systems biology, metabolomics, mass spectrometry

Acknowledgments

First of all, I would like to thank my supervisors, Ross and Ievgeniia. Thank you for giving me this opportunity, for your support and guidance, and for allowing me such a high degree of freedom. Ross, you have taught me not only scientific skills but also how to think and act like a scientist, always with integrity and rigour—lessons I will carry with me for the rest of my life. 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. Alec, Filip, Erik, Prajakta, Gabriel and Beera, thanks for making (almost) every day of my PhD-journey amazing. I am especially grateful to Alec, who started at the same time as me and has shared so many of the same hardships during our time here. Facing them together has made the journey lighter, and you have pushed me to grow—I truly believe I am a much smarter and caring person thanks to you!

A huge 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 my fellow PhD students at DSAI; you have been a pleasure to hang out with, whether we were having coffee in the lunchroom or singing karaoke in Tokyo.

Thanks to all of my friends (with some of you hitting almost every category of this acknowledgement!). I cannot emphasise enough how much you mean to me, and I don't think I would have made it through this with my sanity (or what's left of it) intact without you. I feel truly fortunate to have gotten to know such amazing people—most would be lucky to receive even a fraction of the support you've given me.

To my family, thank you for your never-ending support, no matter the situation. I wouldn't even have considered undertaking something of this magnitude if it weren't for all of you (even though I am still not sure you know what I've been doing for the last few years). I sometimes wonder if Morfar ever imagined I would end up here.

Lastly, I would like to thank my partner, Francine. Thank you for being my best friend and my greatest supporter. You are the best thing that has ever happened to me. Thank you for all the adventures, and for the many more still 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.

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper 1**] **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, Volume 9, Issue 11, April 2023.

[**Paper 2**] **D. Brunnsåker**, F. Kronström, I.A. Tiukova & R.D. King. *Interpreting protein abundance in Saccharomyces cerevisiae through relational learning.* Bioinformatics, Volume 40, Issue 2, February 2024.

[**Paper 3**] G.K. Reder, E.Y. Bjurström, **D. Brunnsåker**, F. Kronström, P. Lasin, I.A. Tiukova, O.I. Savolainen, J.N. Dodds, J.C. May, J.P. Wikswo, J.A. McLean & R.D. King. *AutonoMS: Automated Ion Mobility Fingerprinting.* Journal of the American Society for Mass Spectrometry, Volume 35, Issue 3, February 2024.

[**Paper 4**] F. Kronström, **D. Brunnsåker**, I.A. Tiukova & R.D. King. *Ontology-based box embeddings and knowledge graphs for predicting phenotypic traits in Saccharomyces cerevisiae.* Proceedings of Machine Learning Research, Volume 284, 9th International Conference on Neurosymbolic Learning and Reasoning, Santa Cruz, September 2025.

[**Paper 5**] **D. Brunnsåker**, A.H. Gower, P. Naval, E.Y. Bjurström, F. Kronström, I.A. Tiukova & R.D. King. *Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology.* Manuscript in preparation.

Other publications

The following publications were published during my PhD studies, or are currently in submission. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] A.H. Gower, K. Korovin, **D. Brunnsåker**, I.A. Tiukova & R.D. King. *LGEM$^+$: a first-order logic framework for automated improvement of metabolic network models through abduction.* In International Conference on Discovery Science, Cham: Springer Nature Switzerland, October 2023.

[**b**] A.H. Gower, K. Korovin, **D. Brunnsåker**, F. Kronström, G.K. Reder, I.A. Tiukova, R.S. Reiserer, J.P. Wikswo & R.D. King. *The use of AI-robotic systems for scientific discovery.* Under submission.

[**c**] E.Y. Bjurström, P. Lasin, **D. Brunnsåker**, I.A. Tiukova & R.D. King. *An Investigation of TDA1 Deficiency in Saccharomyces cerevisiae during diauxic growth.* Yeast, Volume 42, June 2025.

[**d**] F. Kronström, A.H. Gower, **D. Brunnsåker**, I.A. Tiukova & R.D. King. *Graph Neural Network based Hierarchy-aware Box Embeddings of Knowledge Graphs.* Under review.

Summary of contributions

Paper 1: Conceived and designed the experiments. Performed the experiments. Wrote the code. Analysed the data. Wrote the manuscript. Designed automated protocols.

Paper 2: Co-conceptualized the study. Created the database. Wrote the code. Analysed the data. Wrote the manuscript.

Paper 3: Co-designed the experiments. Selection of standards and creation of reference databases. Co-performed the experiments. Wrote parts of the manuscript.

Paper 4: Conceived and designed the experiments. Conceptualized application area. Performed the experiments. Designed automated protocols. Wrote parts of the manuscript. Contributed to data analysis.

Paper 5: Conceptualized the study. Wrote most of the code. Conceived and designed the experimental designs. Constructed automated protocols and performed the experiments. Wrote most of the manuscript. Analysed the data.

Summary

The salient points of this thesis are as follows:

Methodological contributions

- Logic programs generated from descriptive ontologies can serve as flexible and testable hypotheses in both high-throughput and automated settings.
- Relational database-derived logic programs can be used to interpretably predict quantified biological abundances (such as protein and metabolite levels) and to infer protein function.
- Knowledge priors structured in semantically meaningful ways can be used to predict phenotypic traits such as digenic and trigenic fitness.
- Ontology-based embeddings, combined with explainable AI techniques, can be used to generate actionable hypotheses in yeast physiology.
- Large Language Models can be used to automatically generate interventions designed to test logically structured hypotheses in a systems biology setting.
- Formalizing knowledge representation from the outset can improve the reliability of automated approaches for hypothesis generation and validation.

Automation and data modalities

- Mass spectrometry–based metabolomics can be automated in both sample preparation and analysis.
- These automated workflows can be cheap, reliable, and high-throughput.
- Ion mobility–based mass spectrometry, particularly when combined with rapid separation methods such as SPE (solid-phase extraction), provides an effective compromise between acquisition speed and resolution in yeast systems biology.
- Untargeted metabolomics can rapidly generate exploratory hypotheses about gene function and contextual regulation.

- It can also provide partial verification of such hypotheses in model organisms such as *S. cerevisiae*.
- Metabolomics can be used to generate testable implications for hypotheses, enabling automated hypothesis-driven experimentation.
- Metabolomics data are a promising candidate for use with automatic hypothesis refinement.
- Phenomics and mass spectrometry integration enables automated discovery cycles that generate and refine hypotheses with improved context.

Biological insights

- The diauxic shift involves not only a metabolic switch between fermentation and respiration, but also major metabolic adjustments to internal and environmental stressors.
- Among these adjustments, reactive oxygen species (ROS) scavenging is particularly pronounced.
- Untargeted metabolomics revealed likely secondary functions for well-annotated genes, such as *FAA1*, *DLD3*.
- Putative proteins YGR067C and *RTS3* likely play indirect roles in the TCA cycle, vitamin B6 metabolism, and amino acid metabolism.
- Many amino acids show synergistic or antagonistic growth inhibition effects when combined with common compounds in *S. cerevisiae*, such as arginine and caffeine or glutamate and spermine.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Science is undergoing a transformation. The traditional, human-driven scientific method—rooted in observation, hypothesis formation, experimentation and analysis—is straining under the complexity and scale of modern research (Musslick et al., 2025). This is especially true in fields like biology, where the systems we study produce enormous volumes of data and exhibit a level of complexity far beyond what a human can experimentally analyze or comprehend within a reasonable timeframe (Dasgupta et al., 2023; Kitano, 2002).

These challenges mirror those seen in other domains that have undergone automation. Just as the industrial revolution mechanized physical labour, and as self-driving systems, such as cars, now handle tasks once thought impossible, the process of scientific discovery, too, is approaching a paradigm shift (Kuhn, 2012). Increasingly, we are exploring how machines might not only assist scientists, but potentially participate in, or even drive, the scientific process itself.

By automating repetitive and time-consuming aspects of research, machines can increasingly handle the scale and precision of modern science. This raises the possibility of a new division of labour: automated systems managing data and experimentation, while humans focus on tasks that demand creativity, interpretation, and intuition. The automation of scientific discovery is therefore not only about speed, but about reconfiguring the process itself. It involves artificial intelligence for hypothesis generation, robotics and high-throughput platforms for experimentation, and new frameworks for structuring knowledge so that both humans and machines can build upon it.

Systems biology serves as a natural proving ground for scientific automation. Through it we seek to understand complex biological systems by integrating computational models with large-scale experimental data across multiple levels of biological organization. The field is, by necessity, interdisciplinary, relying on the integration of methods from biology, computer science, engineering, and mathematics. It also exemplifies the kind of complexity and data-rich environment that exceeds human capacity, making it both a motivation for and a beneficiary of automation. Scientific progress in systems biology increasingly depends on the ability to iterate rapidly across experimental and computational

cycles, integrate heterogeneous datasets, and generate interpretable models.

In this thesis, I investigate scientific automation through its application to the systems biology of *Saccharomyces cerevisiae*, commonly known as baker's yeast. As the most extensively studied eukaryotic organism, yeast systems biology is accompanied by a vast source of structured biological knowledge, curated data, and well-established experimental tools—making it an ideal model system for developing and testing automated approaches.

The research is structured around five studies, which investigates and apply different ideas for automating parts of the scientific process, as illustrated in Fig 1.1:

- **Metabolomics in scientific automation:** Paper 1 investigates the use of untargeted metabolomics (a concept more thoroughly explained in Chapter 3) data as a source of information for future automated approaches. It is an information-rich and automation-friendly data modality, not fully explored in systems biology. We apply it to investigate gene function in the role of a complex temporal phenomena—the diauxic shift. We additionally present a proof of concept for its use in model validation in active-learning-based approaches for metabolic models (Brunnsåker et al., 2023).

- **Explainable AI for hypothesis generation:** Paper 2 applies relational learning and explainable AI techniques to generate regulatory rules and genotype-phenotype relations from structured biological knowledge. We then apply these findings to protein abundances in *S. cerevisiae*, automatically generating human-readable, and logically rooted hypotheses about protein function and genotype-phenotype relations (Brunnsåker et al., 2024).

- **Automated data acquisition:** Paper 3 explores and develops software and hardware integration that automates metabolomics data acquisition, enabling future avenues of downstream automation. The methodology is evaluated on several different biologically interesting biomolecules, and via untargeted metabolomics capture in a biological matrix (*S. cerevisiae*) (Reder et al., 2024).

- **Structured knowledge representation:** Paper 4 utilises structured knowledge about yeast biology to construct a vast resource of information about genetic modifications, and investigates how to represent them efficiently and expressively. This is combined with a graph neural network, and used to predict outcomes of digenic interactions (direct or indirect interactions between two genes). We additionally use the framework as a hypothesis generator, experimentally validating its outputs.

- **Self-driving science:** Paper 5 combines many of the concepts introduced in previous works, and uses them to produce an automated framework for scientific discovery. It combines relational learning for hypothesis generation, large language models for experimental design, and automated

laboratory infrastructure in order to generate and evaluate hypotheses about *S. cerevisiae* phenotype. It does so in an explainable manner, with data representations interpretable by both humans and machines.

Together, these five studies represent a progression toward automating various components of the scientific discovery cycle in systems biology. From data acquisition and hypothesis generation to experimental validation and interpretation. While each study contributes to separate pieces of the process, their combined impact lies in demonstrating how structured knowledge representations, high-throughput experimentation, and machine learning can enable automated scientific inquiry.

The remainder of this thesis builds the context and practical foundation underlying this work. First, I examine the motivations for automating science, including both its opportunities and its limitations. I then delve into the complexity of systems biology, exploring why it is a field which might be a good beneficiary of automation. Finally, I explore how exactly machine learning can be leveraged to support scientific discovery.

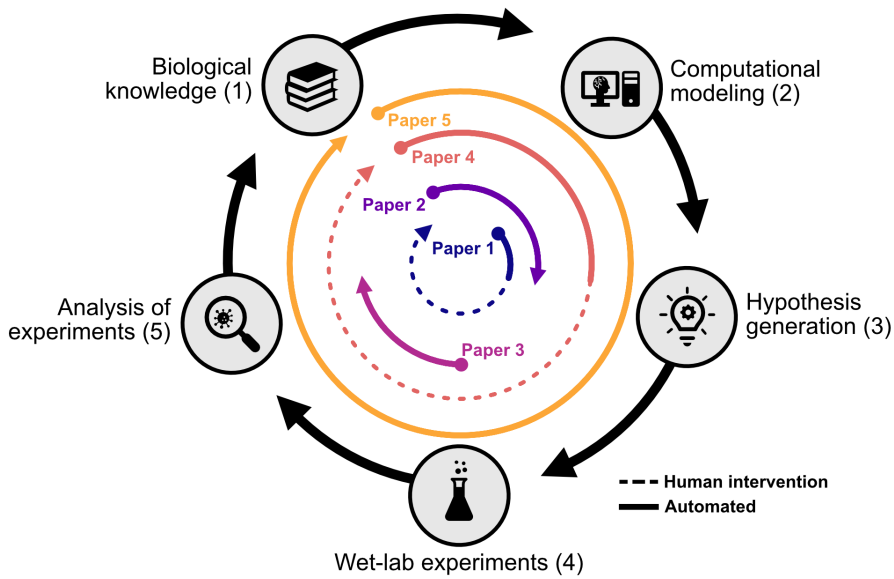


Figure 1.1: The classic iterative cycle of systems biology: (1) utilize existing knowledge, (2) mathematically represent and model it, (3) generate testable hypotheses, (4) test the hypotheses experimentally, (5) analyse the outcomes, and integrate new findings into the current body of knowledge. The arrows represent what factors of the cycle have been performed for each of the papers included in the thesis, and to what extent they have been automated. Dashed arrows represent steps that have been manually performed.

Chapter 2

The Automation of Science

The idea that robots could one day perform scientific reasoning has steadily moved from theory to practice. As modern science tries to deal with increasingly complex systems and overwhelming data volumes, the need to rethink the traditional scientific method has become more and more pressing. Automation offers a compelling alternative.

The following chapter explores the concept of automated science: the use of artificial intelligence, robotics, and structured knowledge to support or even carry out core components of the scientific process. While elements of these concepts have long existed in isolated forms, such as high-throughput screening or laboratory robotics, recent advances now enable integrated, end-to-end systems capable of generating hypotheses, planning and executing experiments, and interpreting results (King et al., 2004; King et al., 2009; Williams et al., 2015; Coutant et al., 2019; Ghareeb et al., 2025; Gottweis et al., 2025).

These ideas are not new. The field of automated scientific discovery traces its roots back to early expert systems like DENDRAL, which interpreted mass spectrometry data to automatically infer molecular structures (Lindsay et al., 1993). It continued with systems like BACON (Langley, 1979), which rediscovered physical relations, such as Kepler's laws. These concepts then eventually transitioned into robotic platforms like Adam (King et al., 2009) and Eve (Williams et al., 2015) that integrated hypothesis generation, experimentation, and analysis. These systems laid the groundwork for what is now a growing and interdisciplinary effort to reshape modern science into a machine-augmented (or even machine-led) process (Gottweis et al., 2025; Ghareeb et al., 2025; C. Lu et al., 2024).

In this chapter, we first examine the key motivations and benefits behind automating science, including managing systems of high scale and complexity, increasing reproducibility, and the potential to further democratize science. We then turn to the functional parts of a robot scientist and the tools needed to perform automated science in systems biology. Lastly, we discuss the challenges, practical and philosophical, including concerns around bias, interpretability, ethical implications, and loss of human oversight.

2.1 Motivation for Automating Science

Automating scientific processes offers numerous advantages that address significant challenges in modern research. By optimizing the use of resources such as reagents, equipment, and time, automation reduces waste and improves efficiency. It also minimizes human exposure to hazardous materials and high-risk environments, enhancing safety (and potentially even regulatory compliance). Many scientific protocols are complex and repetitive, making them prone to human error; automation can ensure that these procedures are executed with precision and consistency. Furthermore, automated systems can operate continuously without breaks, enabling time-efficient experimentation and significantly accelerating the pace of discovery.

While the benefits of automation apply across many scientific domains, they are particularly useful in fields that are data-intensive, require high experimental throughput, and involve highly complex systems. Manual workflows are no longer sufficient to process or make sense of these domains at the speed and scale modern research demands.

Amongst the scientific disciplines, the life sciences present a particularly urgent case for automation, not only due to its scale and complexity, but also because of the transformative potential of automation in areas like sustainability, healthcare, and food security. Biological systems are not just large; they are deeply structured, dynamic, and context-sensitive (Kitano, 2002). These characteristics expose several broader challenges that increasingly limit the impact and sensitivity of traditional scientific workflows. Overcoming them will require more than simply accelerating existing protocols, automation must also be capable of guiding and refining the process itself (Coutant et al., 2019).

In what follows, three challenges are illustrated: the complexity of modern biological systems, difficulties in reproducibility, and the barriers to access and perform science.

Complexity

Biological complexity presents one of the most significant challenges to modern science. Systems-level biology involves the interplay of thousands of genes, proteins, and metabolites—often under dynamic and ever-shifting environmental conditions. Some illustratory examples and challenges in this field include:

- **Combinatorial explosion:** Yeast is a fairly simple organism by eukaryotic standards, but even if accounting for only pairwise interactions among the ∼6,000 genes available in *S. cerevisiae*, it implies $\binom{6000}{2} \approx 18$ million experiments. Far beyond manual capacity, as evidenced in Costanzo et al. (2016).

- **Temporal dynamics:** Regulatory networks and other biological mechanisms shift over time (cell cycle, stress responses, subtle environmental changes), demanding fine-grained sampling that only automated and continuously monitoring platforms are likely to be able to provide.

- **Multiscale interactions:** Data collection and analysis of processes spanning molecular and cellular levels of organization (e.g., kinases regulating enzyme activity, affecting reaction fluxes, which in turn affects environmental sensitivities) require highly coordinated and precise workflows that bridge these scales. Depending on the phenomena, these workflows can easily extend beyond the capabilities of manual human analysis and intervention.

These challenges are not only technical. They define the limits of what human researchers can realistically study without assistance. Addressing these challenges will require more than simply automating existing laboratory procedures. To fully realize modern discovery, automated systems must also be able to guide the scientific process itself, integrating prior knowledge, generating and prioritizing hypotheses and adapting to experiments in real time. In other words, automation must be coupled with intelligent decision-making that keeps the experimental cycles efficient and informative.

Reproducibility

Several studies have shown that biology is facing a reproducibility crisis, where it is difficult to repeat, much less replicate results (Baker, 2016; Munafò et al., 2017). This is eroding confidence in key discoveries and highlighting a large amount of wasted resources and scientific dead-ends (Roper et al., 2022). This crisis can be observed across several different aspects, such as:

- **Protocol variability:** Manual pipetting, differences in timing, subtle (and often undocumented) technician-induced tweaks, slight variations in instrumentation, equipment differences, and the age of reagents are all likely to introduce high degrees of variability.

- **Software drift:** While not restricted to the life sciences, undocumented changes or tweaks in scripts or software versions could lead to non-identical analysis pipelines over time.

- **Metadata loss:** Key contextual details in experimental protocols such as incubator humidity, processing details (e.g., "shake vigorously"), instrument calibration dates, and ambient temperature are often recorded informally (and sometimes even completely omitted), making true replication impossible.

Automation has the ability to directly tackle most, if not all, of these issues. Not only are they essential to proper science, the inclusion of these aspects are a necessity for successful scientific automation:

- **Standardized protocol execution:** Robotics ensure that pipetting volumes, mixing steps, and incubation times are identical across runs. Whilst there could still be errors, they are far more likely to be predictable (and traceable).

- **Version-controlled analysis:** Integration of version management, containerized pipelines, workflow managers and orchestration frameworks can alleviate issues and inconsistencies in data processing and analysis.
- **Comprehensive metadata capture:** Due to the nature of automation, it allows for a much higher degree of detail of performed actions. Consistently recording metadata during both planning and runtime steps could allow for true reproducibility by not omitting any key details.

Note that many of these aspects do not strictly require automation in order to be used in conjunction with scientific workflows, but by more fully integrating automated solutions into the scientific process, many of these issues are likely solved as a byproduct of its inclusion.

Democratizing Science

Despite its global impact, cutting-edge biological research has historically been confined to a relatively small number of well-funded institutions. This is largely due to the high costs of specialized instrumentation, infrastructure, and technical expertise. These continue to grow as experimental platforms evolve and become more complex. As a result, the opportunity to ask and investigate important questions in health, sustainability, and biotechnology are unequally distributed amongst the scientific community.

Scientific automation, if implemented *responsibly*, offers a way to challenge this imbalance. Cloud-connected, open-source platforms and remotely accessible lab automation tools could significantly lower the threshold for conducting rigorous biological experiments. These systems can reduce the need for deep technical specialization, making it easier for a broader range of people to contribute to scientific research in a more consistent and reproducible way.

In this light, automation is not just a way to scale and proceduralize scientific throughput, it is also a means to broaden participation, diversify perspectives, and make the process of science more inclusive and globally accessible.

2.2 Social and Ethical Considerations

While automating the scientific process has great advantages, it also raises important ethical and social questions and considerations, many of which echo long-standing concerns around automation. Historically, automation has led to the displacement of labour, deskilling of workers, and concentration of power (Frey et al., 2017). Similar risks arise in the context of scientific automation and must be addressed carefully.

One concern is the loss of scientific skill and intuition. As machines take on core tasks like hypothesis generation or data analysis, opportunities for hands-on learning and critical thinking may decline—particularly for early-career researchers. Maintaining human expertise and interpretability is essential.

Transparency and accountability are also critical. Scientific processes must remain explainable and auditable; black-box models can obfuscate reasoning, making it harder to validate results or assign responsibility when any errors

occur. Connected to this, bias remains a risk, as automated systems can inherit and amplify historical patterns. Without intervention, this could skew research focus and deepen inequalities. Similarly, with improper deployment, the use of automation could shift control over science to those who own the tools and infrastructure, reinforcing global disparities in research access.

As outlined in the *Stockholm Declaration on the Ethics of AI* by King et al. (2024), these systems must be designed and used responsibly, with transparency, fairness, and human values at the core. Automation should support science, not replace its human foundation.

2.3 Components of an Automated Scientist

Automating the science of biology is more complex than simply replacing manual labour with automated labour. It requires reshaping of the entire scientific workflow into a system that can operate with minimal human intervention (or at select places), while preserving rigour, transparency and robustness. From the initial hypothesis generation steps to the data analysis and sharing, each phase must be computable, scalable and auditable. We note that many, if not most, of these aspects are used in traditional human-led science, but improper implementation of them are exponentially more penalising in an automated context than a manual one. In the following sections, the functional components are explained and example applications are given. We will also discuss the contributions of this thesis in each step. More technical details regarding the different mentioned aspects can be found in Chapters 3 and 4.

2.3.1 Hypothesis Generation

At the core of the hypothetico-deductive method lies the generation and creation of hypotheses. An automated discovery system needs to be capable of proposing hypotheses that are not only understandable, but also scientifically meaningful. These hypotheses can be generated by, for example, leveraging existing knowledge, literature and prior data. For these hypotheses to be useful, they must be testable and preferably grounded in a transparent reasoning process, ideally one that can be reviewed and interpreted by human scientists.

Hypotheses can also exist at different levels of abstraction. At a high level, they may take the form of conceptual statements about relationships or causal influences (e.g., "the gene *SNF1* regulates carbohydrate metabolism"). At a lower level, they can specify precise, testable predictions, such as the expected change in concentration of a specific biochemical under a defined perturbation. An effective automated discovery system should be able to operate across this spectrum, selecting the appropriate level of abstraction for the available data, the completeness of the models, and the intended experimental strategy.

Early work in automated science already demonstrated this variety in abstraction levels. Some systems focused on learning highly specific, mechanistic rules from experimental data, while others generated broader, more conceptual hypotheses. Historical examples illustrate how both ends of this

spectrum have been successfully implemented. Expert systems like DENDRAL generated candidate molecular structures from mass-spectrometry data, and its partner program Meta-DENDRAL went a step further by inducing fragment-to-substructure rules directly from known structure-to-spectrum pairs; each learned rule an hypothesis about chemical fragmentation (Lindsay et al., 1993). More recently, the Robot Scientist "Adam" hypothesized general metabolic roles for orphan yeast genes by simulating knockouts in a metabolic model to predict required rescue metabolites and tested those predictions robotically (King et al., 2009).

These examples highlights two approaches to automated hypothesis generation:

1. Rule induction from data, where patterns or logical rules are mined directly from examples, and
2. Model-driven candidate mining, where a structured model defines the hypothesis space and guides prioritization.

In this thesis, we propose several methods of generating hypotheses bound by these concepts, as they are mainly based on mining literature priors from structured databases to produce rules or extracted from domain-specific models and later prioritised and weighted using empirical data. Paper 1 produced more abstract hypotheses, in the form of regulators likely to be involved in a biological phenomena by assessing uncertainties in their growth rate. In Papers 2 and 5 we apply pattern mining to generate hypothesis bodies, and apply supervised learning to ground them in biological observables (providing the bodies with a testable implication). Likewise, in Paper 4 we embed existing priors using ontology embeddings, and extract viable hypotheses based on data from genetic interactions through explainable AI (XAI) techniques. Both of the latter strategies ensure that the hypotheses themselves follow an interpretable structure, as they are shaped from human-defined, domain-specific ontologies.

2.3.2 Experiment Selection and Planning

Once a hypothesis is generated, the system itself must be able to determine how to test it. This involves selecting (and potentially prioritizing) experiments based on available resources, potential information gain, and feasibility. The planning process should be scalable, accommodate high-throughput experimentation, and incorporate as many data modalities as are needed to, with sufficient confidence, answer the hypothesis. Crucially, this should be a transparent process. Typically, most of the prior planning is designed by humans, but there have been examples where an agent is tasked with designing the experiment, such as in Song et al. (2025) and Williams et al. (2015).

One way to look at this is through the lens of information gain. In information theory, the information content of an outcome x_i is given by

$$I(x_i) = -\log_b p(x_i),$$

where $p(x_i)$ is the probability of observing x_i under the current paradigm.

As such, outcomes that are rare or unexpected have higher information content. From this perspective, one might aim to prioritise experiments with the highest expected information gain (i.e. the expectation of $-\log p(x)$ when taken with respect to some predictive distribution over outcomes $p(x)$), as they offer the most efficient route to reducing uncertainty. The exact implementation is an open question, and can vary from heuristic approaches to fully formalised selection criteria. Regardless, the main principle stays the same: choose experiments that are as informative as possible.

Another important consideration is cost. Even highly informative experiments may be impractical if they require excessive amounts of resources and time. In its simplest form, the cost of an experiment can be represented abstractly as, for example:

$$C(\text{experiment}) = x_1(\text{monetary cost}) + x_2(\text{time spent}) + x_3(\text{resource use}) + \dots$$

One would likely want to perform an experiment that would minimize the cost (C), but at the same time maximize the information content (I). A straightforward conceptual implementation of this weighting could, for example, be:

$$\max_{e \in E} \frac{I(e)}{C(e)},$$

where e is a candidate experiment, $I(e)$ its expected information content, and $C(e)$ its cost.

The key idea being that an automated discovery system should consider both the potential information gain and available resources, especially in high-throughput settings where potential monetary costs could stack up dramatically.

In this thesis, the scales are typically small enough to not warrant full deployment of these aspects. However, they are implicit in many of the selection steps. Such as for Paper 1, where uncertainties in growth rates are used to predict and select candidate genotypes. These factors are also "softly" implemented in Paper 5 through interaction with a large language model, albeit through prompting techniques (e.g., *"select the hypothesis likely to be the most informative"*). More details on on this can be found in Chapter 4.

Experimental design

Beyond selecting which experiments to run, the planning process must also determine how they will be structured to minimise bias and noise. This could include applying established experimental design strategies such as randomisation, blocking, or other types layout optimisation. For example, Latin square or similar arrangements can be used in plate-based assays to control for position effects and confounding variables (King et al., 2009). Randomising sample positions helps mitigate systematic biases, while blocking could allow known sources of variation to be accounted for explicitly. Incorporating these strategies at the planning stage ensures that experiments are informative, robust and reproducible. In an automated setting (in the case of iterative experimentation), these safeguards are especially important in order to prevent the system from

pursuing theoretically optimal but statistically fragile experiments (i.e., locally optimal). Aspects of this are included in Papers 1, 4 and 5.

2.3.3 Agency & Execution of Experiments

Translating an experimental plan into a real-world setting requires reliable physical automation and high degrees of standardisation. The system should be able to perform complex experimental protocols in a reliable and error-free way whilst keeping high fidelity to the selected hypothesis and experimental plan. It must also eliminate human variability, ensuring that results are not only repeatable, but also reproducible (Roper et al., 2022). Additionally, it should capture comprehensive metadata and runtime logs regarding the process, ensuring auditability in case of failure.

The agency of an automated system (i.e., the tools and techniques available to it) directly shapes both the experimental design and hypothesis selection. For example, a system equipped only for liquid handling and incubation will be restricted to experimental strategies compatible with those capabilities, whereas a platform that can also perform automated metabolomics or genetic interventions opens up a much wider design space. Consequently, knowing the systems capabilities in advance allows planning algorithms to propose hypotheses and experiments that are not only scientifically relevant, but also physically executable without extensive manual intervention.

As part of the work done in this thesis, we have mainly created and used custom automation solutions. Papers 1, 4, and 5 have made use of the automated laboratory cell Eve (Williams et al., 2015), enabling basic experimental measures such as liquid handling and cultivation. Additionally, as part of Paper 3, the main product was software enabling the full automation of metabolomics acquisition, making use of partially existing commercial automation software, but also bespoke orchestration to allow for the full process to be automated. Paper 5 combined all of these processes, utilising standardised protocols for both liquid handling and cultivation, whilst making use of developed software and hardware for automatic metabolomics processing and analysis, further increasing agency.

2.3.4 Data Analysis and Integration

Following experimentation, the system must have the capacity to capture and analyse data in a scientifically meaningful way. This includes the ability to handle complex multidimensional datasets, preferably containing many types of biological data modalities (see Chapter 3), or more simply, enough modalities to answer the hypothesis with sufficient confidence. The analysis should extract insights that remain faithful to the original formulation of the hypothesis, whilst being robust, transparent, and reproducible. Results should be human-interpretable and shared openly, along with the underlying empirical data and all metadata regarding the experiment.

A key enabler for such transparency and interoperability is the use of ontologies—formal, shared vocabularies that define concepts and relationships

in a specific domain. By representing both data and metadata using established ontologies, results from different experiments, laboratories, or even organisms can be more easily compared and integrated. Ontologies also allow computational reasoning tools to link experimental outcomes back to prior knowledge, helping to uncover hidden relationships or inconsistencies. More details regarding ontologies will be covered in Chapter 3. For automated systems, this common "language" is of the utmost importance, as it connects hypothesis formulation, experiment execution, and analysis, ensuring that each stage of the scientific cycle can be traced, interpreted, and re-used by both humans and machines.

In this thesis, Papers 1 and 5, and to some extent Paper 3, implemented concepts of automated data acquisition and analysis. Furthermore, Paper 5 involved the construction of a graph database that incorporated ontological annotations for hypotheses, experimental components, and results, enabling both rich querying and interpretability.

Chapter 3

Systems Biology as a Beneficiary of Automation

In the previous chapter, we explored how automation can help address the scale, complexity, and reproducibility challenges of modern science. These issues are especially apparent in systems biology and its related subdomains (Baker, 2016; Roper et al., 2022). The field's very definition makes it both a suitable candidate for, and a major driver of, advances in scientific automation.

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 complementary concept to reductionist biology, driven by the need to integrate data from diverse sources and levels of cellular organization. The goal is to forge a holistic understanding of biological systems (Regenmortel, 2004). Systems biology aims to build models that capture the behaviour of biological systems and predict their responses to perturbations across a variety of conditions (Tavassoly et al., 2018). The approach has revolutionized our understanding of biology and accelerated the development of new biotechnologies (Nielsen et al., 2008). Moreover, systems biology approaches are essential for addressing some of the most pressing challenges in biology today, such as understanding the mechanisms of common diseases and devising strategies to combat cancer (R. Chen et al., 2012; Loscalzo et al., 2011). However, this holistic view of a system comes with challenges. Reductionist biology can be be achieved with localised modelling and data collection, but understanding cells at a systems level requires biological data and modelling at a massive scale. Not only are even the simplest of organisms incredibly complex, the techniques and instrumentation needed to measure their outputs have severe limitations.

In this chapter, we focus on these challenges in the context of *Saccharomyces cerevisiae*, the model organism used throughout this thesis. We will examine why it is ideally suited for studying systems biology at scale, explore the nature of cellular complexity, discuss how controlled perturbations reveal biological roles, review common biological readouts, outline strategies for integrative data analysis and assess computational frameworks and representations.

3.1 Why *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 (Duan et al., 2018). The ease of cultivation and overall resilience of *S. cerevisiae*, combined with the fact that it shares many fundamental biological processes with higher organisms, including cell cycle regulation and core metabolic pathways, 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 (Giaever et al., 2002; Z. Yang et al., 2020). These traits makes it ideal for systems biology, which depends on model organisms that can be systematically perturbed, quantitatively measured, and computationally modelled across scales. This ultimately caused it to be the first eukaryote to have its genome sequenced in 1996 (Goffeau et al., 1996; Botstein et al., 2011). Despite the increasing use of mammalian and multicellular systems in research, *S. cerevisiae* remains uniquely positioned for systems-level investigation due to its genetic tractability, well-annotated genome, and ability to generate reproducible, high-throughput data at low cost. It is also still widely used for designing bio-factory platforms for various industrial uses such as pharmaceuticals, food additives and biofuels (Nielsen et al., 2008; Hong et al., 2012; C. Zhang et al., 2024).

As a result, it has been the premier platform for functional discovery of genes in eukaryotes for many years. Because of this and the relative ease of genetic manipulation, early efforts were focused on creating genome-wide yeast deletion mutant collections (Giaever et al., 2002; Ea et al., 1999). These collections are comprised of large libraries of cells which have 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. Combined with its rapid growth, robustness and low cultivation cost, this enabled massively parallel experimental designs that are still unmatched in throughput and comprehensiveness by more complex organisms. These deletion libraries, along with their associated metadata, form a foundational dataset for many of the analyses and methodologies used in this thesis, and enabled systematic exploration of genotype–phenotype relationships.

Taken together, *S. cerevisiae* offers a uniquely powerful platform for bridging empirical experimentation with computational modelling—making it an indispensable organism for systems biology.

In the following sections, we move from the rationale behind choosing *S. cerevisiae* to the broader biological questions it helps us address: How do cells function as integrated systems? How can we perturb them to infer function? And how can computational tools help us make sense of the resulting complexity?

3.2 Understanding Cellular Complexity

Cells are constructed from nested tiers of biological organization: at the base lie individual molecules—genes encoded in DNA, their RNA transcripts, the proteins they produce, and the metabolites that fuel reactions (Alberts et al., 2007; Haas et al., 2017). The activity of these molecules is dynamically regulated through networks of gene expression, signalling, and feedback (typically comprised of other proteins and metabolites), which together coordinate metabolic pathways and afford cells the ability to react to environmental cues. These molecules can assemble into macromolecular complexes, such as collections of enzymes that catalyse and channel substrates through metabolic pathways. These components can additionally be formed into even larger scale organizational structures such as ribosomes, or even organelles like the mitochondria. All of these building blocks and aggregated modules interact with each other in highly complex spatio-temporal ways. Together, this multi-layered architecture enables the cell to grow, divide, and respond to environmental factors. Each component, regardless of hierarchy, may influence and depend on many others. Understanding each component, big or small, and how they fit into the entire system is one of the central challenges in biology.

3.2.1 The Importance of Functional Genomics

Functional genomics refers to the process of identifying and characterizing the function of specific biological molecules or subsystems, such as genes, proteins and metabolic pathways. 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.

As biological systems are too complex to infer function by intuition or static observation alone, functional genomics uses systematic, large-scale experimental and computational tools to dissect how molecular elements contribute to cellular behaviour. Modern functional genomics combines experimental techniques such as transcriptomics, proteomics, and metabolomics with computational analysis to build models of cellular function (Haas et al., 2017).

Experiments are often designed to test responses to defined perturbations, such as changes in genetic background or nutrient availability. Additionally, biological systems are inherently noisy and subject to environmental variability, yet they maintain remarkable stability. By observing how the system behaves under these controlled changes in conditions, researchers can begin to slowly understand biological function and its intricacies.

3.2.2 Probing Complexity through Perturbation

Given the immense complexity of biological systems, a key strategy for uncovering how they work is through systematic perturbation—deliberately altering specific components, such as knocking out a gene, inhibiting a protein, or

changing environmental conditions, and observing how the system responds. This approach is foundational to biology because it enables us to infer the role of individual components by disrupting them and analysing the resulting changes in phenotype (observable traits) or molecular behaviour.

A useful analogy could be to think of a cell as a car (a machine composed of many interdependent parts). Imagine that you have never seen a car before and have no manual to explain it and its components. One approach to uncover how it functions would be to remove or disable parts one by one and try to drive it. If removing the battery prevents it from starting, you learn something about the battery's role. If taking off the muffler only makes the car louder, you could learn that it is not essential for movement. This is analogous to knocking out genes or inhibiting proteins in a cell to reveal their functions.

Of course, some parts only reveal their importance in combination with others or under specific conditions. A steering wheel is useless without actual wheels, and a car's radiator might seem irrelevant in cool weather, but becomes essential during hot weather. Likewise, some genes show no effect when perturbed alone but are critical when another gene is also disrupted (such as for the example given in Fig. 3.1), or when the organism is under stress. These interactions are key to understanding how biological systems maintain robustness and adapt to their environments.

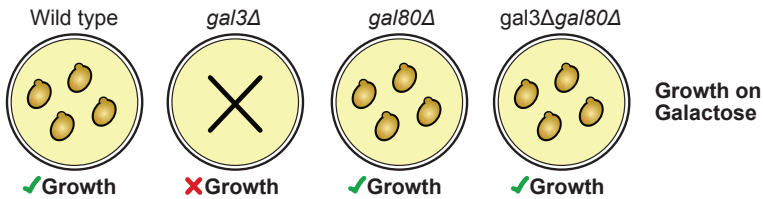


Figure 3.1: Example of a phenotypic suppression between two genes for *S. cerevisiae* growing on galactose. Deleting only the *GAL3* gene causes a severe growth defect, deleting *GAL80* alone has little to no effect but deleting both suppresses the phenotype induced by the *GAL3* deletion, allowing it to grow on galactose.

However, the number of possible perturbation combinations increases rapidly with system size, making exhaustive testing infeasible, as even simpler systems like *S. cerevisiae* has thousands upon thousands of components. This is why experimental designs often prioritize likely interactions, and why computational tools and comprehensive biological readouts are essential to guide and interpret perturbation-based studies.

3.3 Biological Data

What we can learn about biological systems is constrained by what we can observe. Individual cells and their constituent parts exist on a dimensional and temporal scale very unlike our own intuitive concept of time and space, such as in nanometers and microseconds (Heim et al., 2017). Technology is rapidly improving, and with it our potential to understand of biology improves. The science of collecting biological data has given rise to several subdomains of biological science, collectively called "omics".

"Omics" is a 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 and representations of biological systems. Together they represent a type of flow of information through biological systems, as illustrated in Figure 3.3a and Figure 3.3b. By integrating several different types of omics, one can achieve a much more holistic understanding of the biological system in question (Haas et al., 2017; Karczewski et al., 2018).

Genomics typically 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. Note that there are many other types of omics, such as fluxomics and interactomics, that are tangentially relevant, but not directly covered in this thesis.

3.3.1 Challenges in Observability

A fundamental challenge in biological research is that our insights are constrained by what we can observe and quantify. If we cannot measure it, we cannot rigorously analyse or model it. Many processes occur at spatial scales (e.g., intracellular interactions across organelles) or temporal scales (e.g., biochemical reactions or conformational changes in proteins) that are beyond the capacity of current instruments.

Furthermore, living systems are notoriously heterogeneous and dynamic: sampling a few cells or taking a snapshot in time may miss crucial variability or transient events such as cell cycle changes. Even when we can collect data, processing and measuring can by themselves perturb these systems, such as when one takes a sample out of the incubator for sequencing, inadvertently altering its temperature, oxygen levels, and subsequently inducing a stress response. These experimental interventions, though necessary, can introduce artifacts or mask native behaviours. This raises the very real challenge of distinguishing true biological signals from those induced by the act of measurement itself. As a result, designing minimally invasive, high-resolution, and temporally sensitive experimental methods is essential for advancing our ability to capture the true dynamics of living systems. Difficulties in observability are not only bottlenecked by experimental measures, but also on its analysis. Modern data collection techniques used today produce vast amounts of data, from high-throughput sequencing to real-time imaging, that pose their own

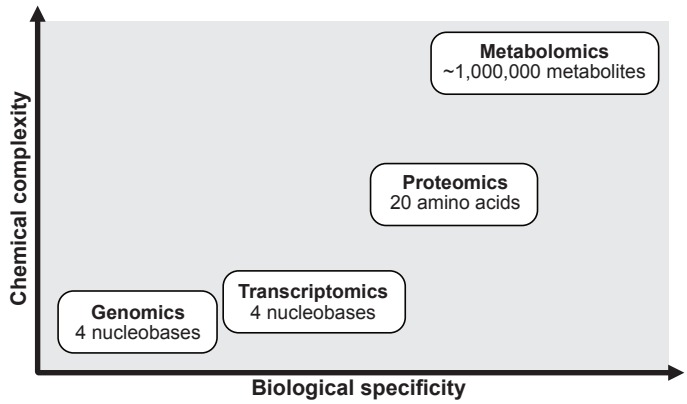


Figure 3.2: Plot illustrating the relationship between chemical complexity and biological specificity across different omics layers. At the bottom, genomics is based on just four nucleotide bases, offering a relatively chemically simple foundation, but with limited direct specificity to cellular state. Transcriptomics, while also built from four bases, adds biological specificity through expression dynamics. Proteomics increases both specificity and complexity, typically being constructed from 20 amino acids, involving diverse post-translational modifications, and dynamic abundances that more directly reflect cellular functions. At the top, metabolomics exhibits the greatest chemical complexity, encompassing a vast and diverse array of small molecules with millions of possible conformations. At the same time, metabolic profiles often provide the most immediate and specific readouts of cellular physiology, linking genotype, environment, and phenotype (Dettmer et al., 2007).

computational and statistical hurdles. As a result, blind spots remain in our understanding of cellular behaviour simply because what we wish to study lies beyond the limits of our experimental capacity and analytical frameworks.

Each of these omics subdomains contain their own sets of problems, each with their own advantages and disadvantages.

3.3.2 Metabolomics

Metabolomics is the study of small molecules called metabolites in a biological system. It typically is used to provide insights into the biochemical pathways and cellular processes that govern metabolism by studying the products and substrates of biochemical reactions. Metabolomics can aid in topics such as identifying biomarkers for disease, elucidating the nature of metabolic pathways, or study responses to environmental factors. It is typically seen as the type of data most closely representing the phenotype (observable state) of the organism (Dettmer et al., 2007). This also means that is a highly volatile, meaning that while it is a highly informative measure, it is also extremely prone to heteroscedastic variation, where the amount of variability depends on the scale or conditions of measurement.

Metabolites are typically identified and quantified using advanced analytical techniques, such as mass spectrometry (Alseekh et al., 2021). 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 (Glish et al., 2003). It typically utilizes several orthogonal separation methods prior to mass analysis to maximize data quality and confidence in metabolite identification. Liquid chromatography (used in Paper 1) could be used to further separate molecules by their chemical traits, such as hydrophilicity and polarity, sharpening chromatographic peaks and potentially mitigate ion suppression (adverse effect on response due to reduced ionisation efficiency) (Harrieder et al., 2022). Additionally, an increasingly widespread technique is ion mobility (used in Papers 3 and 5), which introduces an extra degree of separation based on cross collisional sections (CCS). It essentially separates ions based on their mobility in some type of ideally behaving gas, meaning that this separation metric is typically influenced by their size and shape (Lanucara et al., 2014; Paglia et al., 2022).

Broadly speaking, 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 (Pinu et al., 2017). Intracellular metabolomics typically describes the internal concentrations of metabolites inside the cell, which are involved in various molecular processes governing the cells' functions (A. Zhang et al., 2013).

Additionally, when studying metabolomics through mass spectrometry, it is generally approached in either a targeted (hypothesis-driven) or untargeted (exploratory) manner—or a combination of both. 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 analyse 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 (linking the readout of the machine to an actual biochemical) is not a trivial task, as explained in Monge et al. (2019).

Several papers in this thesis explore different aspects of metabolomics. Paper 1 utilises metabolomics to evaluate regulatory models in an active learning-based setting and extract phenotypic insights from the diauxic shift. Paper 3 is about the automation of data acquisition in metabolomics, specifically regarding ion mobility based mass spectrometry. Paper 5 utilises it to generate and validate biological hypotheses in an automated setting.

3.3.3 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 (Alberts et al., 2007). 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 the previously mentioned mass spectrometry (Messner et al., 2022). It is a field which is much more mature than many other types of omics, such as metabolomics.

High-throughput quantification of proteins has historically been time-consuming, difficult and expensive. However, during the last decade, mass spectrometry-based proteomics has made considerable progress, and it is increasingly able to facilitate biological experiments at scale (Messner et al., 2022; Messner et al., 2023). This is enabling close to genome-wide coverage, measuring many of the organism's proteins, in a high-throughput and relatively inexpensive manner.

Paper 2 utilizes genome-wide proteomic abundances to train supervised machine learning models to evaluate systematized knowledge on *S. cerevisiae* and generate biological hypotheses in a high-throughput manner.

3.3.4 Integrative Analysis

A key methodology in systems biology is integrative analysis: combining several different experimental readouts or levels of biological organization to gain a more holistic understanding of the system in its entirety. By linking observations from multiple omics layers (such as genomics, transcriptomics, proteomics, and metabolomics) one can increase predictive power, identify causal mechanisms, and generate richer, more targeted hypotheses than would be possible from any single data type alone. Despite its promise, integrative analysis faces a number of obstacles:

- **Heterogeneity & quality:** Biological data modalities differ significantly in acquisition, protocols, units, and noise characteristics (even from the same general levels of organization).
- **Dimensionality & scalability:** Adding more data types can dramatically expand the feature space, increasing computational demands and reducing interpretability.
- **Temporal alignment:** Different modalities capture processes on different biological timescales (e.g., shifts in metabolite concentrations occur faster than transcriptional changes), complicating direct integration.
- **Biological relevance & interpretability:** Condition-dependent behaviours require careful contextualization, and integrated models must remain interpretable for domain experts.

- **Missing data & incomplete coverage:** Each modality has unique detection limits and biases, leading to gaps that can propagate through the integration process if not handled carefully.

These issues become even more critical in automated discovery, where the system must be able to reason over multi-modal data with minimal human intervention or intuition.

Common Strategies

The choice of integration method depends heavily on the frameworks, algorithms, and data structures in use. Common strategies include statistical correlation, network-based integration, machine learning models, and graph-based approaches such as knowledge graphs (Gligorijević et al., 2015). Different omics types are also suited for different types of investigations. As visualized in Figure 3.3b, it is crucial to match the data type to the biological question of interest. For example, when investigating metabolism, metabolomics and proteomics together provide robust readouts of the current metabolic state (metabolites) and the effectors of change (proteins).

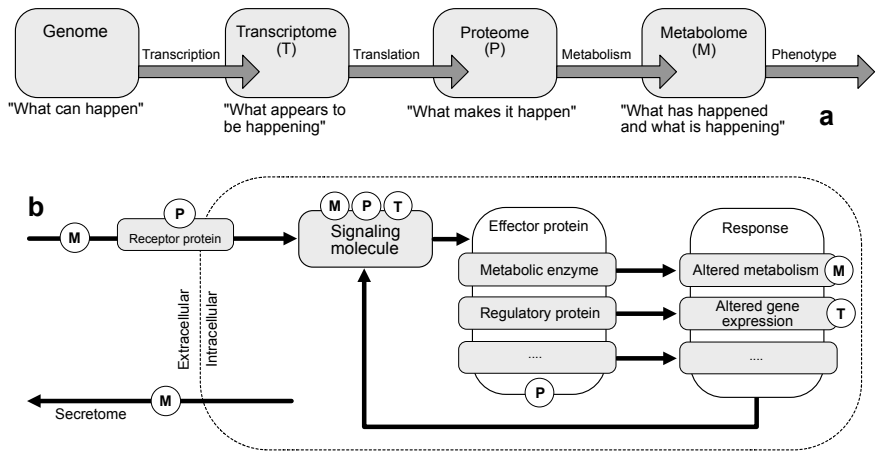


Figure 3.3: Molecular readouts and their roles. a. The "Omics-cascade". Simplified description of the different types of data and levels of organization that could be used to describe the response of biological systems to perturbation (e.g. disease or environmental) (Dettmer et al., 2007). **b.** Simplified 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.

Applications in this Thesis
This thesis applies several approaches to integrative analysis, often using structural knowledge priors to improve interpretability. These concepts will be explained more thoroughly in the next section.

- **Paper 1:** Computational experiment selection using a combined signalling and regulatory network integrated with a genome-scale metabolic model (Coutant et al., 2019). Untargeted metabolomics was used to capture the current metabolic state, then contextualized using curated metabolic networks via topological enrichment, enabling biological interpretation and inference of indirect gene deletion effects.
- **Papers 2 and 5:** Aggregation of multiple organizational levels (e.g., protein interactions, metabolite concentrations, protein abundances) and structure data (e.g., the Gene Ontology, Ascomycete Phenotype Ontology) into a unified logic-program formalism to predict and hypothesise about proteomic and metabolomic states. Paper 2 focused on the proteome; Paper 5 extended the approach to produce testable hypotheses, answerable through phenomics and metabolomics.
- **Paper 4:** Alternative integration via ontological embeddings and knowledge-graphs, representing different datatypes in a more qualitative manner to predict phenotypic outcomes of genetic perturbations.

In the context of automation, integrative analysis is particularly useful when data and models are expressed in machine-readable, semantically consistent formats. This enables automated systems not only to combine heterogeneous data at scale, but also to reason over it and assist in experiment selection and analysis.

3.4 From Biology to Computation

Understanding the complexity of biological systems requires more than just experimental data—it demands abstraction, formalization, and the ability to simulate and reason about system behavior. As the scale and resolution of biological measurements have increased, so too has the need for computational tools that can integrate diverse datasets, extract meaningful patterns, and generate testable predictions. This shift from purely descriptive biology to data-driven, model-based inquiry is a hallmark of modern biology.

The challenge lies in bridging the gap between raw biological complexity and structured computational formalisms. Biological processes are usually non-linear, happens across several levels of hierarchies of organization, and are more often than not context-dependent, making them difficult to represent or simulate without simplification.

This section explores some of the approaches that have been used to enable this translation. From mathematical and statistical modelling techniques to formal knowledge representation methods that allow the encoding, sharing, and automated reasoning across biological knowledge.

3.4.1 Representing Metabolism

Metabolism refers to the set of biochemical processes that occur within living organisms, encompassing the reactions that convert nutrients into energy and generate 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 (M. Kanehisa et al., 2000; Minoru Kanehisa et al., 2023; Minoru Kanehisa, 2019; Gillespie et al., 2022; Karp et al., 2019). These models are typically subdivided into metabolic pathways—modules which perform some sort of localized task in the cell. Examples could encompass catabolic (degradative) pathways like glycolysis or anabolic (biosynthetic) pathways such as amino acid biosynthesis. These representations are highly amenable to computational methods and techniques.

Genome-Scale Metabolic Models (GEMs)
Genome-Scale Metabolic Models are computational reconstructions of the complete metabolic reaction space of an organism. They integrate genomic, biochemical, and physiological data to produce a stoichiometrically consistent network, where reactions are connected to their catalyzing enzymes and the corresponding genes via Gene–Protein–Reaction (GPR) rules. They provide a formal framework for simulating and analysing metabolic activity under various

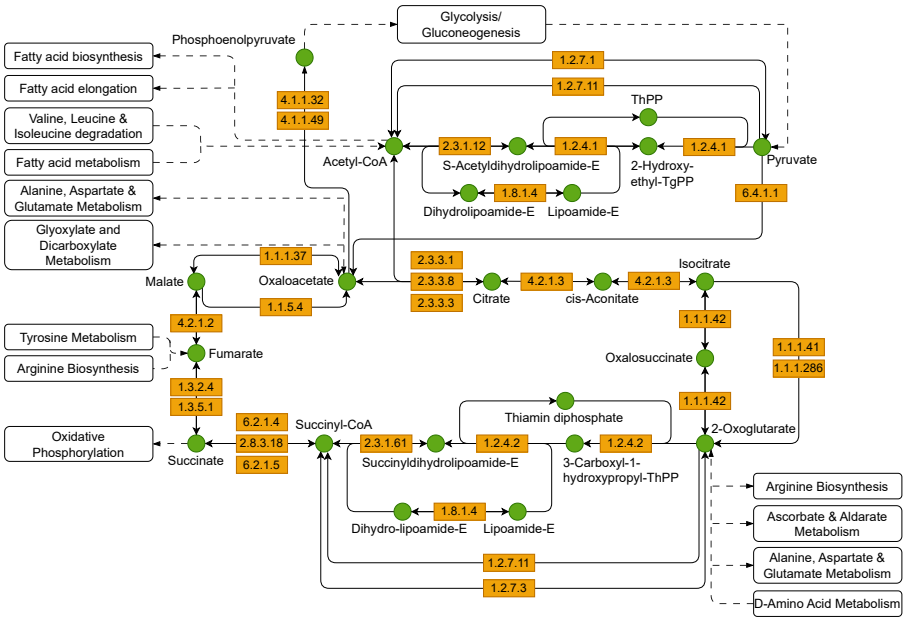


Figure 3.4: Pathway representation (as a directed graph) of the citric acid cycle in *Saccharomyces cerevisiae*. Green circles mark involved metabolites, orange squares represent reactions (via the involved enzyme) and the white squares denote interconnected pathways. Pathway information retrieved from KEGG (2023-08-02) (M. Kanehisa et al., 2000).

conditions (Orth et al., 2010; Mo et al., 2009; C. Zhang et al., 2024). A GEM is represented through, among others, the following concepts:

- S, the stoichiometric matrix, of dimensions $m \times n$ (metabolites $\times$ reactions)
- v, the reaction fluxes (usually in the form of a vector of length n)
- c, an optimization objective (typically biomass or the accumulation or excretion of a specific biochemical).
- lb, ub, lower and upper bounds of fluxes (vectors that decides reaction direction and overall flux capacity of reaction fluxes in v).

GEMs support constraint-based modelling, enabling simulation of metabolic states without explicitly requiring detailed kinetic parameters, which are often unavailable at genome scale—although there exists modelling regimes that incorporate these constraints as well, such as in F. Li et al. (2022).

Flux Balance Analysis (FBA)

This type of structure enables simulation of metabolic activity, and is typically achieved through the use of methodologies such as flux balance analysis (FBA).

FBA is a simulation technique that seeks to model the cell by simulating the flow of metabolites through a GEM under various conditions. This can, for example, enable predictions of growth rates and specific metabolite production rates based on nutrient and environmental conditions (Orth et al., 2010).

One of the main design choices made for overall tractability, is that it assumes a steady state for intracellular metabolites (no internal accumulation of biochemicals), as in $Sv = 0$. Reaction fluxes (flow of mass) are additionally bound the lower and upper bounds, enabling the solving the fluxes through linear programming, i.e.:

$$\max_{v} c^T v \ s.t. \ Sv = 0, \ lb \leq v \leq ub \tag{3.1}$$

Note that the methodology is not without its limitations. The steady-state assumption is severely limiting, as it is not a realistic reflection of intracellular states. Additionally, the framework does not typically account for regulatory effects (such as transcription or phosphorylation, see Fig. 3.3). This concept can however be extended through manipulation of lower and upper reaction bounds when the acting enzyme is exposed to regulatory effects, which could be simulated, as in Coutant et al. (2019). This methodology was heavily used in Paper 1, both for experiment selection and model validation.

Abstracting GEMs for reasoning
Beyond their use for direct metabolic simulation, GEMs can straightforwardly be transformed into higher-level graph structures that link metabolism to other biological processes. For example:

- **Reaction–gene graphs:** where reactions are linked to the genes encoding their catalyzing enzymes, as shown in Fig. 3.5B.
- **Gene–metabolite graphs:** which connect gene products to metabolic changes through intermediate reactions, as seen in Fig. 3.5C.

These abstractions are especially useful when connecting changes in genotype to phenotype. In Paper 1 and its associated work, it used to connect regulatory perturbations to predicted flux changes and growth phenotypes. Additionally, in Papers 2 and 5, the metabolic relationships are encoded as relational facts, enabling structured reasoning across gene-to-metabolite relations. In these cases, these abstractions allow metabolic information to be combined with otherwise heterogeneous datasets, enabling richer systems-level analyses.

3.4.2 Gene Regulation

A gene regulatory network (GRN) describes how genes, their products, and regulatory elements interact to control when and how strongly genes are expressed. In essence, it maps the cellular decision-making layer: transcription factors can activate or repress specific targets, kinases can modulate protein activity post-translationally, and signalling pathways can cascade regulatory effects across multiple genes (Lee et al., 2002; Coutant et al., 2019).

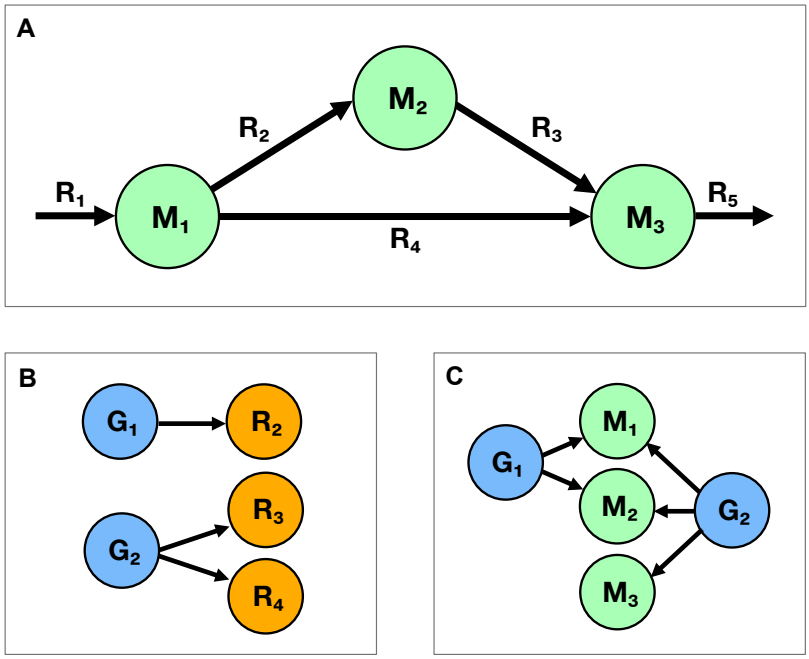


Figure 3.5: Metabolic network along with examples of network structures in GEMs. **A.** Toy metabolic, where M_1, M_2 and M_3 correspond to metabolites, and R_1-R_5 denote metabolic reactions. The reactions are: $R_1 :\rightarrow M_1$ (import), $R_2 : M_1 \rightarrow M_2$, $R_3 : M_2 \rightarrow M_3$, $R_4 : M_1 \rightarrow M_3$, $R_5 : M_3 \rightarrow$ (export) **B.** Reaction-gene graph for exemplified network. G_1 and G_2 denote genes whose gene product catalyzes reactions. **C.** Metabolic graph, demonstrating the connections between genes and metabolites (through reactions).

GRNs are typically represented as directed graphs in which nodes denote genes or gene products, and edges indicate regulatory relationships such as activation, repression, or modulation. These can be inferred from data using approaches like Bayesian networks, mutual-information methods, time-series models (e.g. dynamic Bayesian networks, DBNs), or large-scale deep learning frameworks (Margolin et al., 2006; Murphy, 2002; Z. Li et al., 2023).

In this thesis, GRNs provide a way to connect regulatory changes to metabolic consequences. In Paper 1, a DBN of the diauxic shift was coupled to a genome-scale metabolic model to simulate the impact of gene deletions on metabolic fluxes and phenotypes. In Paper 2, regulatory relationships were encoded as logical propositions, enabling integration with metabolic abstractions for reasoning across genotype–phenotype links.

3.4.3 Structured Knowledge Representation

Representing biological knowledge in a structured and "computable" form is essential for enabling integration, automated reasoning, and large-scale

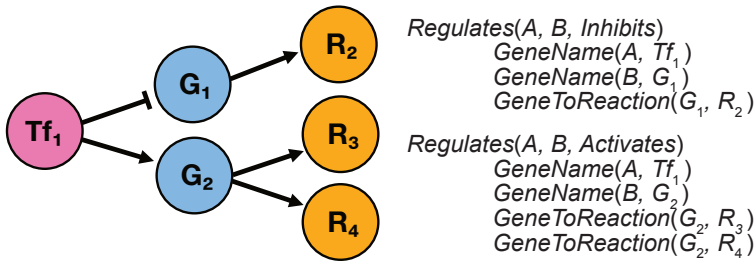


Figure 3.6: Simplified graphical example of regulation-metabolism integration. Tf_1 is an example transcription factor activating G_1 and repressing G_2, both of which in turn catalyse reactions R_2-R_4. Note, that in the context of the toy network presented in Fig. 3.5, higher abundance of this transcription factor could force the network to take an alternative path to the end product (via R_4). Also shown are example propositional descriptions of the toy regulatory network (starting from the transcription factor), similar to the abstractions used in Papers 2 and 5.

analysis. Biological systems are complex, and functional knowledge about genes, phenotypes, metabolites, and interactions must be usable across different datasets, tools, and domains. Ontology-based representations (structured vocabularies with formally defined relationships) are a widely adopted solution, as they provide semantic interoperability and organization of concepts. Below are a descriptions of a few of the ontologies used in this work.

The Gene Ontology (GO)

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 (Ashburner et al., 2000):

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**, a category describing 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,

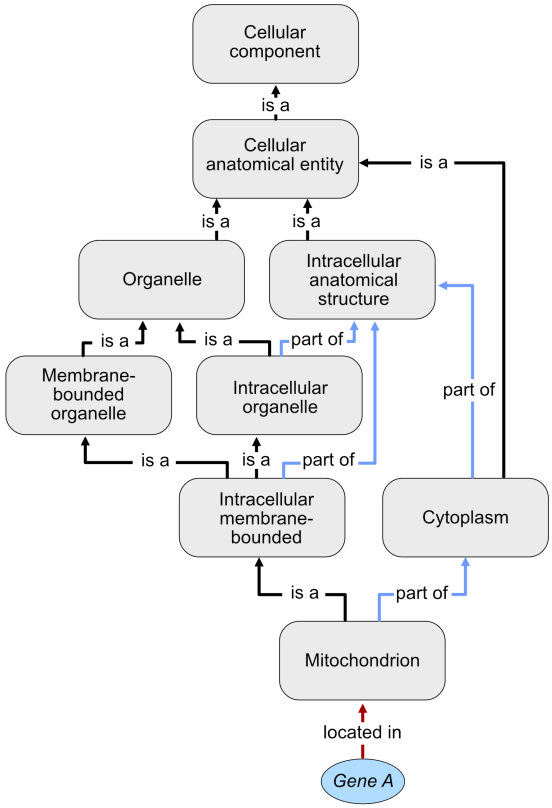


Figure 3.7: Example ancestor chart for the "mitochondrion" gene ontology term (in the cellular compartment category). Each arrow corresponds to a different semantically meaningful relation with the parent term. The hierarchical structure allows for classification of gene function at different levels of specificity.

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.

Other ontologies and resources

Beyond GO, several complementary ontologies and structured databases exist to capture other facets of molecular and cellular function, phenotype, and interaction. For example, an ontology that is heavily used in this thesis is the Ascomycete Phenotype Ontology (APO), which formalizes observed yeast phenotypes—such as growth defects, colony morphology, and chemical sensitivities—using a controlled vocabulary (Cherry et al., 2012; Engel et al., 2025).

The Chemical Entities of Biological Interest (ChEBI) ontology provides standardized identifiers and hierarchical classification for small molecules and ions encountered in biology (Hastings et al., 2016). To link proteins both to

each other and to small molecules, resources like STRING-DB curate and score protein–protein associations from experimental data and prediction methods, while STITCH-DB focuses on protein–chemical interactions, both with interoperable descriptions of the interactions themselves (Szklarczyk et al., 2023; Szklarczyk et al., 2016). Together, these ontologies and interaction databases enable rich, multilayered knowledge graphs that have the potential to assist in systems-level analyses of gene function and phenotype.

While not used in this thesis, there are many different widely used biological ontologies, such as KEGG Brite and Panther (Mi et al., 2013; Minoru Kanehisa et al., 2023).

Beyond domain-specific resources, there are also more general ontologies for representing the scientific process itself. The Ontology for Biomedical Investigations (OBI) defines terms for describing many aspects of an investigation such as protocols, instrumentation, data transformations, and analysis methods, facilitating annotation and integration of experimental workflows (Bandrowski et al., 2016).

Many of the aforementioned ontologies were used in Papers 2, 4 and 5, either as instantiations of genes, or to describe and annotate biological hypotheses and experimental investigations.

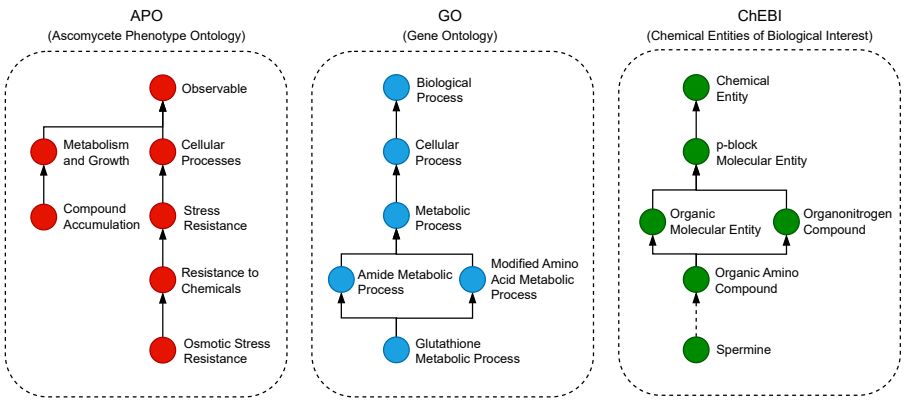


Figure 3.8: Simplified examples of ontologies used in this thesis. APO (Ascomycete Phenotype Ontology) standardizes the description of mutant phenotypes in fungal species. GO (Gene Ontology) describes functional aspects of gene products. ChEBI (Chemical Entities of Biological Interest) describe compounds relevant for biological processes.

Chapter 4

Machine Learning for Biological Discovery

The complexity of biology demands analytical approaches that go beyond traditional human intuition. Systems biology seeks to unravel the dynamics and interactions that give rise to cellular behaviour and emergent properties. High dimensionality, non-linear dynamics and relationships, context dependent regulation and overall data volume produce a setting in which interesting biological signals are buried underneath noise and complex interdependencies. These challenges quickly overwhelm traditional statistical and mathematical approaches, motivating the need for more advanced computational strategies.

Machine learning provides tools to detect complex patterns, produce accurate prediction models, and suggest testable hypotheses. By learning directly from experimental observations, these algorithms can detect subtle and complex properties of biological systems at a level of detail beyond human capability.

In this chapter, we introduce the core machine learning techniques that were used in this thesis, and illustrate how they can be used to aid in biological discovery and the automation of biological science.

4.1 Learning from Observations

4.1.1 Supervised Learning

Supervised learning is a fundamental concept in machine learning. It is a category of learning algorithms where labelled training data are used in order to learn. In this paradigm, a dataset typically consists of pairs of input samples (features) and their corresponding outputs—often called labels or targets. The goal is then typically to train a model (a learner) that can learn from the training data in order to make accurate predictions about the phenomena of choice. The model learns from the labelled examples by identifying patterns, relationships, or statistical dependencies between the input and output variables.

Formally, in supervised learning we are given a set of (n) training examples (input-output pairs), i.e. $\{(x_i, y_i)\}_{i=1}^{n}$, where each $x_i \in \mathcal{X}$ is an input (feature vector) and each $y_i \in \mathcal{Y}$ is an associated output. The goal is to then learn a function that can accurately map the relation between the two, such as $g : \mathcal{X} \to \mathcal{Y}$. In biological contexts, $\mathcal{X}$ could, for example, represent transcriptomic profiles, protein concentrations, or other types of descriptors, while $\mathcal{Y}$ might correspond to phenotypic outcomes such as metabolite accumulations or drug responses. The specific choice of $\mathcal{X}$ and $\mathcal{Y}$ will depend on the framing of the biological question.

Beyond prediction, supervised learning can be used as a tool for biological insight. By analysing how input features contribute to model performance, it enables us to identify which properties are most associated with the output label. This is particularly useful when features are biologically meaningful and interpretable because it allows the model to act not just as a predictor, but also as a hypothesis generator.

Supervised learning has been successfully used in biology for decades, and is now a staple in most bioinformatic workflows, with applications in for example:

- Classification of cancer through gene expression by using a classifier on microarray data to distinguish acute myeloid leukemia from acute lymphoblastic leukemia (Golub et al., 1999).
- Predicting clinical drug response from gene expression using regularised linear models on cancer cell-line panels and applied them to patient tumors to predict sensitivity to chemotherapy (Geeleher et al., 2014).
- Gaining mechanistic insight into antibiotic resistance using biochemical reaction data and regularised linear models to generate mechanistic, interpretable explanations of antibiotic resistance (J. H. Yang et al., 2019).

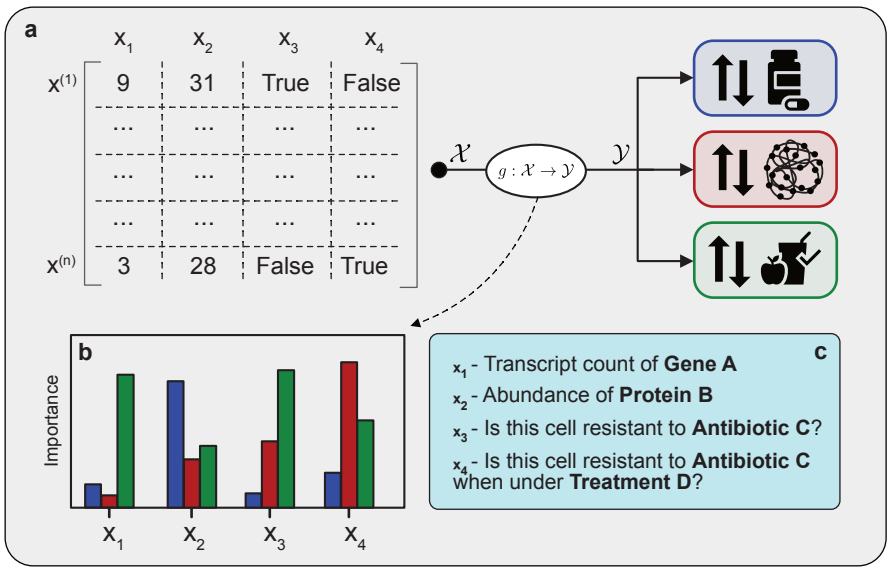


Figure 4.1: Supervised learning to extract testable features. a. General methodology behind supervised learning. Using an input $\mathcal{X}$ (e.g. gene counts) and a target label $\mathcal{Y}$ (e.g., drug resistance, protein levels or nutrient uptake), learn a function g that maps the relation between the two types of variables. **b.** Relative importances of features for different targets, extracted from the learner (g). For example, feature 4 (x_4) is very important for predicting protein levels (red). **c.** Description of features. Features can be continuous measurements, like RNA counts or protein abundances, but could also be represented as propositions. Propositions such as in x_3 and x_4 have been used extensively in this thesis. These could allow for more interpretable (and testable) multimodal features, e.g. a nutrient uptake could be commonly associated with antibiotic resistance, but during a conditional treatment a specific protein could have a stronger connection to the predicted outcome.

4.1.2 Explainable Machine Learning

Explainable AI (XAI) seeks to make machine-learning decisions transparent and trustworthy by revealing how input features contribute to model outputs. In biology, interpretability is essential for turning predictions into hypotheses, mechanisms, or actionable insights. Unlike purely predictive settings, biological research often seeks to explain the underlying processes, identify causal factors, and potentially even guide further experimentation. An interpretable model could allow researchers to trace predictions back to specific features, biological pathways, or relevant interactions, thereby transforming statistical associations into biologically meaningful hypotheses.

In this thesis, we make use of several different types of learners (with varying degrees of interpretability). Below, we distinguish between models that are intrinsically interpretable (where the model itself serves as its own explanation) and post-hoc explanation techniques that approximate the behaviour of less interpretable learners.

Explainable Models

In the simplest case—a linear regression model—interpretation is direct. For an input vector $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})$, the model predicts:

$$\hat{y}^{(j)} = \beta_0 + \sum_{i=1}^{m} \beta_i x_i^{(j)}. \tag{4.1}$$

where β_i quantifies the effect of feature $x_i^{(j)}$ on the prediction for sample j. The value of β_i reflects the strength of the association, while its sign indicates the direction (i.e., negative or positive association with the dependent variable). Because of this explicit structure, the best explanation for the model is the model itself.

Another family of interpretable models are those based on decision trees. Here, predictions are generated by traversing through a sequence of "decisions", typically based on feature thresholds (e.g., if feature $x_i > 0.5$, then predict high $\hat{y}$). This structure makes it straightforward to trace how any given prediction was made.

However, some variants of this, such as random forest (Breiman, 2001) or even gradient-boosted trees (T. Chen et al., 2016) sacrifice this transparency for improved accuracy. While each individual tree is interpretable, an ensemble of hundreds or even thousands of trees make it difficult to aggregate traces or decisions. In these cases, one typically leverages built-in aspects of the training to facilitate interpretability, such as:

- **Decrease in impurity:** A measure of the reduction in impurity (i.e., how cleanly did this decision split the data) for a feature, aggregated across all trees.

- **Gain:** A measure of the total improvement in the model's loss given a feature.

These are essentially summary statistics of the models learning process, offering interpretability at a high level.

Post-hoc explanations

However, for more complex learners—such as particularly large decision-tree ensembles or deep neural networks—the model might not be able to reliably serve as its own explanation because its internal structure could be too intricate. In such cases, post-hoc explanation methods can be employed to approximate how input features influence predictions.

One widely adopted technique is SHAP (Shapley Additive Explanations). Drawing on cooperative game theory, SHAP assigns each feature a "credit" for a given prediction, decomposing the model's output into a sum of locally accurate, feature-specific contributions (Scott M Lundberg et al., 2017; Scott M. Lundberg et al., 2020). Then, for a given sample $(x^{(j)})$, the predictions can be decomposed in the following relation:

$$f(x^{(j)}) = \phi_0 + \sum_{i=1}^{m} \phi_i^{(j)}, \tag{4.2}$$

where $\phi_0 = \mathbb{E}[f(x)]$ is the expected prediction (baseline), and $\phi_i^{(j)}$ is the contribution of feature i to the prediction for sample j. $\phi_i^{(j)}$ is computed as the Shapley value of feature i, i.e. its average marginal contribution across all possible spaces of features (see Scott M Lundberg et al., 2017). This ensures local accuracy, as the baseline and all feature contributions sum exactly to the model's prediction for that sample.

In biology, these tools let us peer inside complicated models to uncover potential mechanisms. For example, a classifier trained to predict cellular fitness from gene-expression profiles might highlight a small subset of genes whose SHAP values consistently drive fitness predictions upward. Those genes then become candidates for downstream validation, hypothesis generation, or even therapeutic targeting; transforming opaque predictions into actionable biological insight.

Feature Explanations

It is important to distinguish feature interpretability from mechanistic explanation. A feature may reliably predict an outcome without itself being part of the causal mechanism. Such predictive associations may reflect correlation, shared upstream drivers, or other indirect relationships rather than true mechanistic involvement. To generate truly mechanistic explanations, studies like the one produced by J. H. Yang et al. (2019) have combined mechanistic data (fluxomics) with linear models, yielding transparent explanations for antibiotic resistance that directly reflect biochemical reaction rates. This is an example of a complementary strategy for improving interpretability. Namely to focus on the features themselves. By engineering or selecting descriptors that have clear, semantically meaningful definitions such as reaction fluxes, or actions of regulatory binding, the explanation task is simplified. When each input feature corresponds directly to a recognizable biological concept, even a complex model's decisions become more understandable. In practice, this can be achieved by, for example:

- Domain-guided feature construction, where expert knowledge is used to group sets of measurements into more higher-level summaries (such as pathway enrichment scores) (Barbie et al., 2009; Golriz Khatami et al., 2021).
- Rule-based or logic-derived features, for example from inductive logic programming, that encode specific relationships between entities in human-readable form (Orhobor et al., 2020; Brunnsåker et al., 2024).
- Ontology embeddings, where each dimension of a learned representation aligns with a known category or term (J. Chen et al., 2025).

By coupling these interpretable features with any learner, be it a linear model or deep neural network, you shift much of the burden of explanation onto the

features themselves. The model then only needs to combine a set of already-meaningful inputs, greatly simplifying both global and local interpretability analyses.

In several works presented in this thesis (Papers 2, 4 and 5), we proceduralized this principle by generating interpretable features via inductive logic programming (ILP) or ontological embeddings and then applying XAI techniques to them. By grounding each feature in explicit domain knowledge, we produce explanation models that are both faithful to the original learner and expressed in human-readable terms. This not only enhances interpretability but, in the case of ILP (see section 4.2.1), can also yield directly testable hypotheses expressed in natural language, bridging the gap between predictive power and scientific understanding.

4.2 Learning from Community Knowledge

A reoccurring theme of this thesis is the usage of structural knowledge priors, typically based on accumulated community knowledge constructed with structured ontologies. These priors can be leveraged for biological discovery using algorithms designed for use with relational data representations.

4.2.1 Inductive Logic Programming

Inductive logic programming (ILP) is a subfield of artificial intelligence that aims to learn logic programs from examples. This is typically done by constructing hypotheses (h) to explain examples (E) with the aid of background knowledge (B). Formally, the goal is to infer h such that, together with B, it correctly accounts for the observed data (Muggleton et al., 1994; Muggleton, 1999; Muggleton et al., 2012).

In what follows, we use the symbols $\wedge$ (logical and), $\vee$ (logical or), $\neg$ (logical negation), $\models$ (logical entailment), $\not\models$ (non-entailment), and $\square$ (falsity or contradiction). With this notation, a correct hypothesis is usually expected to satisfy four conditions (Muggleton, 1999):

$$B \not\models E^+ \qquad \text{(necessity)} \tag{4.3}$$

$$B \wedge h \models E^+ \qquad \text{(sufficiency)} \tag{4.4}$$

$$B \wedge h \not\models \square \qquad \text{(weak consistency)} \tag{4.5}$$

$$B \wedge h \wedge E^- \not\models \square \qquad \text{(strong consistency)} \tag{4.6}$$

Here, E^+ denotes the set of positive examples and E^- the set of negative examples. Necessity ensures that the background knowledge alone does not already entail the positives (so that the hypothesis is not redundant). Sufficiency requires that the hypothesis, together with the background, entails all positive examples. Weak consistency requires that the combination of background and hypothesis is satisfiable, i.e. free of contradiction. Strong consistency further requires that this remains true even when the negative examples are added, ensuring that no negatives are entailed. In practice, particularly in noisy domains, sufficiency and strong consistency are sometimes relaxed (or disregarded completely) and replaced with statistical criteria used to rank hypotheses that approximately satisfy these conditions (King et al., 2001).

Among the four conditions described above, sufficiency is the most central in practice, as it directly concerns whether a hypothesis explains the positive examples. In this thesis, however, sufficiency is often applied in a relaxed form, and consistency is usually interpreted in its weaker version—requiring only that hypotheses remain free of contradiction while covering the positives. Necessity is generally regarded as an auxiliary safeguard against redundancy.

A logic program acquired through these types of methods usually takes the following form:

$$a \leftarrow b_1, ..., b_n \tag{4.7}$$

where a is an atom (the head of the rule), and each b_i is a literal (an atom or its negation) in the body. Atoms represent basic propositions, from which more complex logical statements can be constructed. In ILP, a hypothesis h is such a logic program (i.e. a set of rules).

To enable systematic hypothesis construction, ILP often relies on the concept of *inverse entailment*, derived from the deduction theorem applied to the condition of sufficiency (Muggleton, 1999).

$$\begin{aligned} & B \wedge h \models E^+ \\ \Leftrightarrow\ & B \models (h \rightarrow E^+) \\ \Leftrightarrow\ & B \models (\neg E^+ \rightarrow \neg h) \\ \Leftrightarrow\ & B \wedge \neg E^+ \models \neg h \end{aligned} \tag{4.8}$$

Inverse entailment provides a way to construct candidate hypotheses from examples: given positive (and sometimes negative) examples, it allows the generation of a most-specific clause that is entailed by the data, from which hypotheses can then be derived (Muggleton et al., 1994; Muggleton, 1999).

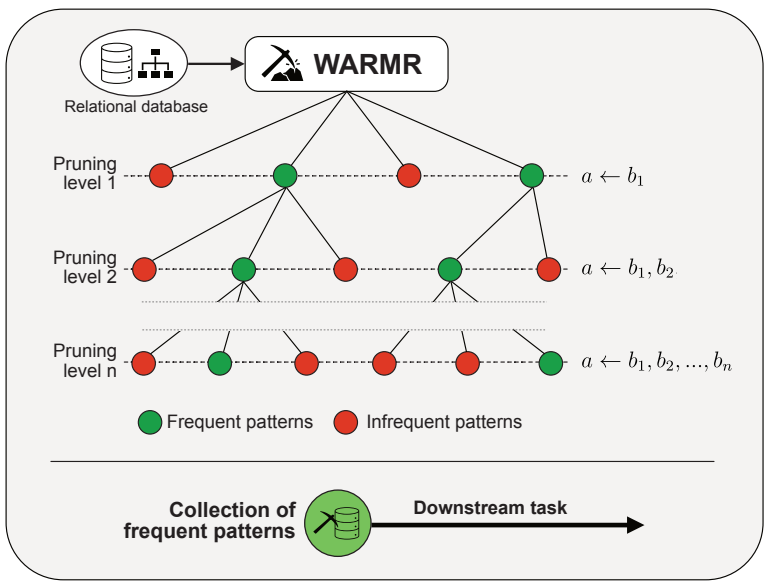


Figure 4.2: Overview of frequent pattern mining using WARMR. Given a relational (Datalog) database of the phenomena of interest defined by the head (a) and its associated relations (e.g., genes and their phenotypes), WARMR then performs a level-wise search over candidate queries defined by a user-specified language bias. At each step, existing patterns are refined by adding literals (b), and candidates that do not meet a pre-specified frequency (support) threshold are pruned. The logic program on the right side denotes the general structure of the program at the current level. The resulting frequent patterns can then be used as candidate hypotheses or as features for downstream tasks such as regression or classification.

Generating Candidate Logic Programs

In this thesis, candidate hypotheses are often generated using the WARMR algorithm, a level-wise ILP approach for discovering frequent Datalog patterns under a user-specified language bias (Dehaspe et al., 1999; King et al., 2001).

At a high level (also see Fig. 4.2), WARMR explores the space of possible queries allowed by the language bias in a breadth-first manner, retaining only those that occur frequently in the data. The language bias (restrictions on which predicates and argument types may appear in a clause) plays a central role, since it determines both the scope of the search and the interpretability of the resulting rules. In our setting, such biases are typically defined using biological relations (e.g. regulatory interactions, phenotypes), which ensures that the resulting candidate programs are meaningful within the domain.

WARMR provides an efficient way of generating structured hypotheses, which can then be evaluated using the ILP conditions introduced above. For a more detailed description of the algorithm, including data representation, search strategy and bias specification, please read the original descriptions as written by King et al., 2001 and Dehaspe et al., 1999.

An example logic program as defined in Prolog (using three of the biological concepts mentioned in Section 3.4.3 and used in Paper 2) could take the following form:

$$\begin{aligned} &\mathsf{Gene(A)} :- \\ &\quad \mathsf{Regulated\_By(A,\ B,\ Transcription\ factor)}, \\ &\quad \mathsf{Located\_In(B,\ Mitochondrion)}, \\ &\quad \mathsf{Enzyme\_Metabolite(A,\ Glutamine)} \end{aligned} \tag{4.9}$$

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.

An advantage of this method is the ability control the structure of the logic program according to a specified task. In Paper 5 we leverage this to shape the search in order to produce logic programs that adhere to typical experimental biological hypotheses, that could be easily testable with available laboratory infrastructure. An example of this can be seen below:

$$\begin{aligned} &\mathsf{Cells(A)} := \\ &\quad \mathsf{ExhibitsPhenotype(A,\ Increased\ Resistance,\ B,\ C)}, \\ &\quad \mathsf{CompoundName(B,\ Formic\ acid)}, \\ &\quad \mathsf{Condition(C,\ Treatment:\ 10mM\ Formic\ acid)}. \end{aligned} \tag{4.10}$$

This can be interpreted as: Cells (A) with increased resistance to formic acid (B) when exposed to a concentration of 10 mM (C).

Note that these patterns have no testable or actionable implications and instead only describe coverage of positive examples. Whilst the programs themselves can be identified and evaluated, they do not yet have a biological quantity associated to them outside of the examples they cover.

4.2.2 Finding Useful Patterns in Data

Whilst one can mine many patterns from structured databases, it could be difficult to know whether they are biologically relevant (or rather, relevant for the scientific question). For several works that has been done as part of this thesis, we assign importance to extracted patterns using XAI-techniques in order to assess their relevance.

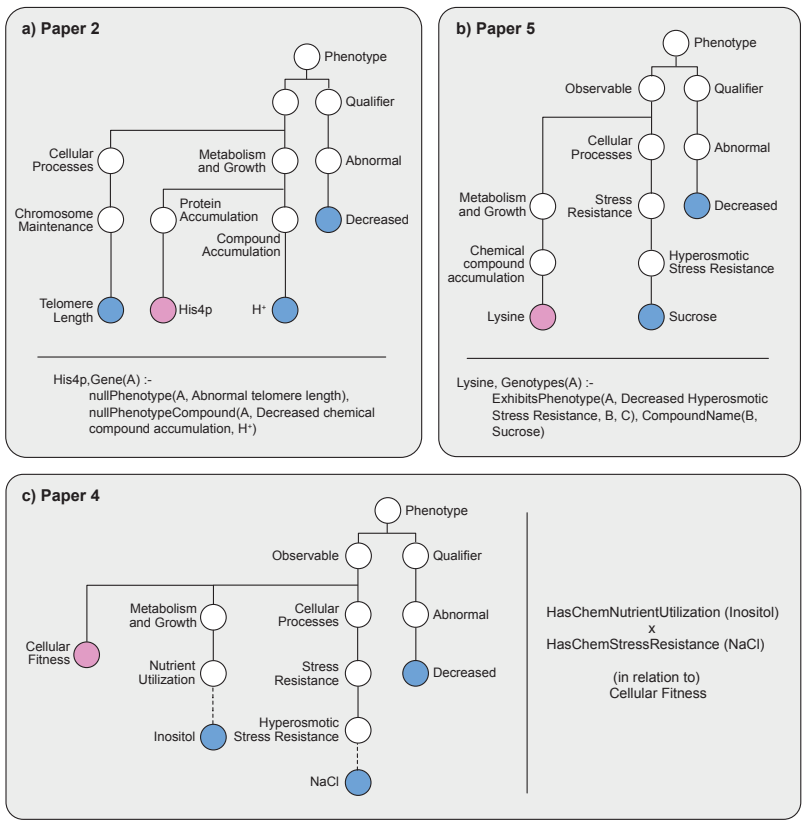


Figure 4.3: Extracted patterns used to generate hypotheses. This thesis has explored different uses of patterns from existing ontologies using various types of representation. **a.** Pattern used in Paper 2 to infer clues about general phenotypical associations with His4p abundance (examples with both abmormal telomere lengths and H^+ accumulations tend to be associated with changed His4p levels). **b.** Pattern and association used as the basis for an hypothesis in Paper 5. Involves the association between the amino acid lysine and overall tolerance to hyper-osmotic stress. **c.** Extracted structure used to predict an association between NaCl tolerance and Inositol utilisation (in relation to cellular fitness) in Paper 4. Dashed lines indicate a simplification of the underlying ontology for brevity. Blue indicates the concepts used to form the pattern itself, whilst pink signifies the biological readout used to define the relevance. The patterns and associations presented for Paper 4 and 5 were experimentally validated.

In Paper 4 we utilise an input-gradient method to extract useful patterns in the context of cellular fitness (Shrikumar et al., 2017; Costanzo et al., 2016). This allowed us to extract synergistic phenotypes, and generate a testable combination of traits—in this case the association between inositol utilization and NaCl tolerance—ass illustrated in Fig. 4.3.

For Papers 2 and 5, we use a similar approach, but in a propositional setting. The patterns themselves are propositionalised (i.e., by instantiating each predicate over the available constants to produce ground atoms, and then encoding those atoms as Boolean features) enabling the use of more efficient (and highly interpretable) attribute-value learners, such as linear regressors or decision trees (as explained in Section 4.1.2) (Kramer et al., 2001). In the context of Paper 2, biologically relevant patterns were found by predicting for protein abundances and connecting the qualitative patterns (logic programs) to this quantified biological entity through model explainability techniques (Scott M. Lundberg et al., 2020). For the latter paper, we used regularized linear models to force a testable implication unto the logic programs, allowing us to get clauses like the following (this was experimentally validated in the paper, as seen in Fig. 4.4):

$$
\begin{aligned}
+\,&\mathsf{Aminoadipate},\ \mathsf{Cells(A)} :=\\
&\quad \mathsf{ExhibitsPhenotype(A,\ Increased\ Resistance,\ B,\ C)},\\
&\quad \mathsf{CompoundName(B,\ Formic\ acid)},\\
&\quad \mathsf{Condition(C,\ Treatment:\ 10mM\ Formic\ acid)}.
\end{aligned}
\tag{4.11}
$$

Additional examples of underlying patterns used in this thesis can be seen in Fig. 4.3, highlighting several patterns used for experimental evaluation.

An additional advantage of this family of methods is that they can easily be used on top of existing ontologies, producing patterns that directly adhere to the underlying semantics.

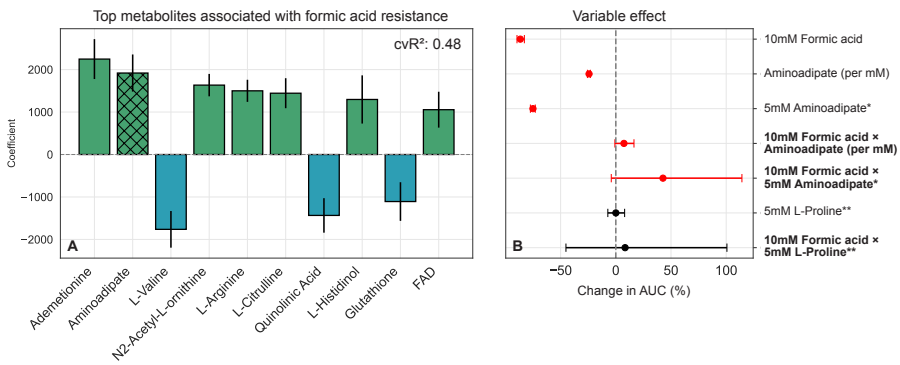


Figure 4.4: Example of data underlying the generation and evaluation of a pattern. In Paper 5, endpoint metabolic profiles and time-series growth data was used to infer compounds associated with formic acid resistance (panel **A**). A logic program was then procedurally generated (see Equation 4.11) and automatically experimentally evaluated, as seen in panel **B**. Significant interactions in red.

4.3 Large Language Models in Science

Large language models (LLMs) are deep neural networks typically built on transformer architectures (Vaswani et al., 2017). These are pre-trained on vast amounts of text to learn statistical patterns in language. Through a process of self-supervised learning, they internalize grammar, facts, and potentially even reasoning heuristics without explicit labels. These models can perform a wide variety of downstream tasks: drafting and summarizing text, translating between formats (e.g., natural language to code or experimental protocols), extracting structured data from unstructured sources, and (in the case of this thesis) generating hypotheses or experimental designs in response to a few examples or natural-language instructions.

Because they learn from such diverse data, LLMs exhibit remarkable adaptability. In scientific workflows, this flexibility has been harnessed for tasks including:

- Multi-agent AI co-scientist frameworks, where LLMs are organized into pipelines that iteratively propose and refine hypotheses. This approach has been used for drug repurposing for acute myeloid leukemia, identifying epigenetic targets for liver fibrosis and finding novel therapeutic targets for macular degeneration (Gottweis et al., 2025; Ghareeb et al., 2025).

- Drug-synergy hypothesis generation, using prompt-based exploration of chemical spaces to suggest combinations of FDA-approved drugs to treat cancer. Several pairs of drugs were tested on MCF7 breast-cancer cells, finding several with synergy exceeding standard controls. (Abdel-Rehim et al., 2025).

- Augmented autonomous chemistry, integrating LLMs with web searches, code execution, and robotic interfaces to design, execute, and optimize chemical reactions. The system autonomously improved yields in palladium-catalysed cross-couplings and successfully completed several other diverse synthetic tasks (Boiko et al., 2023).

- Retrieval-augmented knowledge grounding, combining LLMs with ontology-backed vector stores to extract structured hypotheses from literature and map entities to standard ontologies. This was applied to several publications in yeast biology, showing promise in automated hypothesis summarization and entity grounding (Reder et al., 2025).

Despite their power, LLMs can produce hallucinations (plausible-sounding statements unsupported by data) and may vary in consistency across outputs. While not always the case, generated hypotheses often lack a degree of validation that is standard in most experimental sciences, and model outputs may not conform to community data standards unless carefully guided.

In this thesis (Paper 5), we integrate LLMs into a logic-driven discovery framework where we leverage them to translate interpretable—ILP-derived—logic programs into detailed experimental plans that can then be executed robotically. This hybrid approach combines the flexibility of LLMs with

the precision and reproducibility afforded by formal logic (described in the previous sections) and standardized data practices. We show that this has many advantages over previously used methods, such as the ones seen in Gottweis et al., 2025.

Chapter 5

Summary of Included Papers

In this chapter, the five papers included in this thesis are summarized.

Paper 1 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—the diauxic shift. It highlights several metabolic pathways involved in the shift itself, whilst evaluating the use of mass spectrometry-based metabolomics for model validation and phenotyping for regulatory deletants.

Paper 2 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, enabling high-throughput hypothesis generation.

Paper 3 involves the creation of software pipeline for automated high-throughput metabolic profiling, enabling downstream scientific automation. It involves evaluation on several chemical standards and intracellular yeast matrices.

Paper 4 investigates and validates the use of ontology-based box embeddings and knowledge graphs to predict and interpret cellular fitness. It also investigates the frameworks ability to generate testable hypotheses. Experimental evaluation of hypotheses were performed, extracting insights on hyper-osmotic stress.

Paper 5 is about the design and experimental validation of an automated framework for biological discovery using logic programming and concepts from agentic AI. It combines ideas from **Paper 1-4**, involving automated hypothesis generation, laboratory automation, mass spectrometry-based metabolomics and automated hypothesis testing. It extracted insights from several metabolic interactions regarding amino acids, generating growth-data, metabolic profiles and detailed metadata in the process.

5.1 Paper 1: High-throughput metabolomics for the design and validation of a diauxic shift model

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 (Geistlinger et al., 2013). 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, where it instead typically ferments glucose into lactate (Liberti et al., 2016). Despite extensive research, the regulation of the diauxic shift remains poorly understood (Coutant et al., 2019).

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 the diauxic shift.

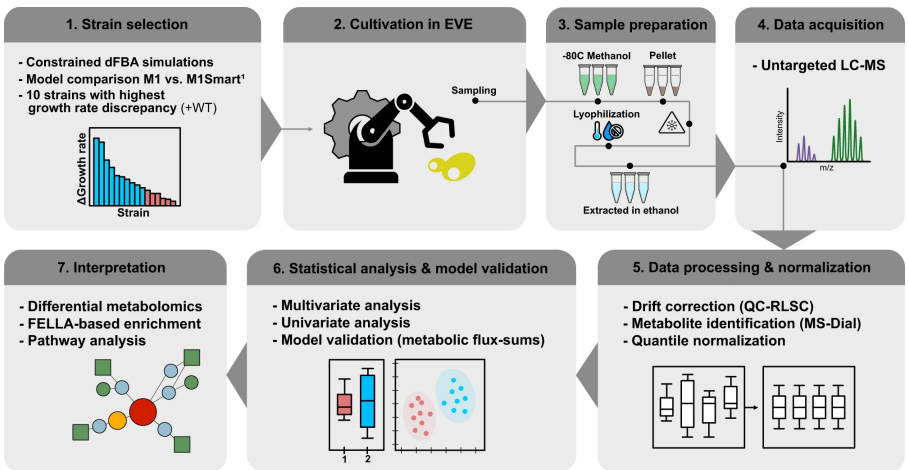


Figure 5.1: Workflow demonstrated in Paper 1. dFBA (dynamic Flux Balance Analysis) simulation suggests deletant strains which are subsequently cultivated in an automated laboratory cell and then and analysed using mass spectrometry-based metabolomics and various bioinformatics tools. [1]Simulations using models proposed in Coutant et al. (2019)

Genes of interest were selected based on the simulated impact of specific types of gene deletions on metabolism. This selection was performed using a combined signalling and regulatory network (explained in Sections 3.4.1 and 3.4.2), along with flux balance analysis, within a framework established by previous iterations of the robot scientist concept, developed by Coutant et al. (2019). The selection criteria were based on differences in growth phenotype given the structural changes caused by the semi-autonomous model improvements performed in the original work. These were then investigated through the phenotyping

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 Fig. 5.1.

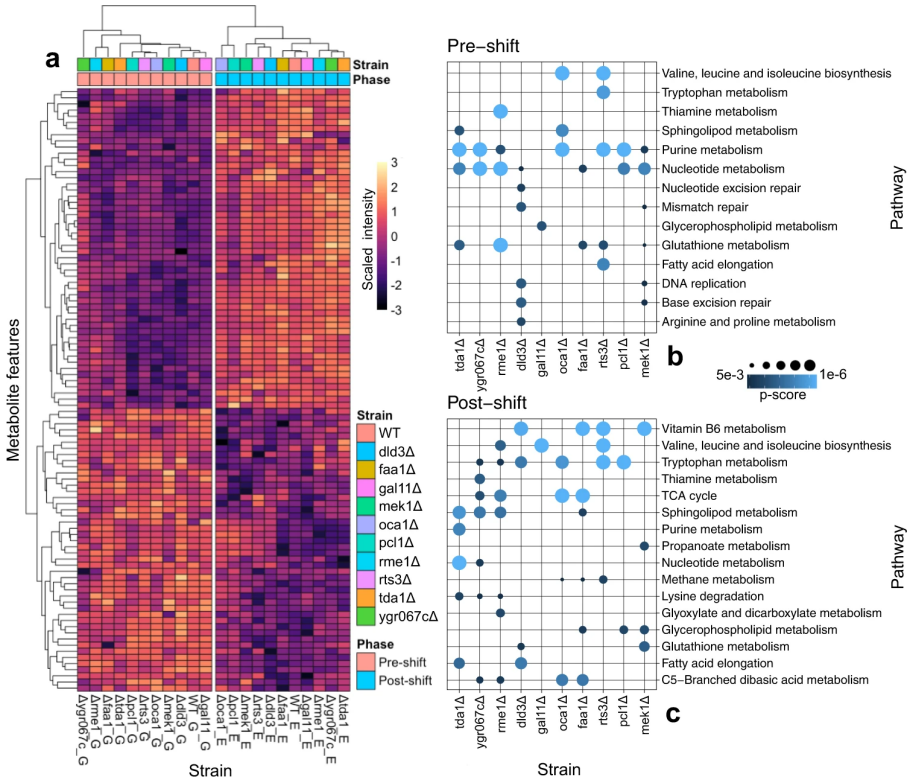


Figure 5.2: Overview of the effects of gene deletion on metabolic profiles. **a.** Metabolic profiles (levels of observable metabolites) for the deletant mutants pre and post diauxic shift. **b**. Pathway enrichment (with the KEGG-derived yeast metabolic network as the background) for the deletant strains 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).

We demonstrate the suggested workflow by successfully characterizing several genes involved in the diauxic shift, as seen in Figure 5.2. Three of these are of either or contested 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*) (Bjurström et al., 2025). Additionally, we further phenotype 5 other genes (*RME1*, *OCA1*, *PCL1*, *GAL11* and *MEK1*).

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, such as glycerophospholipid metabolism and the intersection between arginine,

proline and glutathione metabolism (see Fig. 5.3c). Additionally, we find that—unsurprisingly—the diauxic shift itself involves a major metabolic transformation, clearly visible through metabolomics data, as seen in Fig. 5.3a and 5.3b.

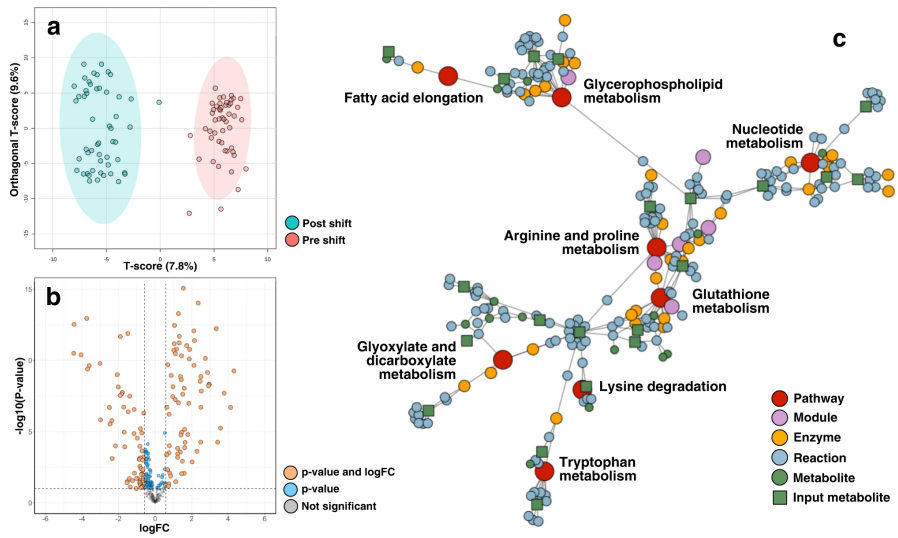


Figure 5.3: a. Diauxic shift phase classification and 95% confidence intervals using orthogonal partial least squares discriminatory analysis (oPLS-DA) with identified peaks as features. **b.** Volcano plot showing differentially expressed metabolites across the shift. **c.** Diffusion based topological enrichment with significantly enriched pathways in red.

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. We concluded that whilst it could be a useful tool, the steady state assumptions going into the simulation framework (see Section 3.4) are in conflict with the nature of measured metabolite accumulations. As such, it should likely be seen as a more holistic assessment, rather than be used for specific reactions.

We conclude that untargeted intracellular metabolomics is well suited to generating data and hypotheses about gene function due to its high information content. Moreover, it is ideally positioned to support automated approaches due to the relative simplicity of processing and the potential to make massively high-throughput.

5.2 Paper 2: Interpreting protein abundance in *Saccharomyces cerevisiae* through relational learning

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 a challenge (Wood et al., 2019).

In this study, we investigate proteomic and metabolic profiles derived from a collection of *S. cerevisiae* deletants, utilizing structured priors, relational learning, and supervised machine learning (both described in Chapter 4).

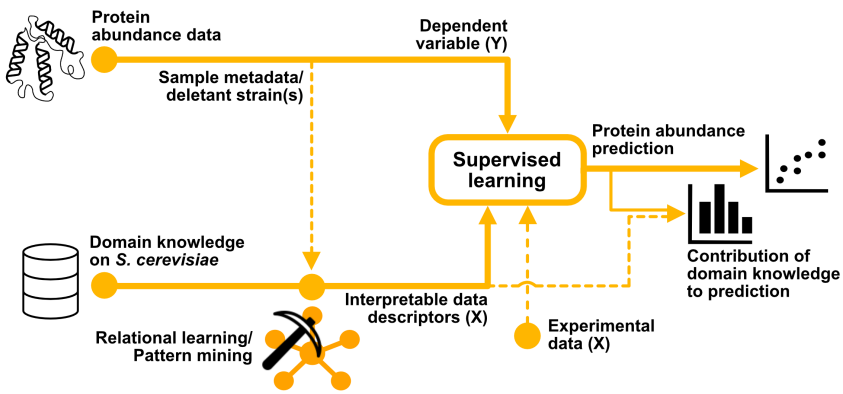


Figure 5.4: Overview of methodology applied in Paper 2. Metadata (genotype) from data sets on proteomic abundances is used to identify frequent patterns in a relational database. The frequent patterns are propositionalized and used to predict protein levels in an explainable manner.

S. 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 (Engel et al., 2025; Cherry et al., 2012; H. Lu et al., 2019). This work subsequently makes use of this prior to learn predictive relationships between proteomic profiles (generated by Messner et al. (2023)) and the functional characterization of the yeast genome. This is done by translating this knowledge into a Datalog database, and using frequent pattern mining (applied through inductive logic programming) to generate logic programs—representing biologically relevant regulatory rules. These were then evaluated using supervised learning and feature analysis. See Fig. 5.4 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, protein structure, 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 4.2.1. These hypotheses/relational features are then assessed by evaluating their predictive power (in terms of proteomic abundances) as seen in Fig. 5.5.

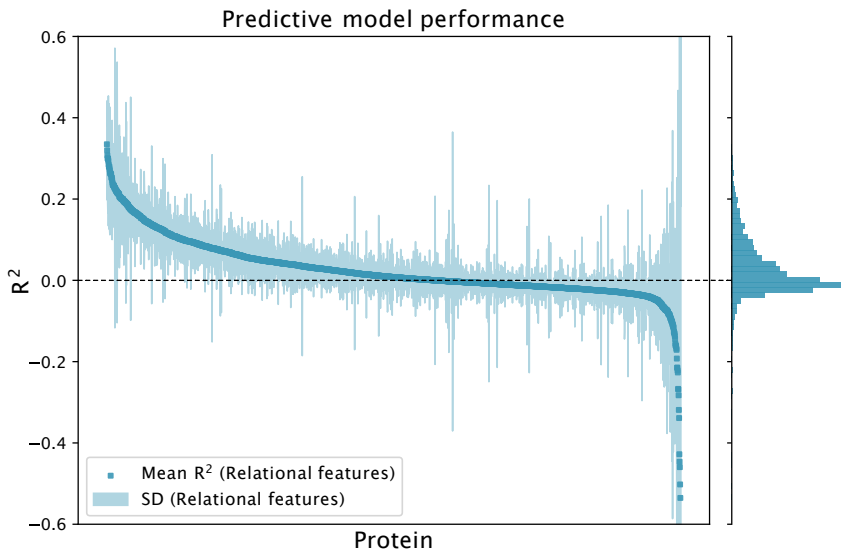


Figure 5.5: Predictability of protein abundance given relational features. R^2 denotes the coefficient of determination (proportion of variance explained).

By assessing predictive logic programs across all of the 2292 proteins present in the study by Messner et al. (2023), we could evaluate which biological concepts contributed most to protein abundance in general (according to our framework) as seen in Fig. 5.6. For example, highly impacting concepts such as abnormal growth rates severely affect protein abundances in many cases, typically connected to abnormalities in abundances of amino acids—molecules crucial for a functional metabolism and metabolic signalling as well as being fundamental building blocks of proteins.

We also learnt several predictive relationships between specific 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 His4p and Ilv2p (see Fig. 5.7); the methodology successfully validated existing literature, but also inferred their roles as regulatory elements for neighbouring processes.

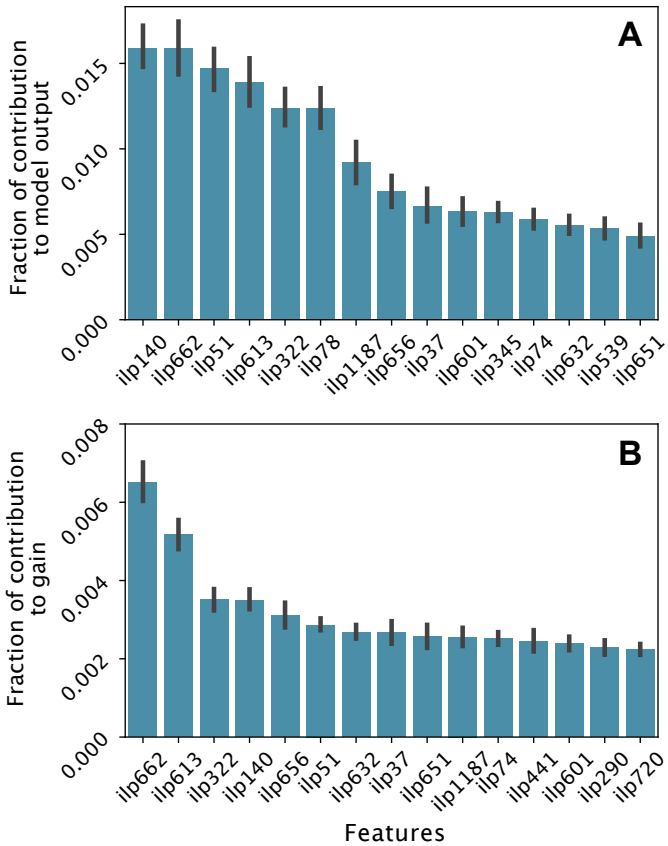


Figure 5.6: A. Normalized average SHAP-values of relational features across all available protein models with a positive coefficient of determination. **B.** Normalized gain of relational features across all available protein models with a positive coefficient of determination. Both of these concepts of explainability (SHAP and gain) are explained in more detail in Chapter 4

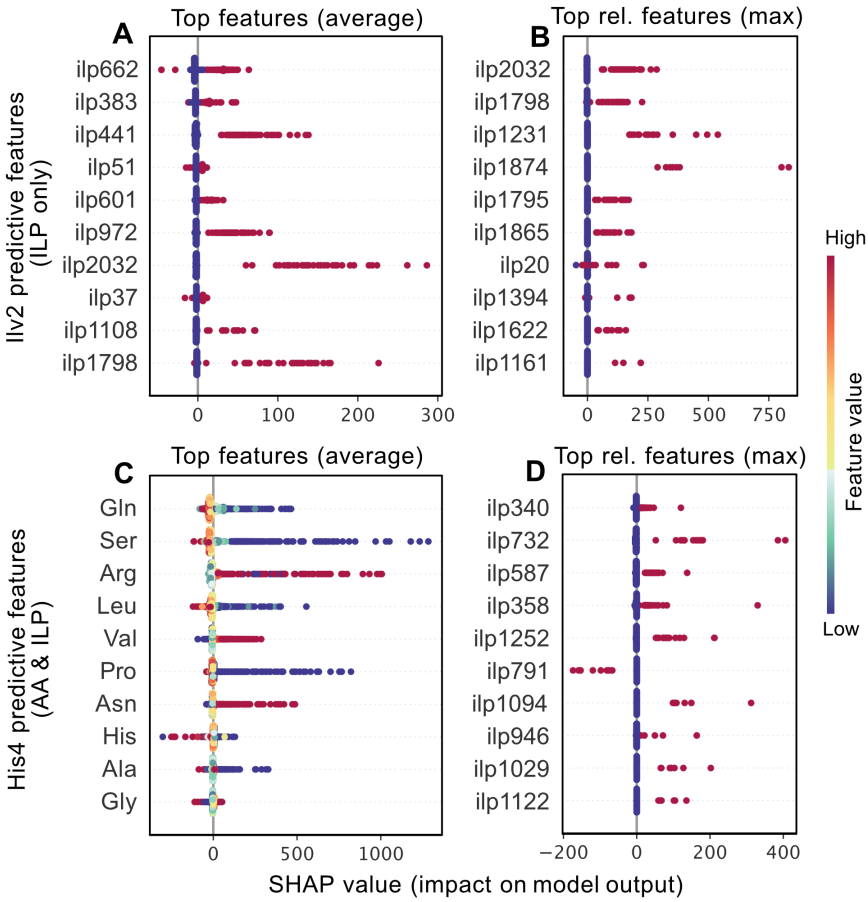


Figure 5.7: A. Top features for the prediction of Ilv2p, given only relational features (the frequent patterns). Sorted by average contribution in descending order. **B.** Top relational features for Ilv2p abundance, according to maximum change in model output (i.e facts that severely changed the outcome of the prediction for a subset of proteins), given only relational features. **C.** Top features for the prediction of His4p abundance, 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).

5.3 Paper 3: AutonoMS: Automated Ion Mobility Metabolomic Fingerprinting

Modern life science laboratories are transitioning from manual work to high-throughput, data-centric discovery platforms (Musslick et al., 2025; Lobentanzer et al., 2025; Coutant et al., 2019). This transformation is driven by the convergence of robotic automation and integrated software systems. It is becoming increasingly common for automated laboratory instruments to handle labour-intensive tasks like sample preparation and assay execution, increasing both data quantity and quality (Holland et al., 2020; Bai et al., 2022).

Mass spectrometry (MS), especially when integrated with ion mobility (IM-MS, further explained in Chapter 3), is a widely used analytical tool in life sciences due to its sensitivity and resolution. IM-MS adds an additional separation dimension based on molecular structure, enhancing its utility for complex sample analysis. Despite advancements in automation and analytical techniques, many mass spectrometry workflows remain complex, time-intensive, and require significant manual intervention. In areas like metabolomics, analytical instrumentation are often late adopters of automation, limiting throughput. The complexity introduced by high-throughput systems like IM-MS further exacerbates this challenge.

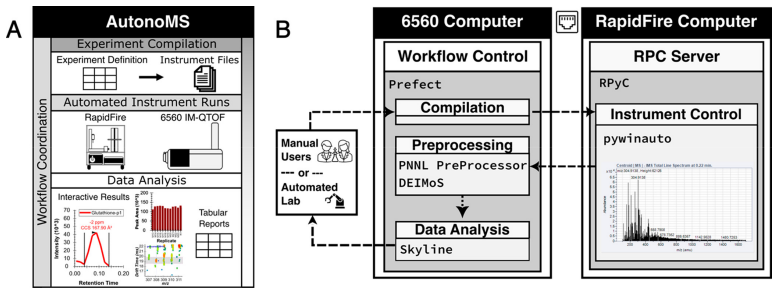


Figure 5.8: AutonoMS enables walkaway automation of ion mobility mass spectrometry data collection and analysis. **A.** AutonoMS integrates software control layers with an Agilent RapidFire-6560 ion mobility mass spectrometry system to provide automated data acquisition, raw data handling, data processing, and metabolomic end-to-end analysis. **B.** The AutonoMS software stack is shared between the 6560 and RapidFire control computers. Human users or an upstream software agent may trigger AutonoMS runs using a pre-specified experiment definition file.

In this work, we introduce *AutonoMS*. It is an automated platform for mass spectrometry experimentation. It automates sample injection, data acquisition, and metabolomics analysis using open-source libraries and integrates with an Agilent RapidFire and 6560 DTIMS-QTOF systems (see Fig. 5.8). AutonoMS allows experiments to be planned and executed programatically, supporting upstream automation agents.

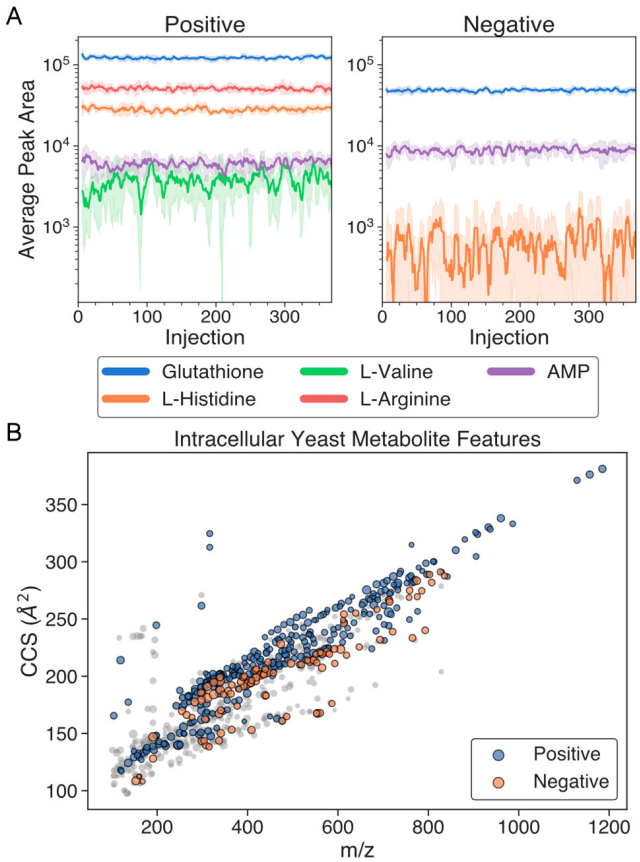


Figure 5.9: Automated analysis of the extracted intracellular yeast samples with AutonoMS. **A.** Detected peak areas in extracted yeast samples across injections of the 5 ions used in the chemical standards analysis. Peak areas shown as the moving average (solid lines) together with the standard deviation (shaded areas). **B.** Untargeted metabolite features found across all extracted yeast samples across positive (blue) and negative (orange) ionization modes. Single ionization state ion features are shown in gray, and marker size is scaled according to abundance.

We demonstrate autonomous operation on biologically relevant chemical standards, and extract untargeted metabolomics data from complex biological samples (*S. cerevisiae*) as seen in Fig. 5.9. The platform addresses the need for greater automation at the analytical end of omics workflows, supporting large-scale screening and discovery efforts. Its ability to bridge experiment execution with informatic analysis highlights several use-cases for downstream closed-loop automation.

5.4 Paper 4: Ontology-based box embeddings and knowledge graphs for predicting phenotypic traits in *Saccharomyces cerevisiae*

Despite decades of study, many aspects of yeast biology are poorly understood. Many genes are still unannotated, and complex interplay between genes and proteins can produce unexpected phenotypes (Wood et al., 2019; Costanzo et al., 2016; Kuzmin et al., 2018). Fully exploring even simple organisms experimentally is an extremely resource-heavy and time-intensive undertaking due to the sheer complexity of biological systems. Therefore, scalable, information-rich, methods for hypothesis generation are needed to accelerate biological discovery (Coutant et al., 2019).

Saccharomyces cerevisiae is a widely studied model eukaryotic organism due to its industrial relevance and biological similarity to higher eukaryotic organisms. Due to its history as a widely used research organism, there exists extensive structured resources like the Saccharomyces Genome Database (SGD), BioCyc, and ontologies such as Gene Ontology (GO), Ascomycete Phenotype Ontology (APO) which provide rich biological knowledge about yeast (Cherry et al., 2012; Engel et al., 2025; Karp et al., 2019; Ashburner et al., 2000).

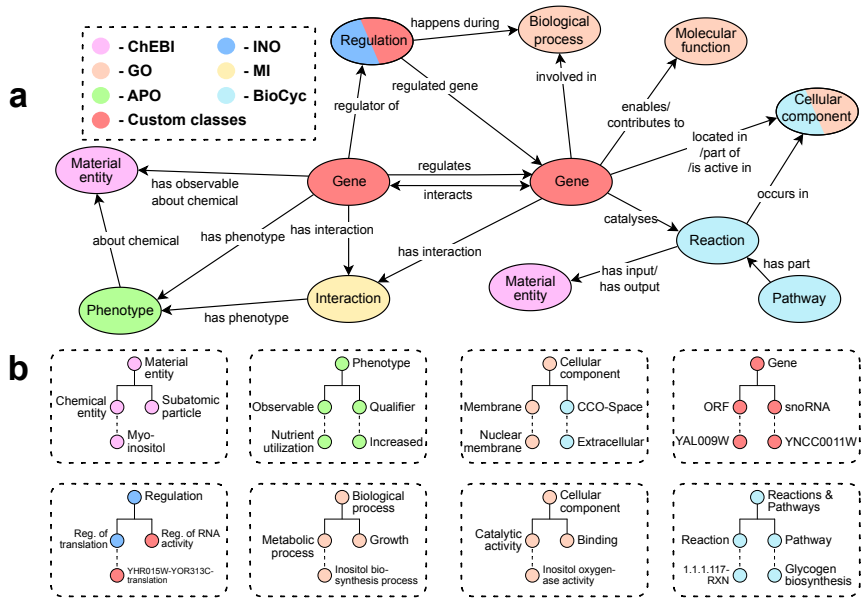


Figure 5.10: a. An overview of the data representation used in this study, and how it is connected in the knowledge graph. The colour of the nodes specifies where the classes (nodes) themselves are defined. **b.** Examples from the defined hierarchies.

Knowledge graphs (KGs) offer a structured way to integrate heterogeneous biological data. These KGs can be embedded into vector spaces to enable

computational tasks like link prediction. Techniques such as TransE, as well as graph neural networks, can generate useful representations of entities and relationships. These representations can facilitate downstream tasks such as phenotype prediction and hypothesis generation about biological function.

In this work, we present a method that uses graph neural networks (GNNs) to predict and interpret the effect of gene deletions in the yeast *S. cerevisiae*. It makes use of a knowledge graph (KG) and ontology-based box embeddings, utilising several widely used ontologies combined with bespoke integration terms (see Fig. 5.10 and 5.11). From the class hierarchies in the ontologies, box embeddings are learnt as low dimensional representations of the nodes in the graph, which are used together with GNNs to predict cell growth for double gene knockouts from the dataset generated by Costanzo et al. (2016). We show that high level qualitative information can be used to predict experimental data (such as cellular fitness).

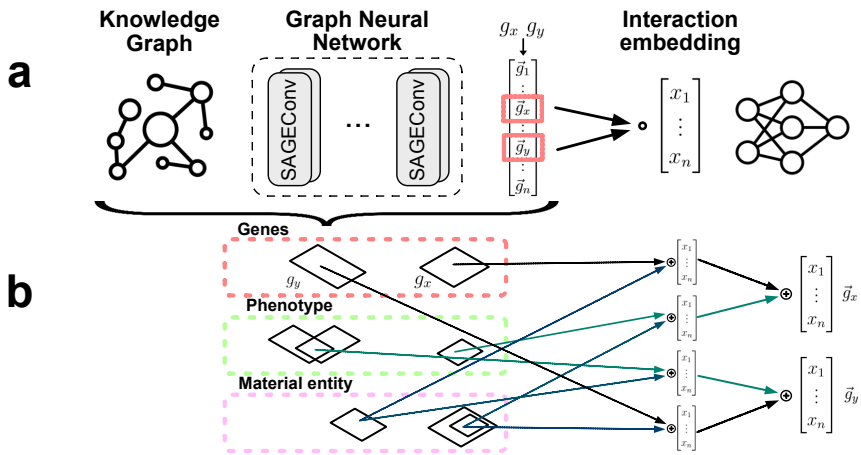
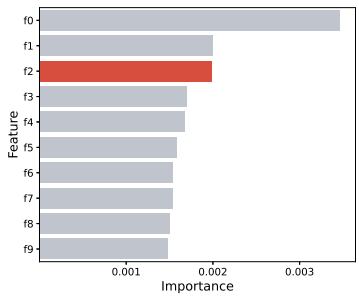


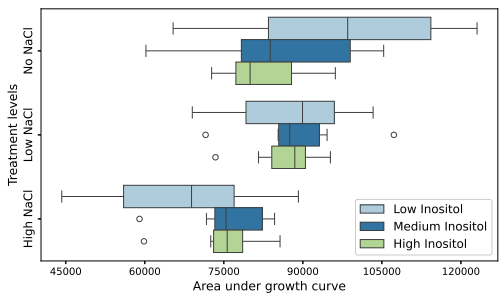
Figure 5.11: a. An overview of the framework for predicting the fitness when deleting pairs of genes. **b.** Representation of different domains and how they are aggregated in the GNN. Arrows from the boxes represent SAGE modules.

We also demonstrate that the model can generalise beyond the task it was trained for by evaluating its performance on triple knockouts (Kuzmin et al., 2018). Additionally, we apply model interpretability techniques (Kokhlikyan et al., 2020) to identify co-occurring edges important for fitness predictions. Highlighted results can be seen in Fig. 5.12(a).

We additionally use the outcomes of this to computationally generate an hypothesis about inositol utilisation and sodium chloride (table salt) tolerance (highlighted in Fig. 5.12(a)). This hypotheses was validated in an automated biological experiment revealing a dose-dependent rescuing effect, as seen in Fig 5.12(b). Potentially highlighting and further validating inositols association with cellular membrane stability.



(a) Feature importances derived from model explainability techniques.



(b) Boxplot of growth AUC (Area Under Curve) for experimental conditions tested in study.

Figure 5.12: Overview of the feature selection and experimental results. **a.** Highest ranked features, selecting based on phenotypes testable in our experimental setup through filtering of ontological terms. Highly ranked features in grey were not selected due to safety constraints (e.g., cancerogenic compounds). **b.** Box plot showing the distribution of AUC (a holistic measure of growth dynamics) for all of the tested experimental conditions. Results show a significant (dose dependent) rescuing effect of inositol during NaCl treatment.

5.5 Paper 5: Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology

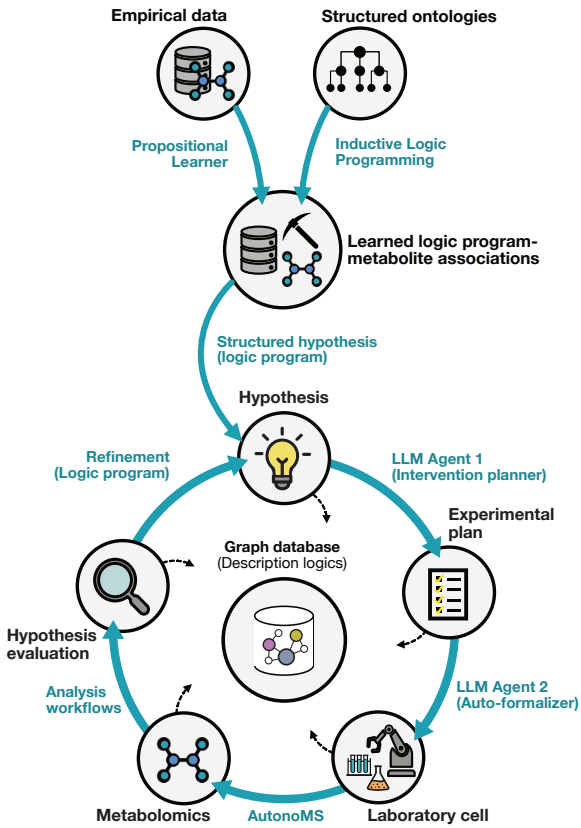


Figure 5.13: An automated framework for biological discovery, covering the full cycle from generating hypotheses, executing experiments and integrating results. Hypotheses are derived from structured yeast knowledge using ILP and the linked to metabolomics data. The hypotheses are then tested through experiments designed by a large language model (LLM). A robotic lab system automatically executes experiments, measuring growth over time and collecting metabolic profiles through mass spectrometry. All outcomes, including metadata, are stored in a graph database for analysis, transparency, and reuse.

Artificial intelligence (AI) combined with laboratory automation offers the possibility of transforming how science is done. Advances in AI, robotics, and high-throughput technologies now make it possible to imagine research systems that design, execute, and interpret experiments with minimal human intervention (King et al., 2009; Bai et al., 2022; Musslick et al., 2025). Biology,

in particular, poses challenges of scale and complexity that far exceed human analytical capacity. Even in simple model organisms such as *S. cerevisiae*, the number of interacting components and possible experimental conditions makes manual exploration infeasible (Coutant et al., 2019). To address this, we present a framework that integrates logic-based reasoning, large language models (LLMs), and laboratory automation for end-to-end scientific discovery (as illustrated in Fig. 5.13).

Our system combines the flexibility of LLMs with relational learning, grounded in community-adopted ontologies. Hypotheses (see Fig. 5.14) are generated from structured biological data, automatically formalized, and then prioritized. Experimental plans are designed with the aid of LLMs, which are then executed on robotic platforms. All data—including metadata and intermediate outputs—are stored in a graph database to ensure traceability, reproducibility, and reuse (King et al., 2011).

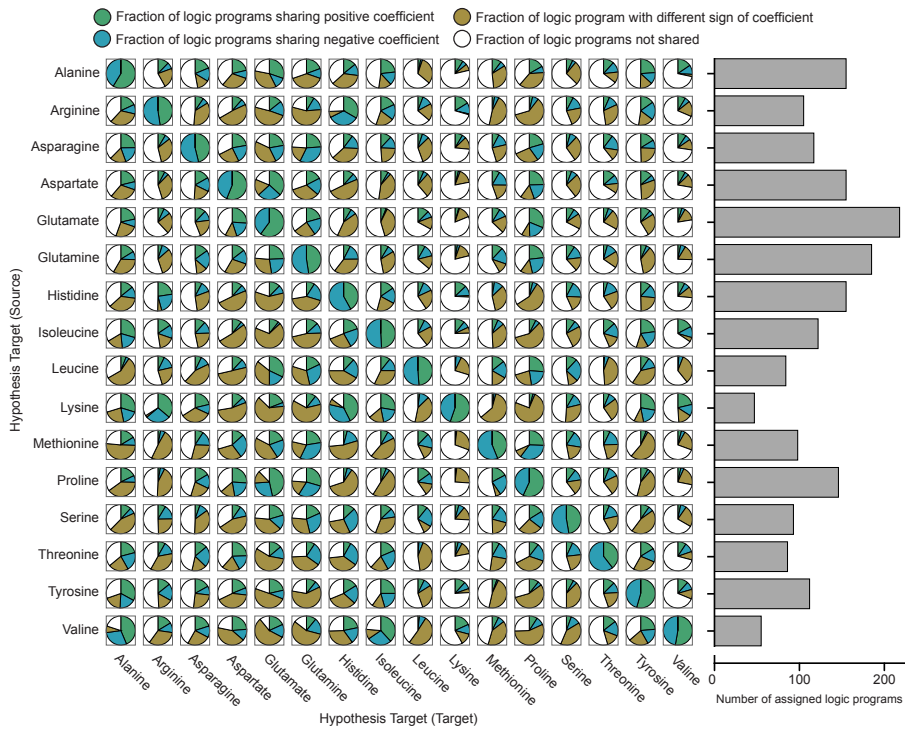


Figure 5.14: Asymmetric overlap between amino acid hypothesis spaces. Each pie chart shows, for a given amino acid (vertical axis), the fraction of its logic programs that are also linked to another amino acid (horizontal axis). Shared programs are colored by their regression outcome: blue = positive, green = negative, yellow = differing signs. The bar chart indicates the total number of logic programs assigned to each amino acid on the vertical axis.

We validated this framework in *S. cerevisiae*, focusing on interactions between amino acid supplementations and stress conditions. Automated experiments

revealed previously underexplored phenomena, such as glutamate-induced growth inhibition in spermine-treated cells and arginine-mediated enhancement of caffeine toxicity. Additionally, through iterative hypothesis refinement, we discover an association between aminoadipate and formic acid stress.

Even when predictions failed, the system produced valuable insights. For instance, highlighting when phenotypes were likely driven by downstream metabolites rather than the supplemented compounds themselves.

The framework's modular design makes it readily extensible to other hypothesis types and data modalities, such as transcriptomics or proteomics. While some minimal human input still remains, these steps are easily automated if need be.

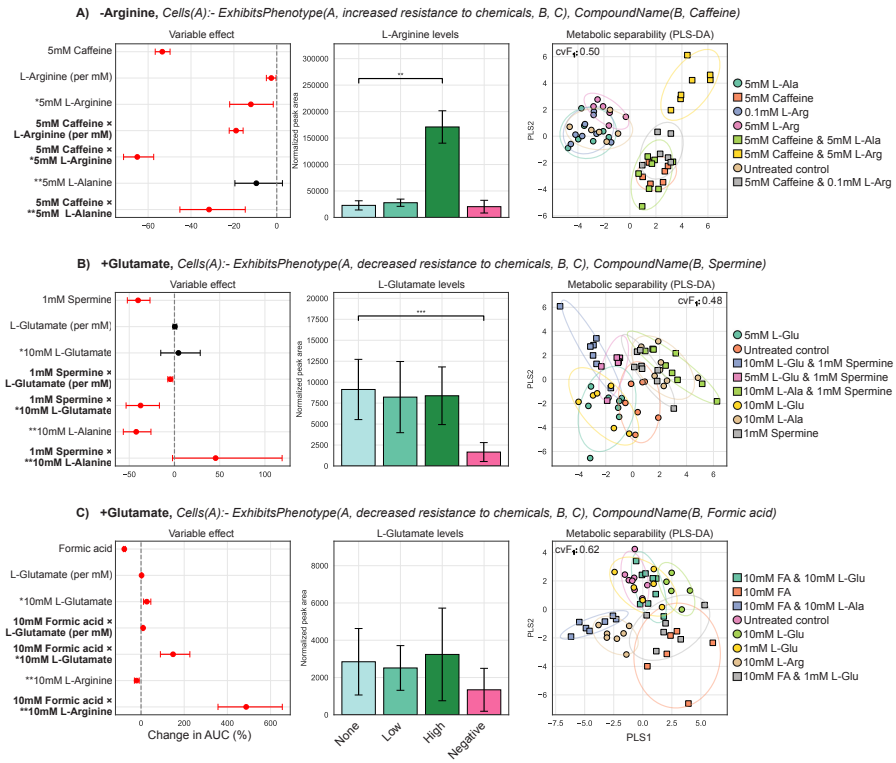


Figure 5.15: Results from three automated interaction experiments with different outcomes. The title shows the tested logic program. The first plot shows how the interventions affected growth, with red marking significant effects. The bar plot shows how much of the tested compounds built up inside cells. The final plot illustrates how the metabolic profiles of the groups differed, confirming that the interventions caused detectable changes in metabolism.

More broadly, this work demonstrates the feasibility of a logic-driven, agentic, multi-component framework for automated hypothesis generation and experimental validation. By combining symbolic reasoning, language-driven flexibility, and robotics with metabolomics as a scalable readout, we show how

AI scientists can become reliable partners in systems biology—accelerating discovery, ensuring reproducibility, and operating at scales beyond traditional human reach. Leveraging mass spectrometry-based metabolomics as a central data source provides an informative readout for scientific discovery, while remaining fully compatible with high-throughput robotic experimentation. This integration of reasoning, automation, and metabolomics shows how AI-driven discovery cycles can be more comprehensive than those relying on a single measurement type, and more precise than those built on loosely typed message passing. We believe this platform lays the groundwork for a more reliable, machine-driven discovery process in systems biology, extensible to other omics modalities and experimental domains.

Chapter 6

Concluding Remarks

This thesis has explored the automation of scientific discovery in yeast systems biology, focusing on the development and integration of computational and robotic systems for hypothesis generation, experimental design, data acquisition, and analysis. The collective works illustrates how automation and machine learning can be combined to accelerate functional genomics, reduce human bias, and improve reproducibility in biology.

A central theme of this work has been the automation of experimental workflows. From semi-automated experiment selection and metabolomics profiling (Paper 1) to fully integrated systems for hypothesis generation and robotic execution (Paper 5), the research demonstrates how robotics, computational algorithms and experimental design can replace or augment manual laboratory processes. This contributes not only to increased scientific throughput, but also to reduced variability and mitigation of reproducibility concerns.

Another major contribution lies in making better use of existing biological knowledge. Papers 2 and 4 demonstrate how large-scale experimental data and structured ontologies can be used to infer new functional relationships and generate interpretable predictions. Even when predictive accuracy is imperfect, these approaches provide insight by grounding predictions in prior knowledge and data-driven relationships.

A third explored aspect is the development of frameworks for interpretable and flexible hypothesis generation. By combining relational learning with structured representations and language models (Papers 4 and 5), the thesis showcases how machine learning systems can both formulate and reason about biological hypotheses in ways that are highly accessible to human researchers with a basic grasp of the domain.

Together, these works represent a step toward closed-loop scientific systems that span from data to discovery. They show how diverse tools—ranging from inductive logic programming and graph neural networks to large language models and laboratory automation—can be composed into modular workflows that assist in, or even drive, the scientific process end to end.

While this work highlights some progress, there are challenges that remain. Data heterogeneity, model generalizability, and system integration all

require further refinement. Nonetheless, the thesis lays a foundation for more autonomous and interpretable systems biology. While these results represent a step forward, they also highlight open challenges and opportunities for further development.

6.1 Limitations

While the individual studies in this thesis demonstrates technical feasibility and proof of concept, it also contains many methodological limitations, constraining broader applicability.

Work in this thesis has relied extensively on mass spectrometry-based readouts (especially metabolomics). Whilst metabolomics provide a rich and functionally relevant readout of cellular states, it is inherently difficult to connect it to biological mechanisms. It is a highly volatile measure, and many aspects of modelling relies on completely disregarding their accumulation (FBA, steady-state assumptions). Additionally, in mass spectrometry-based workflows, compound coverage is typically limited due to factors like ionization efficiency, instrument sensitivity or matching database quality (Alseekh et al., 2021). As a result, a large fraction of the metabolome can not be reliably covered, much less quantified. This incomplete picture risks leading to incorrect conclusions, for example, inferring that a pathway is inactive simply because its metabolites are not observed.

Another key limitation in this work is the heavy emphasis on single measurement modalities. While many of the approaches described here implicitly integrate heterogeneous data sources through logic formalisms or other qualitative frameworks, these do not replace the need for direct, quantitative measurements. For example, metabolomics can capture functional endpoints of cellular processes, but without parallel quantitative readouts from other layers (e.g. transcriptomics, proteomics or phosphoproteomics), important regulatory or signalling events may be entirely missed. This absence of multi-layer quantitative data can obscure causal relationships, weaken mechanistic interpretation, and limit the ability to model dynamic, multi-scale processes.

ILP and other logic-based approaches, while highly interpretable, are bounded by the expressivity of the underlying logic and the completeness of the background knowledge (Muggleton et al., 2012). They typically work in discrete, symbolic spaces, which can oversimplify the inherently quantitative and dynamic nature of cellular biology. Search spaces must also be heavily curated and pruned to remain computationally tractable, which risks excluding relevant but complex patterns.

The use of LLMs for hypothesis generation and planning (Paper 5) introduces reliability concerns. While these models can synthesize coherent, domain-relevant text, they are not inherently grounded in factual accuracy or domain constraints. They may hallucinate entities, misinterpret context, or propose experiments that are syntactically correct but scientifically infeasible. Whilst considerations have been taken to reduce this unreliability (such as reliance on logical structures), it still necessitates careful validation (and often

human oversight) before acting on their output.

Finally, many of the methods described here depend on the rich annotation and densely mapped networks available for *S. cerevisiae* (Cherry et al., 2012; Engel et al., 2025). In less-characterized organisms—where genome-scale models, interaction networks, and ontologies are much more incomplete—the same approaches may yield weaker results, with greater uncertainty in both hypothesis generation and interpretation. Real-world validation can also become substantially more challenging for non-model organisms, as cultivation conditions, genetic tools, and assay protocols are often less standardized or more technically demanding. While far from insurmountable, these combined factors pose significant barriers to extending the pipelines described here to other species, such as higher eukaryotes.

6.2 Future directions

While this thesis demonstrates how various components of scientific discovery can be automated and integrated, much remains to be done to realize the full potential of autonomous systems biology.

The automation of data acquisition, as explored in Papers 1 and 3, provides a strong foundation for high-throughput experimentation. However, these systems remain limited in scale, scope, and adaptability. Expanding beyond metabolomics to incorporate other data modalities, such as transcriptomics, proteomics, or fluxomics, will be necessary to support more comprehensive models of cellular behaviour.

Papers 2 and 4 demonstrate the utility of structured reasoning and interpretable modelling for drawing insights from existing biological data. Yet, challenges remain in bridging the gap between qualitative biological knowledge and quantitative experimental outputs. Future research should focus on tighter integration of ontologies, formalized metadata, and context-aware reasoning frameworks. Improving the robustness of predictions, especially in the presence of noisy or incomplete data, will also be critical. Such work would enable more nuanced and informed inferences, moving beyond simple pattern recognition toward deeper biological understanding.

Paper 5 presents a combination of many of these previous ideas, producing a system that generates hypotheses, plans experiments, and executes them in the lab. However, the autonomy of these types of systems is limited. Feedback from experimental outcomes is still underutilised, as the system does not explicitly update its own models. Achieving truly closed-loop discovery will require the ability to learn from failure to a much larger extent. The systems would also need to reason more under uncertainty, and adapt to new evidence, all while maintaining interpretability and transparency. This remains one of the most important and difficult challenges in scientific automation.

Across all these domains, several limitations need to be addressed. The heterogeneity of biological data, the lack of standardized metadata, and the lack of robustness of current reasoning systems limit the generalizability and reliability of automated approaches. Moreover, while automation can increase

scale, it also introduces risks. This is particularly true if decisions made by more "opaque" models are not subject to human oversight, such as in the case of parts of the workflow in Paper 5.

Additionally, future systems must be designed not only to operate autonomously but to collaborate effectively with human researchers. This could be integrated in many different ways, from iterative learning to human-in-the-loop type solutions.

In conclusion, this thesis outlines one possible path toward machine-assisted scientific discovery. It provides some tools and frameworks, but also exposes many of the unresolved questions that future research must tackle. These include questions of scale, reasoning, collaboration, trust, and fairness. Addressing them will be essential not just for automating existing processes, but for fundamentally transforming how scientific discovery is conducted.

Bibliography

Abdel-Rehim, Abbi, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova and Ross King (June 2025). "Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment". In: *Journal of The Royal Society Interface* 22.227, p. 20240674. DOI: 10.1098/rsif.2024.0674.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter (Dec. 2007). *Molecular Biology of the Cell*. 5th ed. New York: W.W. Norton & Company. ISBN: 978-0-203-83344-5. DOI: 10.1201/9780203833445.

Alseekh, Saleh, Asaph Aharoni, Yariv Brotman, Kévin Contrepois, John D'Auria, Jan Ewald, Jennifer C. Ewald, Paul D. Fraser, Patrick Giavalisco, Robert D. Hall, Matthias Heinemann, Hannes Link, Jie Luo, Steffen Neumann, Jens Nielsen, Leonardo Perez de Souza, Kazuki Saito, Uwe Sauer, Frank C. Schroeder, Stefan Schuster et al. (July 2021). "Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices". In: *Nature Methods* 18.7, pp. 747–756. DOI: 10.1038/s41592-021-01197-1.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin and Gavin Sherlock (May 2000). "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29. DOI: 10.1038/75556.

Bai, Jiaru, Liwei Cao, Sebastian Mosbach, Jethro Akroyd, Alexei A. Lapkin and Markus Kraft (Feb. 2022). "From Platform to Knowledge Graph: Evolution of Laboratory Automation". In: *JACS Au* 2.2, pp. 292–309. DOI: 10.1021/jacsau.1c00438.

Baker, Monya (May 2016). "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604, pp. 452–454. DOI: 10.1038/533452a.

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen et al. (Apr. 2016). "The

Ontology for Biomedical Investigations". In: *PLOS ONE* 11.4, e0154556. DOI: 10.1371/journal.pone.0154556.

Barbie, David A., Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta et al. (Nov. 2009). "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1". In: *Nature* 462.7269, pp. 108–112. DOI: 10.1038/nature08460.

Bjurström, Erik Y., Praphapan Lasin, Daniel Brunnsåker, Ievgeniia A. Tiukova and Ross D. King (2025). "An Investigation of TDA1 Deficiency in Saccharomyces cerevisiae During Diauxic Growth". In: *Yeast* 42.5-7, pp. 142–156. DOI: 10.1002/yea.4004.

Boiko, Daniil A., Robert MacKnight, Ben Kline and Gabe Gomes (Dec. 2023). "Autonomous chemical research with large language models". In: *Nature* 624.7992, pp. 570–578. DOI: 10.1038/s41586-023-06792-0.

Botstein, David and Gerald R. Fink (Nov. 2011). "Yeast: An Experimental Organism for 21st Century Biology". In: *Genetics* 189.3, pp. 695–704. DOI: 10.1534/genetics.111.130765.

Breiman, Leo (Oct. 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Brunnsåker, Daniel, Filip Kronström, Ievgeniia A Tiukova and Ross D King (Feb. 2024). "Interpreting protein abundance in Saccharomyces cerevisiae through relational learning". In: *Bioinformatics* 40.2, btae050. DOI: 10.1093/bioinformatics/btae050.

Brunnsåker, Daniel, Gabriel K. Reder, Nikul K. Soni, Otto I. Savolainen, Alexander H. Gower, Ievgeniia A. Tiukova and Ross D. King (Apr. 2023). "High-throughput metabolomics for the design and validation of a diauxic shift model". In: *npj Systems Biology and Applications* 9.1, pp. 1–9. DOI: 10.1038/s41540-023-00274-9.

Chen, Jiaoyan, Olga Mashkova, Fernando Zhapa-Camacho, Robert Hoehndorf, Yuan He and Ian Horrocks (July 2025). "Ontology Embedding: A Survey of Methods, Applications and Resources". In: *IEEE Transactions on Knowledge and Data Engineering* 37.7, pp. 4193–4212. DOI: 10.1109/TKDE.2025.3559023.

Chen, Rui and Michael Snyder (Oct. 2012). "Systems biology: personalized medicine for the future?" In: *Current Opinion in Pharmacology* 12.5, pp. 623–628. DOI: 10.1016/j.coph.2012.07.011.

Chen, Tianqi and Carlos Guestrin (Aug. 2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16, pp. 785–94. DOI: 10.1145/2939672.2939785.

Cherry, J. Michael, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison et al. (Jan.

2012). “Saccharomyces Genome Database: the genomics resource of budding yeast”. In: *Nucleic Acids Research* 40.Database issue, pp. D700–705. DOI: 10.1093/nar/gkr1029.

Costanzo, Michael, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D. Lee, Vicent Pelechano, Erin B. Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S. Piotrowski, Tharan Srikumar et al. (Sept. 2016). “A global genetic interaction network maps a wiring diagram of cellular function”. In: *Science* 353.6306, aaf1420. DOI: 10.1126/science.aaf1420.

Coutant, Anthony, Katherine Roper, Daniel Trejo-Banos, Dominique Bouthinon, Martin Carpenter, Jacek Grzebyta, Guillaume Santini, Henry Soldano, Mohamed Elati, Jan Ramon, Celine Rouveirol, Larisa N. Soldatova and Ross D. King (Sept. 2019). “Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast”. In: *Proceedings of the National Academy of Sciences* 116.36, pp. 18142–18147. DOI: 10.1073/pnas.1900548116.

Dasgupta, Abhijit and Rajat K. De (Jan. 2023). “Chapter 6 - Artificial intelligence in systems biology”. In: *Handbook of Statistics*. Ed. by Steven G. Krantz, Arni S. R. Srinivasa Rao and C. R. Rao. Vol. 49. Artificial Intelligence. Elsevier, pp. 153–201. DOI: 10.1016/bs.host.2023.06.004.

Dehaspe, Luc and Hannu Toivonen (Mar. 1999). “Discovery of frequent DATALOG patterns”. In: *Data Mining and Knowledge Discovery* 3.1, pp. 7–36. DOI: 10.1023/A:1009863704807.

Dettmer, Katja, Pavel A. Aronov and Bruce D. Hammock (2007). “Mass spectrometry-based metabolomics”. In: *Mass Spectrometry Reviews* 26.1, pp. 51–78. DOI: 10.1002/mas.20108.

Duan, Shou-Fu, Pei-Jie Han, Qi-Ming Wang, Wan-Qiu Liu, Jun-Yan Shi, Kuan Li, Xiao-Ling Zhang and Feng-Yan Bai (July 2018). “The origin and adaptive evolution of domesticated populations of yeast from Far East Asia”. In: *Nature Communications* 9.1, p. 2690. DOI: 10.1038/s41467-018-05106-7.

Ea, Winzeler, Shoemaker Dd, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke Jd, Bussey H, Chu Am, Connelly C, Davis K, Dietrich F, Dow Sw, El Bakkoury M, Foury F, Friend Sh, Gentalen E, Giaever G et al. (Aug. 1999). “Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis”. In: *Science* 285.5429, pp. 901–906. DOI: 10.1126/science.285.5429.901.

Engel, Stacia R, Suzi Aleksander, Robert S Nash, Edith D Wong, Shuai Weng, Stuart R Miyasato, Gavin Sherlock and J Michael Cherry (Mar. 2025). “Saccharomyces Genome Database: advances in genome annotation, expanded biochemical pathways, and other key enhancements”. In: *Genetics* 229.3, iyae185. DOI: 10.1093/genetics/iyae185.

Frey, Carl Benedikt and Michael A. Osborne (Jan. 2017). “The future of employment: How susceptible are jobs to computerisation?” In: *Technological Forecasting and Social Change* 114, pp. 254–280. DOI: 10.1016/j.techfore.2016.08.019.

Geeleher, Paul, Nancy J. Cox and R. Stephanie Huang (Mar. 2014). "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines". In: *Genome Biology* 15.3, R47. DOI: 10.1186/gb-2014-15-3-r47.

Geistlinger, Ludwig, Gergely Csaba, Simon Dirmeier, Robert Küffner and Ralf Zimmer (Oct. 2013). "A comprehensive gene regulatory network for the diauxic shift in Saccharomyces cerevisiae". In: *Nucleic Acids Research* 41.18, pp. 8452–8463. DOI: 10.1093/nar/gkt631.

Ghareeb, Ali Essam, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks and Samuel G. Rodriques (May 2025). *Robin: A multi-agent system for automating scientific discovery.* DOI: 10.48550/arXiv.2505.13400. (Visited on 20/07/2025).

Giaever, Guri, Angela M. Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno André, Adam P. Arkin, Anna Astromoff, Mohamed El Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer et al. (July 2002). "Functional profiling of the Saccharomyces cerevisiae genome". In: *Nature* 418.6896, pp. 387–391. DOI: 10.1038/nature00935.

Gillespie, Marc, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman et al. (Jan. 2022). "The reactome pathway knowledgebase 2022". In: *Nucleic Acids Research* 50.D1, pp. 687–692. DOI: 10.1093/nar/gkab1028.

Gligorijević, Vladimir and Nataša Pržulj (Nov. 2015). "Methods for biological data integration: perspectives and challenges". In: *Journal of The Royal Society Interface* 12.112, p. 20150571. DOI: 10.1098/rsif.2015.0571.

Glish, Gary L. and Richard W. Vachet (Feb. 2003). "The basics of mass spectrometry in the twenty-first century". In: *Nature Reviews Drug Discovery* 2.2, pp. 140–150. DOI: 10.1038/nrd1011.

Goffeau, A., 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 (Oct. 1996). "Life with 6000 Genes". In: *Science* 274.5287, pp. 546–567. DOI: 10.1126/science.274.5287.546.

Golriz Khatami, Sepehr, Sarah Mubeen, Vinay Srinivas Bharadhwaj, Alpha Tom Kodamullil, Martin Hofmann-Apitius and Daniel Domingo-Fernández (Oct. 2021). "Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures". In: *npj Systems Biology and Applications* 7.1, p. 40. DOI: 10.1038/s41540-021-00199-1.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander (Oct. 1999). "Molecular Classification of Cancer: Class

Discovery and Class Prediction by Gene Expression Monitoring". In: *Science* 286.5439, pp. 531–537. DOI: 10.1126/science.286.5439.531.

Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias et al. (Feb. 2025). *Towards an AI co-scientist*. DOI: 10.48550/arXiv.2502.18864. (Visited on 08/05/2025).

Haas, Robert, Aleksej Zelezniak, Jacopo Iacovacci, Stephan Kamrad, StJohn Townsend and Markus Ralser (Dec. 2017). "Designing and interpreting 'multi-omic' experiments that may change our understanding of biology". In: *Current Opinion in Systems Biology* 6, pp. 37–45. DOI: 10.1016/j.coisb.2017.08.009.

Harrieder, Eva-Maria, Fleming Kretschmer, Sebastian Böcker and Michael Witting (Jan. 2022). "Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics". In: *Journal of Chromatography B* 1188, p. 123069. DOI: 10.1016/j.jchromb.2021.123069.

Hastings, Janna, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes and Christoph Steinbeck (Jan. 2016). "ChEBI in 2016: Improved services and an expanding collection of metabolites". In: *Nucleic acids research* 44.1, pp. 1214–9. DOI: 10.1093/nar/gkv1031.

Heim, Noel A., Jonathan L. Payne, Seth Finnegan, Matthew L. Knope, Michał Kowalewski, S. Kathleen Lyons, Daniel W. McShea, Philip M. Novack-Gottshall, Felisa A. Smith and Steve C. Wang (June 2017). "Hierarchical complexity and the size limits of life". In: *Proceedings of the Royal Society B: Biological Sciences* 284.1857, p. 20171039. DOI: 10.1098/rspb.2017.1039.

Holland, Ian and Jamie A. Davies (Nov. 2020). "Automation in the Life Science Research Laboratory". In: *Frontiers in Bioengineering and Biotechnology* 8. DOI: 10.3389/fbioe.2020.571777.

Hong, Kuk-Ki and Jens Nielsen (Aug. 2012). "Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries". In: *Cellular and Molecular Life Sciences* 69.16, pp. 2671–2690. ISSN: 1420-9071. DOI: 10.1007/s00018-012-0945-1.

Kanehisa, M. and S. Goto (Jan. 2000). "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30. DOI: 10.1093/nar/28.1.27.

Kanehisa, Minoru (2019). "Toward understanding the origin and evolution of cellular organisms". In: *Protein Science* 28.11, pp. 1947–1951. DOI: 10.1002/pro.3715.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Masayuki Kawashima and Mari Ishiguro-Watanabe (Jan. 2023). "KEGG for taxonomy-based analysis of pathways and genomes". In: *Nucleic Acids Research* 51.1, pp. 587–92. DOI: 10.1093/nar/gkac963.

Karczewski, Konrad J. and Michael P. Snyder (May 2018). "Integrative omics for health and disease". In: *Nature Reviews Genetics* 19.5, pp. 299–310. DOI: 10.1038/nrg.2018.4.

Karp, Peter D., Richard Billington, Ron Caspi, Carol A. Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M. Keseler, Markus Krummenacker, Peter E. Midford, Quang Ong, Wai Kit Ong, Suzanne M. Paley and Pallavi Subhraveti (July 2019). "The BioCyc collection of microbial genomes and metabolic pathways". In: *Briefings in Bioinformatics* 20.4, pp. 1085–1093. DOI: 10.1093/bib/bbx085.

King, Ross D., Maria Liakata, Chuan Lu, Stephen G. Oliver and Larisa N. Soldatova (Apr. 2011). "On the formalization and reuse of scientific research". In: *Journal of The Royal Society Interface* 8.63, pp. 1440–1448. DOI: 10.1098/rsif.2011.0029.

King, Ross D., Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan and Amanda Clare (Apr. 2009). "The Automation of Science". In: *Science* 324.5923, pp. 85–89. DOI: 10.1126/science.1165620.

King, Ross D., Teresa Scassa, Stefan Kramer and Hiroaki Kitano (Feb. 2024). "Stockholm declaration on AI ethics: why others should sign". In: *Nature* 626.8000, pp. 716–716. DOI: 10.1038/d41586-024-00517-7.

King, Ross D., Ashwin Srinivasan and Luc Dehaspe (Feb. 2001). "Warmr: a data mining tool for chemical data". In: *Journal of Computer-Aided Molecular Design* 15.2, pp. 173–81. DOI: 10.1023/A:1008171016861.

King, Ross D., Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell and Stephen G. Oliver (Jan. 2004). "Functional genomic hypothesis generation and experimentation by a robot scientist". In: *Nature* 427.6971, pp. 247–252. ISSN: 1476-4687. DOI: 10.1038/nature02236.

Kitano, Hiroaki (Mar. 2002). "Systems Biology: A Brief Overview". In: *Science* 295.5560, pp. 1662–1664. DOI: 10.1126/science.1069492.

Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan and Orion Reblitz-Richardson (Sept. 2020). *Captum: A unified and generic model interpretability library for PyTorch*. DOI: 10.48550/arXiv.2009.07896. (Visited on 10/07/2025).

Kramer, Stefan, Nada Lavrač and Peter Flach (2001). "Propositionalization Approaches to Relational Data Mining". In: *Relational Data Mining*. Berlin, Heidelberg: Springer, pp. 262–91. ISBN: 978-3-662-04599-2.

Kuhn, Thomas S. (Apr. 2012). *The Structure of Scientific Revolutions: 50th Anniversary Edition*. Ed. by Ian Hacking. Chicago, IL: University of Chicago Press. ISBN: 978-0-226-45812-0.

Kuzmin, Elena, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda van Leeuwen, Elizabeth N. Koch, Carles Pons, Andrius J. Dagilis, Michael Pryszlak, Zi Yang Wang, Julia Hanchard, Margot Riggi, Kaicong Xu, Hamed Heydari, Bryan-Joseph San Luis et al. (Apr. 2018). "Systematic analysis of complex genetic interactions". In: *Science* 360.6386, eaao1729. DOI: 10.1126/science.aao1729.

Langley, Pat (Aug. 1979). "Rediscovering physics with BACON.3". In: *Proceedings of the 6th international joint conference on Artificial intelligence - Volume 1*. IJCAI'79. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 505–507. ISBN: 978-0-934613-47-7.

Lanucara, Francesco, Stephen W. Holman, Christopher J. Gray and Claire E. Eyers (Apr. 2014). "The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics". In: *Nature Chemistry* 6.4, pp. 281–294. DOI: 10.1038/nchem.1889.

Lee, Tong Ihn, Nicola J. Rinaldi, François Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Thompson, Itamar Simon, Julia Zeitlinger, Ezra G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Ernest Fraenkel, David K. Gifford et al. (Oct. 2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae". In: *Science* 298.5594, pp. 799–804. DOI: 10.1126/science.1075090.

Li, Feiran, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin K. M. Engqvist, Eduard J. Kerkhoven and Jens Nielsen (Aug. 2022). "Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction". In: *Nature Catalysis* 5.8, pp. 662–672. DOI: 10.1038/s41929-022-00798-z.

Li, Zhongxiao, Elva Gao, Juexiao Zhou, Wenkai Han, Xiaopeng Xu and Xin Gao (Jan. 2023). "Applications of deep learning in understanding gene regulation". In: *Cell Reports Methods* 3.1, p. 100384. DOI: 10.1016/j.crmeth.2022.100384.

Liberti, Maria V. and Jason W. Locasale (Mar. 2016). "The Warburg Effect: How Does it Benefit Cancer Cells?" In: *Trends in Biochemical Sciences* 41.3, pp. 211–218. DOI: 10.1016/j.tibs.2015.12.001.

Lindsay, Robert K., Bruce G. Buchanan, Edward A. Feigenbaum and Joshua Lederberg (June 1993). "DENDRAL: A case study of the first expert system for scientific hypothesis formation". In: *Artificial Intelligence* 61.2, pp. 209–261. DOI: 10.1016/0004-3702(93)90068-M.

Lobentanzer, Sebastian, Shaohong Feng, Noah Bruderer, Andreas Maier, Cankun Wang, Jan Baumbach, Jorge Abreu-Vicente, Nils Krehl, Qin Ma, Thomas Lemberger and Julio Saez-Rodriguez (Feb. 2025). "A platform for the biomedical application of large language models". In: *Nature Biotechnology* 43.2, pp. 166–169. DOI: 10.1038/s41587-024-02534-3.

Loscalzo, Joseph and Albert-Laszlo Barabasi (2011). "Systems biology and the future of medicine". In: *WIREs Systems Biology and Medicine* 3.6, pp. 619–627. DOI: 10.1002/wsbm.144.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune and David Ha (Sept. 2024). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. DOI: 10.48550/arXiv.2408.06292. (Visited on 21/07/2025).

Lu, Hongzhong, Feiran Li, Benjamín J. Sánchez, Zhengming Zhu, Gang Li, Iván Domenzain, Simonas Marcišauskas, Petre Mihail Anton, Dimitra Lappa, Christian Lieven, Moritz Emanuel Beber, Nikolaus Sonnenschein, Eduard J. Kerkhoven and Jens Nielsen (Aug. 2019). "A consensus S. cerevisiae

metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism". In: *Nature Communications* 10.1, p. 3586. DOI: 10.1038/s41467-019-11581-3.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal and Su-In Lee (Jan. 2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Margolin, Adam A., Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera and Andrea Califano (Mar. 2006). "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bioinformatics* 7.1, S7. DOI: 10.1186/1471-2105-7-S1-S7.

Messner, Christoph B., Vadim Demichev, Julia Muenzner, Simran K. Aulakh, Natalie Barthel, Annika Röhl, Lucía Herrera-Domínguez, Anna-Sophia Egger, Stephan Kamrad, Jing Hou, Guihong Tan, Oliver Lemke, Enrica Calvani, Lukasz Szyrwiel, Michael Mülleder, Kathryn S. Lilley, Charles Boone, Georg Kustatscher and Markus Ralser (Apr. 2023). "The proteomic landscape of genome-wide genetic perturbations". In: *Cell* 186.9, pp. 2018–34. DOI: 10.1016/j.cell.2023.03.026.

Messner, Christoph B., Vadim Demichev, Ziyue Wang, Johannes Hartl, Georg Kustatscher, Michael Mülleder and Markus Ralser (Nov. 2022). "Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology". In: *Proteomics* 23. DOI: 10.1002/pmic.202200013.

Mi, Huaiyu, Anushya Muruganujan, John T. Casagrande and Paul D. Thomas (Aug. 2013). "Large-scale gene function analysis with PANTHER Classification System". In: *Nature protocols* 8.8, pp. 1551–66. DOI: 10.1038/nprot.2013.092.

Mo, Monica L., Bernhard Ø Palsson and Markus J. Herrgård (Mar. 2009). "Connecting extracellular metabolomic measurements to intracellular flux states in yeast". In: *BMC Systems Biology* 3.1, p. 37. DOI: 10.1186/1752-0509-3-37.

Monge, María Eugenia, James N. Dodds, Erin S. Baker, Arthur S. Edison and Facundo M. Fernández (June 2019). "Challenges in Identifying the Dark Molecules of Life". In: *Annual review of analytical chemistry (Palo Alto, Calif.)* 12.1, pp. 177–199. DOI: 10.1146/annurev-anchem-061318-114959.

Muggleton, Stephen (Oct. 1999). "Inductive Logic Programming: Issues, results and the challenge of Learning Language in Logic". In: *Artificial Intelligence* 114.1, pp. 283–296. DOI: 10.1016/S0004-3702(99)00067-3.

Muggleton, Stephen and Luc De Raedt (1994). "Inductive logic programming: Theory and methods". In: *The Journal of Logic Programming* 19, pp. 629–679.

Muggleton, Stephen, Luc De Raedt, David Poole, Ivan Bratko, Peter Flach, Katsumi Inoue and Ashwin Srinivasan (Jan. 2012). "ILP turns 20". In: *Machine Learning* 86.1, pp. 3–23. DOI: 10.1007/s10994-011-5259-2.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware and John P. A. Ioannidis (Jan. 2017). "A manifesto for reproducible science". In: *Nature Human Behaviour* 1.1, p. 0021. DOI: 10.1038/s41562-016-0021.

Murphy, Kevin Patrick (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. University of California, Berkeley.

Musslick, Sebastian, Laura K. Bartlett, Suyog H. Chandramouli, Marina Dubova, Fernand Gobet, Thomas L. Griffiths, Jessica Hullman, Ross D. King, J. Nathan Kutz, Christopher G. Lucas, Suhas Mahesh, Franco Pestilli, Sabina J. Sloman and William R. Holmes (Feb. 2025). "Automating the practice of science: Opportunities, challenges, and implications". In: *Proceedings of the National Academy of Sciences* 122.5, e2401238121. DOI: 10.1073/pnas.2401238121.

Nielsen, Jens and Michael C. Jewett (Feb. 2008). "Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae". In: *FEMS Yeast Research* 8.1, pp. 122–131. DOI: 10.1111/j.1567-1364.2007.00302.x.

Orhobor, Oghenejokpeme I., Joseph French, Larisa N. Soldatova and Ross D. King (2020). "Generating Explainable and Effective Data Descriptors Using Relational Learning: Application to Cancer Biology". In: *Discovery Science*. Ed. by Annalisa Appice, Grigorios Tsoumakas, Yannis Manolopoulos and Stan Matwin, pp. 374–85. ISBN: 978-3-030-61527-7.

Orth, Jeffrey D., Ines Thiele and Bernhard Ø Palsson (Mar. 2010). "What is flux balance analysis?" In: *Nature Biotechnology* 28.3, pp. 245–248. DOI: 10.1038/nbt.1614.

Paglia, Giuseppe, Andrew J. Smith and Giuseppe Astarita (2022). "Ion mobility mass spectrometry in the omics era: Challenges and opportunities for metabolomics and lipidomics". In: *Mass Spectrometry Reviews* 41.5, pp. 722–765. DOI: 10.1002/mas.21686.

Pinu, Farhana R. and Silas G. Villas-Boas (Aug. 2017). "Extracellular Microbial Metabolomics: The State of the Art". In: *Metabolites* 7.3, p. 43. DOI: 10.3390/metabo7030043.

Reder, Gabriel K., Erik Y. Bjurström, Daniel Brunnsåker, Filip Kronström, Praphapan Lasin, Ievgeniia Tiukova, Otto I. Savolainen, James N. Dodds, Jody C. May, John P. Wikswo, John A. McLean and Ross D. King (Mar. 2024). "AutonoMS: Automated Ion Mobility Metabolomic Fingerprinting". In: *Journal of the American Society for Mass Spectrometry* 35.3, pp. 542–550. DOI: 10.1021/jasms.3c00396.

Reder, Gabriel K., Carl Collins, Abbi Abdel Rehim, Larisa Soldatova and Ross D. King (2025). "LLM-retrieval based scientific knowledge grounding". In: *2nd International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025)*. URL: https://ceur-ws.org/Vol-3977/NSLP-01.pdf.

Regenmortel, Marc H.V. Van (Nov. 2004). "Reductionism and complexity in molecular biology". In: *EMBO Reports* 5.11, pp. 1016–1020. DOI: 10.1038/sj.embor.7400284.

Roper, Katherine, A. Abdel-Rehim, Sonya Hubbard, Martin Carpenter, Andrey Rzhetsky, Larisa Soldatova and Ross D. King (Apr. 2022). "Testing the reproducibility and robustness of the cancer biology literature by robot". In: *Journal of The Royal Society Interface* 19.189, p. 20210821. DOI: 10.1098/rsif.2021.0821.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina and Anshul Kundaje (Apr. 2017). *Not Just a Black Box: Learning Important Features Through Propagating Activation Differences*. DOI: 10.48550/arXiv.1605.01713. (Visited on 15/07/2025).

Song, Tao, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Guoqing Zhang, Fei Zhang, Weiwei Shang, Yao Fu, Jun Jiang and Yi Luo (Apr. 2025). "A Multiagent-Driven Robotic AI Chemist Enabling Autonomous Chemical Research On Demand". In: *Journal of the American Chemical Society* 147.15, pp. 12534–12545. DOI: 10.1021/jacs.4c17738.

Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen and Christian von Mering (Jan. 2023). "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest". In: *Nucleic Acids Research* 51.D1, pp. D638–D646. DOI: 10.1093/nar/gkac1000.

Szklarczyk, Damian, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork and Michael Kuhn (Jan. 2016). "STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data". In: *Nucleic Acids Research* 44.D1, pp. D380–D384. DOI: 10.1093/nar/gkv1277.

Tavassoly, Iman, Joseph Goldfarb and Ravi Iyengar (Oct. 2018). "Systems biology primer: the basic methods and approaches". In: *Essays in Biochemistry* 62.4, pp. 487–500. DOI: 10.1042/EBC20180003.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin (Dec. 2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., pp. 6000–6010. ISBN: 978-1-5108-6096-4.

Williams, Kevin, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N. Soldatova, Kurt De Grave, Jan Ramon, Michaela de Clare, Worachart Sirawaraporn, Stephen G. Oliver and Ross D. King (Mar. 2015). "Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases". In: *Journal of the Royal Society, Interface* 12.104, p. 20141289. DOI: 10.1098/rsif.2014.1289.

Wood, Valerie, Antonia Lock, Midori A. Harris, Kim Rutherford, Jürg Bähler and Stephen G. Oliver (Feb. 2019). "Hidden in plain sight: what remains to be discovered in the eukaryotic proteome?" In: *Open Biology* 9.2, p. 180241. DOI: 10.1098/rsob.180241.

Yang, Jason H., Sarah N. Wright, Meagan Hamblin, Douglas McCloskey, Miguel A. Alcantar, Lars Schrübbers, Allison J. Lopatkin, Sangeeta Satish, Amir Nili, Bernhard O. Palsson, Graham C. Walker and James J. Collins (May 2019). "A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action". In: *Cell* 177.6, 1649–1661.e9. DOI: 10.1016/j.cell.2019.04.016.

Yang, Zhiliang and Mark Blenner (Dec. 2020). "Genome editing systems across yeast species". In: *Current Opinion in Biotechnology*. Tissue, Cell and Pathway Engineering 66, pp. 255–266. DOI: 10.1016/j.copbio.2020.08.011.

Zhang, Aihua, Hui Sun, Hongying Xu, Shi Qiu and Xijun Wang (Oct. 2013). "Cell Metabolomics". In: *OMICS : a Journal of Integrative Biology* 17.10, pp. 495–501. DOI: 10.1089/omi.2012.0090.

Zhang, Chengyu, Benjamín J Sánchez, Feiran Li, Cheng Wei Quan Eiden, William T Scott, Ulf W Liebal, Lars M Blank, Hendrik G Mengers, Mihail Anton, Albert Tafur Rangel, Sebastián N Mendoza, Lixin Zhang, Jens Nielsen, Hongzhong Lu and Eduard J Kerkhoven (Oct. 2024). "Yeast9: a consensus genome-scale metabolic model for S. cerevisiae curated by the community". In: *Molecular Systems Biology* 20.10, pp. 1134–1150. DOI: 10.1038/s44320-024-00060-7.