Thesis for The Degree of Licentiate of Engineering

Frameworks for Automated Discovery in Systems Biology

First-order logic models of metabolism and ontology-driven databases

Alexander H. Gower

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2024

Frameworks for Automated Discovery in Systems Biology
First-order logic models of metabolism and ontology-driven databases

Alexander H. Gower

ISSN 1652-876X

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

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2024.

To my teachers.

Frameworks for Automated Discovery in Systems Biology

First-order logic models of metabolism and ontology-driven databases

Alexander H. Gower

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

Abstract

Systems biology is an integrationist approach to biological science, meaning we treat organisms as complex systems whose behaviour is dictated by the interaction of their constituent parts. Because eukaryotic organisms are extremely complex systems, research progress in systems biology can be slow. Recent advances in robotics, and more importantly in artificial intelligence (AI), offer great opportunity for automating scientific discovery in this field.

Using the model organism *Saccharomyces cerevisiae*, baker's yeast, this thesis explores: the philosophical and practical motivations for the use of automation in biological research; the structure of knowledge models, experimental data, and hypotheses in systems biology; and computational models of metabolism, a core component of systems biology.

The first main contribution of this thesis is a set of ontologies and accompanying database software for enabling an autonomous discovery platform. The second main contribution is a first-order logic framework for modelling cellular physiology, which we call LGEM$^+$. Abduction of hypotheses for improvement of knowledge models is enabled by LGEM$^+$, which couples a set of predicates and clauses expressing biochemical reaction processes with an efficient automated theorem prover (ATP), iProver.

Results from these studies show automated improvement of knowledge models in systems biology can be achieved using general purpose tools, in this case ATPs, by using a first-order logic formalism faithful to domain ontologies. More work is needed to integrate these techniques with laboratory robotics and inductive reasoning agents, building on the work presented in this thesis, to achieve the goal of autonomous discovery in systems biology.

Keywords

Machine learning, first-order logic, abduction, automated theorem provers, knowledge modelling, ontologies, systems biology, metabolic modelling

List of Publications

Appended Publications

This thesis is based on the following publications:

[**Paper I**] **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.

[**Paper II**] G. K. Reder, **A. H. Gower**, F. Kronström, R. Halle, V. Mahamuni, A. Patel, H. Hayatnagarkar, L. N. Soldatova, R. D. King, *Genesis-DB: a database for autonomous laboratory systems*.
Bioinformatics Advances, Volume 3, Issue 1 (August 2023).
https://doi.org/10.1093/bioadv/vbad102

[**Paper III**] F. Kronström, **A. H. Gower**, I. A. Tiukova, R. D. King, *RIMBO - An Ontology for Model Revision Databases*.
In International Conference on Discovery Science (pp. 523–534). Cham: Springer Nature Switzerland. (October 2023).
https://doi.org/10.1007/978-3-031-45275-8_35

[**Paper IV**] **A. H. Gower**, K. Korovin, D. Brunnsåker, E. Y. Bjurström, P. Lasin, I. A. Tiukova, R. D. King, *LGEM^{+}: Automated Improvement of Metabolic Network Models and Model-Driven Experimental Design through Abduction*.
Under Submission.

Other Publications

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

[**a**] 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 1, Article number 11 (April 2023). `https://doi.org/10.1038/s41540-023-00274-9`

[**b**] **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 (pp. 628–643). Cham: Springer Nature Switzerland. (October 2023). `https://doi.org/10.1007/978-3-031-45275-8_42`

Acknowledgment

My first thanks are to my supervisor, Professor Ross King, for the opportunity to pursue such interesting research, for the academic direction, and for the many valuable discussions we have had in the past three years. Thank you to my co-supervisor, Dr Ievgeniia Tiukova, for our discussions and her feedback, and for her encouragement and support. Thank you to my examiner, Professor David Sands, for his advice in the preparation of this thesis.

Thank you to all the co-authors of the papers included in this thesis. I am particularly grateful to Dr Konstantin Korovin for the collaboration we have shared and the informal supervision he has given me. Thanks to the members of the King Lab—Daniel, Erik, Filip, and Dr Praphapan Lasin (Beera)—for being such excellent colleagues, for our discussions, and for all the feedback on the drafts for this thesis and my other projects. Thanks to Dr Gabriel Reder for our time as colleagues, and for his guidance and advice on navigating life as a PhD student.

Thank you to everyone at the division of Data Science and AI for making me feel welcome, and for a lively and supportive environment in which to pursue my research. In particular, thanks to the PhD students and post-docs for sustaining an enthusiastic and diverse academic community, and for inspiring one another.

Thank you to the Chalmers Library and its staff for access to many of the resources that informed this thesis. Thank you to the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Alice Wallenberg Foundation for supporting my PhD studies.

Thanks to my friends for your support and for making our world a better place. To Annick, Lennart, and Nyasha for helping me feel welcome in Sweden and your family. For your courage, our joy and our strength, thank you to my siblings, Tom and Anna, and to Ella. To Anaëlle, for walking beside me. To my parents, John and Diana, for your love, ever felt.

Contents

II Appended Papers 29

Part I

Introductory Chapters

Chapter 1

Introduction

Progression toward understanding eukaryote biology is one of the most important areas of modern scientific effort. The 20th century saw many advances in our understanding of the fundamental components and processes of eukaryotic life. In turn, society has benefited greatly from application of this knowledge to medicine, agriculture, and engineering. However, we are still without an accurate predictive model of the physiology of one organism, let alone broadly applicable theories and laws for the behaviour of eukaryotic systems such as we have for physics.

Part of the reason why progress in biology is limited by today's scientific methods is the diversity and complexity of the systems. Hundreds of research hours can be spent in the study of one particular gene, yet the limits of human capability and of course the economic resource available to the researcher will hamper progression to a complete understanding of the gene and its roles. Scientific discovery automation has therefore great potential in biology. And this is particularly the case when adopting the systems biology paradigm.

Systems biology is an integrationist approach to biological science, meaning we treat organisms as complex systems whose behaviour is dictated by the interaction of their constituent parts. Research progress in systems biology can be slow, partly because eukaryotic organisms are extremely complex systems. Recent advances in robotics, and more importantly in artificial intelligence (AI), offer great opportunity for automating scientific discovery in this field.

Using the model organism *Saccharomyces cerevisiae*, baker's yeast, this thesis explores: the philosophical and practical motivations for the use of automation in biological research; the structure of knowledge models, experimental data, and hypotheses in systems biology; and computational models of metabolism, a core component of systems biology.

The first main contribution of this thesis is a set of ontologies and accompanying database software for enabling an autonomous discovery platform. The second main contribution is a first-order logic framework for modelling cellular physiology, which we call LGEM$^+$. Abduction of hypotheses for improvement of knowledge models is enabled by LGEM$^+$, which couples a set of predicates and clauses expressing biochemical reaction processes with an efficient automated theorem prover (ATP), iProver.

Results from these studies show automated improvement of knowledge models in systems biology can be achieved using general purpose tools, in this case ATPs, by using a first-order logic formalism faithful to domain ontologies. More work is needed to integrate these techniques with laboratory robotics and inductive reasoning agents, building on the work presented in this thesis, to achieve the goal of autonomous discovery in systems biology.

Structure of this thesis. Chapter 2 introduces the philosophical and pragmatic motivations for automating scientific discovery, and some of the relevant machine learning concepts. Chapter 3 serves as an introduction to the scientific domain: systems biology. We begin with a statement of the overarching problem, and proceed with a discussion of the main parts and processes of biological systems. The remainder of this chapter is dedicated to an overview of different computational modelling frameworks for yeast physiology. Chapter 4 contains a summary of contributions made in the published papers upon which this thesis is based—these papers are appended in Part II. Finally, Chapter 5 is a discussion of the themes and conclusions we can draw from the contributions of this thesis, and of potential future research directions.

Chapter 2

The Automation of Scientific Discovery

Scientific discovery is the generation of new knowledge through organised enquiry. The methods used to generate this knowledge, and the ways that knowledge is stored and communicated, vary widely across domains. However there are common values and tools that have enabled philosophers to characterise scientific enquiry to a certain degree (Schindler 2022).

In **Paper I**, we explore the philosophical background to, and tools for, scientific discovery, which provides important context for the contents of this thesis. **Paper I** also provides an introduction to the automation of scientific discovery and the concept of a robot scientist, defined by King et al. (2009) as

> a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experimentation.

The cycles of experimentation that are executed by robot scientists will vary depending on the domain, but have a core similarity. The general structure of these cycles is to form hypotheses, perform deductive simulations using a computation model to obtain predicted behaviour, and test these predictions by performing experiments and collecting data. With each cycle the robot scientist seeks to improve the quality of its predictions by forming better hypotheses. This is a supervised learning problem, specifically a type of active learning.

This active learning process has many constituent tasks, and thus requires a combination of domain-specific and general-purpose tools. In this thesis we present tools that address several of these tasks, for example: hypothesis generation (**Paper IV**); hypothesis storage (**Paper II**, **Paper III**); simulation (**Paper IV**); experimental data curation (**Paper II**); and model evaluation (**Paper IV**).

In addition to the introduction given in **Paper I**, we now treat a couple of concepts regarding logic and reasoning that are relevant for the contributions of this thesis, in particular **Paper IV**.

2.1 Logic in Science

Logics are mathematical languages that relate premises and conclusions and enable formal reasoning (Ben-Ari 2012). There are various logics that can be used for reasoning in science, and here we present a brief overview of propositional logic, the concept of satisfiability, normal forms, and first-order logic.

Propositional logic

The elementary form of logic is propositional logic, which deals in assigning truth value to statements about a world (propositions). An example of a proposition would be:

> "Uppsala is the capital city of Sweden."

In the world we live in, we would assign this proposition the truth value 'false'. But it is possible to consider a world where this statement would be assigned the truth value 'true'. Indeed if we consider the world as it was in the 14th Century then this proposition would be assigned 'true', and in common relaxation of the language we would say that the proposition is true.

Propositions are atoms; they can be used to build complex formulas using Boolean operators such as negation (`NOT`,$\neg$), conjunction (`AND`, $\wedge$), and disjunction (`OR`, $\vee$).

Satisfiability

An interpretation for a formula is a function that assigns truth values to each atom that appears in the formula. A formula is considered unsatisfiable if it is false in all interpretations. If there is at least one interpretation for which the formula is true, then it is satisfiable. We can extend the notion of satisfiability to a set of formulas naturally, by requiring that there exists an interpretation under which each formula in the set is true.

Normal forms

For any given logical formula there are many possible ways to express an equivalent formula. An example is the following tautology, one of De Morgan's laws:

$$\neg(A \vee B) \iff (\neg A) \wedge (\neg B) \tag{2.1}$$

A way to express logical formulae which, and has advantages for automated theorem proving, is in conjunctive normal form (CNF). To be in CNF, a formula is written as a conjunction of disjunctions of literals (atoms or negated atoms); the right hand side of (2.1) is in CNF. It is possible to express every formula in propositional logic in CNF.

A different notation for CNF is clausal normal form. Instead of using the symbols for disjunction and conjunction to construct the formula, we appeal to the structure of the normal form and express a formula as a set of sets of literals. For example we could express the right hand side of (2.1) in clausal normal form as $\{\{\neg A\}, \{\neg B\}\}$. A theory is a set of closed formulae (formulae with no free variables). Clausal normal form is an efficient representation of logical theories that is often used in automated theorem proving.

First-order logic

First-order logic (FOL) extends propositional logic to allow for relations between variables. Relations are represented using predicate symbols, and quantifiers such as *for all* ($\forall$) and *there exists* ($\exists$) are introduced to allow for formulae that express general statements about relations. Instead of atoms being propositional statements, atoms in FOL are a predicate symbol with a list of arguments. These arguments are either constants or variables in the domain of the relation the predicate symbol denotes. For example, FOL allows for a statement such as:

> "A dog is happy if it has a bone."

This could be expressed with the following logical formula, containing new predicates $\mathsf{dog}\backslash 1$, $\mathsf{bone}\backslash 1$, $\mathsf{has}\backslash 2$, and $\mathsf{happy}\backslash 1$:

$$\forall \mathsf{x} \forall \mathsf{y}(\mathsf{dog}(\mathsf{x}) \land \mathsf{bone}(\mathsf{y}) \land \mathsf{has}(\mathsf{x}, \mathsf{y}) \to \mathsf{happy}(\mathsf{x}))$$

The atoms in this formula are $\mathsf{dog}(\mathsf{x})$, $\mathsf{bone}(\mathsf{y})$, $\mathsf{has}(\mathsf{x}, \mathsf{y})$, and $\mathsf{happy}(\mathsf{x})$.

FOL has many desirable properties for automated reasoning. It is an expressive language and can therefore be used to represent a wide variety of concepts and theories. However, the fact that it is not as expressive as higher-order logics means it is semi-decidable, meaning that there is an efficient procedure for checking if a formula is in a theory (but no such method for checking that a formula is not in a theory).

We stated earlier that it is possible to express every formula in propositional logic in CNF. The same is not true in FOL, however Skolem's theorem states that for any closed formula ϕ there exists a formula ϕ' such that ϕ is satisfiable if and only if ϕ' is satisfiable (Ben-Ari 2012). A useful consequence of this is that we can always map our theory to one in CNF or clausal normal form to assess satisfiability.

2.1.1 Reasoning

Reasoning on logical theories can be broken down into a few main categories. Deductive reasoning derives conclusions from premises and laws. Inductive reasoning and abductive reasoning seek to provide explanations for facts by generating either laws (induction) or facts (abduction); for this we need statistical inference. For a more detailed explanation of the types of reasoning, see **Paper I**, Section 2.1.

2.1.2 Automated theorem provers

Automated theorem provers (ATPs) are software that can perform reasoning tasks on logical theories. A common task for ATPs is deciding whether a conjecture, C, is entailed by a given theory T and optional hypotheses H. In particular in science, the conjecture usually takes the form of some statement rooted in empirical data, and we are interested if our theory, perhaps together with a hypothesis, entails the data. ATPs most often take inputs in the form of FOL theories, and these are commonly written in clausal normal form. Formally, the problem posed to the ATP is:

$$T \wedge H \overset{?}{\models} C.$$

Conjectures are often posed in negated form and then submitted to a SAT solver, an algorithm to decide satisfiability. In which case the proof takes the form of a refutation of the negated form that shows the unsatisfiability of $T \wedge H \wedge \neg C$. Or a demonstration of the satisfiability of $T \wedge H \wedge \neg C$ which shows that C is not entailed by $T \wedge H$ under any interpretation. Constructing such arguments for FOL theories requires heuristic search, and the design and implementation of algorithms for this task is the core activity in developing ATPs for FOL (Korovin 2008).

ATPs are primarily applied to mathematical reasoning tasks (Urban and Vyskočil 2013). But they have also been used in engineering applications including software verification (Georgiou et al. 2022) and hardware verification (Goel and Ray 2022; Khasidashvili et al. 2015).

Chapter 3

Mathematical Models of Yeast Physiology

3.1 Systems Biology

Systems biology is an approach to studying biological systems that aims to understand how the behaviour of the system arises from the interaction of its constituent parts (Kohl et al. 2010). We describe systems biology in detail, and its link to complex systems, in Section 4 of **Paper I**. In this book chapter we also outline why systems biology is a good domain in which to apply the automation of scientific discovery. Partly this is because of the inherent complexity of the systems involved, which limits the progress that can be made with non-automated scientific methods. Robot scientists can manage complexity more precisely and at larger scales than humans, both in experimentation and in analysis.

A core concept in systems biology is the connection of *genotype*—the DNA sequence information of an organism—to its *phenotype*—the observed characteristics of the organism. The traditional reasoning behind treating DNA sequence information as the source for phenotypical observations comes from two hypotheses, proposed by Crick (1958) as the "Sequence Hypothesis", and the so-called "Central Dogma" of molecular biology, that the primary form of information transfer within cells is as follows: DNA is *transcribed* to RNA which is in turn *translated* to proteins. Specifically, the "Central Dogma" stated that information passes from nucleic acids to proteins, but is not transferred from proteins to nucleic acids. Later, Crick (1970) clarified this hypothesis stating that this hypothesis was intended to apply to the general case for living organisms, and that though cases of information transfer from proteins to nucleic acids was theoretically possible there was no evidence for these interactions at the time.

Developments in molecular biology since the 1970s have demonstrated numerous potential violations of this mechanistic view of information transfer between molecules in biological systems. Some examples are: reverse transcription, from RNA to DNA (Baltimore 1970; Temin and Mizutani 1970); and post-translational protein modification, for example through phosphorylation (Krebs and Graves 2000). Opinion is divided on whether these truly represents violations of the original hypothesis, with its detractors raising examples such as those above, and its defenders arguing that the original source of the sequence information remains the DNA.

Regardless, either interpretation justifies structuring enquiry around the genotype-phenotype relationship. Systems biology enables a nuanced approach allowing for feedback loops and incorporating additional classes of molecules, such as sugars and lipids, into a complex system of signalling, gene regulation, and metabolism. We now proceed to discuss each of these concepts in more detail.

3.1.1 Eukaryote cellular metabolism

Metabolism refers to the consumption, transformation and production of chemical compounds by an organism, through various biochemical reactions. Metabolism has three main purposes: to make energy available, to make building blocks for structures, and to eliminate waste.

Biochemical reactions come in various types. Many require active catalysis via enzymes, formed of proteins. Which reactions are feasible in a particular organism is therefore largely determined by its genome. Reactions will be feasible in different organisms, though the specific gene for the reaction may differ as there are often several different enzymes that can catalyse the same class of reaction (isoenzymes).

A metabolic pathway is a set of metabolic reactions, usually a sequence or a cycle. When studying metabolism, one approach is to study pathways, as they facilitate particular functions subordinate to the purposes listed above, yet are often small enough to be tractable for detailed methods and analysis. However pathways do not exist in isolation, so it is also necessary to consider so-called superpathways (collections of pathways) and the metabolism of the whole cell or organism.

In systems biology, the knowledge about metabolism is drawn from various sources and compiled in metabolic network models (MNMs). Metabolic network models are covered in more detail in Section 3.3, and in **Paper IV**.

3.1.2 Gene regulation in eukaryotes

Genes are segments of DNA and the process of synthesising functional products (e.g. proteins) from the DNA is referred to as gene expression. The state of gene expression in a cell dictates which processes occur within the cell. A significant part of this control is through the expression of metabolic genes, those which encode proteins that form enzymes. Eukaryotes regulate genes and their activity in response to environmental stimuli and also within the organism. This is achieved by eukaryotic cells at several levels.

1. Transcription is controlled by limiting the amount of messenger RNA (mRNA) that is produced from a given gene.

2. Post-transcription there are events that regulate the translation of RNA into proteins.

3. Post-translation there are mechanisms which modify proteins, which can affect their activity.

Transcription factors are proteins that recognise and bind to a segment of DNA adjacent to the genes they regulate. There are a variety of processes through which transcription factors regulate genes, but essentially they control the rates of transcription. Messenger RNA (mRNA) transcription most often cannot occur without the help of transcription factors. As the ground state for transcription is restrictive, positive regulation is the predominant form of control. Transcription factors are themselves regulated, resulting in a complex interaction network.

An example of a mechanism of gene regulation specific to eukaryotes is physically restricting access to DNA promoters through the structure of chromatin. DNA is wound tightly on nucleosomes that form the chromatin fibre, and modifications to the chromatin structure can regulate gene expression.

3.1.3 Cell signalling

A cell, whether a single-celled organism such as *Saccharomyces cerevisiae* or one part of a multicellular organism, needs to send and receive signals to interact with its environment, other cells, and indeed itself. This process of cell signalling is enabled by the binding of small molecules known as ligands to an effector molecule, frequently a protein. The binding of the ligand to a particular site on the effector causes a change which allows the effector to perform some function. This could be up- or down-regulating (increasing or decreasing expression of) a particular gene product, or modifying an enzyme complex to increase or decrease its activity. Similarly to gene regulation, signalling is a complex mode of control in eukaryotes, with many individual signalling molecules interacting with each other, and other systems such as metabolism and gene expression.

As a result, though many individual signalling interactions such as the binding of adrenaline to adrenoreceptors are well-studied, cell signalling networks are in general poorly understood in most eukaryotes. Yet they can provide promising explanations for phenomena such as ageing (Greer and Brunet 2008), and effective therapeutic options for diseases, for example leukaemia (Weisberg et al. 2005).

3.2 The Yeast *Saccharomyces cerevisiae* as a Model Organism for Eukaryote Biology

It should by now be clear that eukaryote cellular physiologies are extremely complex systems. Gene regulation and enzymatic catalysis are but two of many processes that enable information propagation and feedback, both within cells and across cell boundaries; and the entities and processes involved occur across a huge range of timescales and length scales.

As discussed in **Paper I**, when studying systems directly is impractical or infeasible, we should use a model. This is often the case with eukaryotes; for example, there are many experiments that would be undesirable or immoral to conduct using human subjects, and *in vitro* experiments often do not recreate the inherent complexity of the object systems. And the yeast *Saccharomyces cerevisiae* is the model organism for the eukaryotic cell, for a number of reasons.

Firstly, there are tools available for easy genetic manipulation of yeast (and fewer ethical and legal issues in doing so than with higher eukaryotes). Cultivation cost is relatively low, in terms of the key resources: money, time, space and human resource. And *S. cerevisiae*'s was the first eukaryotic genome to be fully sequenced (Goffeau et al. 1996). There is also a wealth of experience and knowledge on *S. cerevisiae* that can be used as a rich prior for discovery.

Using *S. cerevisiae* as a model organism means that we aim to understand processes relevant to other eukaryotes through yeast. These could be evolutionarily conserved functions, or it could be that we transplant genes that we wish to study. Yeast is also heavily used for bioengineering purposes, where genotype and conditions are manipulated to efficiently produce a desired product, for example a pharmaceutical.

3.3 Modelling Frameworks

Models of yeast vary from deterministic and high-resolution to descriptive or pedagogical. As discussed in **Paper I**, there are various desirable qualities of scientific models, including: predictive power; parsimony; explanatory usefulness; consistency across contexts; and consistency with different scientific models. For most models of *S. cerevisiae* it is desirable to be able to exploit the models to make predictions about real-world behaviour of the system, and we will discuss briefly some of the mathematical frameworks commonly used.

3.3.1 Differential equations models

One common technique is to model the abundances of genes, proteins and chemical species using systems of coupled ordinary differential equations (ODEs) with time as the dynamic variable. These models have been successfully employed to model various biological processes including: central carbon metabolism; batch fermentative growth; and toxicity responses. The differential equations are often based around reaction kinetics paradigms, commonly Michaelis-Mentin kinetics. A great challenge with these models is parametrisation; the models can have tens of thousands of parameters, only a fraction of which have experimentally obtained values, which are often condition specific. Machine learning techniques have recently been employed to predict these parameter values.

Another challenge is that timescales and length scales across the different systems involved vary across at least five orders of magnitude, from the molecular scale (1×10^{-10} m) to the cellular scale (1×10^{-5} m) (Castiglione et al. 2014; Southern et al. 2008). ODE models based on Michaelis-Mentin kinetics do not explicitly model the length dimension, but the timescale of reaction kinetics can vary over orders of magnitude depending on the reaction (Resat et al. 2009), and is vastly quicker than the timescale of gene expression (Carthew 2021).

A third challenge is that biological processes are stochastic in nature, so deterministic differential equations models will likely be unable to capture a great deal of behaviour. For certain classes of molecules the assumption of continuous concentrations may not be valid due to the low count of molecules and their compartmentalisation (Resat et al. 2009).

3.3.2 Constraint-based modelling

Another popular modelling framework is a constrained optimisation method, where the objective function is defined as some real-valued function representing a biologically realistic quantity. Examples of commonly used objectives are growth maximisation or minimisation of uptake of a particular carbon source (Orth et al. 2010; García Sánchez and Torres Sáez 2014). Other knowledge about the cell is encoded via constraints on this optimisation. The fundamental constraints are the rates of flux through chemical reactions in the cell. A common approach is to assume the system is in a steady state and the sum of the fluxes for any given chemical species is zero, hence the technique is known as flux balance analysis (FBA). A more detailed explanation of the mathematics of FBA is provided in **Paper IV**.

Various extensions and modifications to FBA have been proposed that relax or work around this assumption, for example: dynamic flux balance analysis (dFBA) which couples FBA solutions with ordinary differential equations (Mahadevan et al. 2002); flux variability analysis (FVA) which characterises the space of fluxes that give rise to an optimal solution (Mahadevan and Schilling 2003); and two techniques, regulatory on/off minimisation (ROOM, Shlomi et al. 2005) and minimisation of metabolic adjustment (MOMA, Segrè et al. 2002) which seek to find a likely flux distribution—in comparison to a reference distribution—after a genotypic change. Some techniques extend FBA to more nuanced constraints, for example through the inclusion of setting reaction flux constraints through abundance of relevant enzymes (ecFBA, Sánchez et al. 2017).

3.3.3 Logical models

Formal mathematical logic has also been used to model different parts of cellular biology. A common form of logic modelling is propositional logic (Boolean) models of gene regulation and signalling. First-order logic models of metabolism have also been developed previously. For more detail see the introduction to **Paper IV**.

3.3.4 Hybrid models

Some approaches seek to extend FBA models yet further by integrating them with separate models for other cellular processes, for example genetic regulation and signalling processes. These hybrid models employ different mathematical formalisms for different cellular processes and rely on bespoke techniques at the boundary between models to integrate them together. One example of a hybrid model is presented in Österberg et al. (2020), which combined enzyme constrained FBA with a Boolean signalling model to predict non-trivial yeast phenotypes. A review of hybrid modelling approaches in systems biology was conducted by Cruz and Kemp (2021), showing a diverse approach to modelling, including combinations of modelling techniques listed here and others. And these models were to applied a variety of biological applications.

Chapter 4

Summary of Included Papers

4.1 Paper I: The Use Of AI-Robotic Systems For Scientific Discovery

In this book chapter, we aim to provide machine learning researchers with an introduction to the research problem of the automation of scientific discovery. We address considerations for the design of robot scientists to achieve this aim, systems that combine artificial intelligence with robotics. We begin by examining the scientific method using concepts and models from the philosophy of science. We define the concept of a scientific model and their use by robot scientists. We present three components of the scientific method—logical inference, statistical inference, and parsimony—and how they are applied to the development of scientific models. We finish the opening section by presenting a set of scientific values for examining the quality of a given model which enables comparison between competing models, the central aspect of scientific method that allows for progress.

Section 3 of this book chapter analyses scientific discovery using machine learning paradigms. We discuss how aspects of machine learning algorithms could be mapped to aspects of the scientific method. We conclude that scientific discovery should be viewed as a supervised learning problem, with input-output pairs for training data coming either from controlled experiments or from observational studies. The material in this section is highly relevant to Chapter 2 of this thesis. Section 4 is primarily an introduction to the domain of systems biology, covered in Chapter 3 of this thesis. We also cover two examples of robot scientists being applied in systems biology: ADAM (King et al. 2009) and Eve (Williams et al. 2015). Section 5 is a case study using the example of the robot scientist Genesis and LGEM$^+$, covered in Chapter 3 of this thesis and in appended **Paper IV** respectively.

This manuscript has been submitted for consideration for publication in a volume aimed at introducing the automation of scientific discovery, primarily for machine learning researchers. This chapter is also a partial introduction to the contents of this thesis, and is recommended to be read first as it provides context to other matters discussed in the introductory chapters and in the appended papers.

Author contributions

The conceptualisation of the project was done by Ross D. King, and Alexander H. Gower. Content included in this chapter was informed by discussions between **A.H.G.** and each of the co-authors (Konstantin Korovin, Daniel Brunnsåker, Filip Kronström, Gabriel K. Reder, Ronald S. Reiserer and John P. Wikswo). The manuscript was written by **A.H.G.** and edited by D.B. and R.D.K. The project was supervised by R.D.K., and Ievgeniia A. Tiukova. The funding for the project was acquired by R.D.K.

4.2 Paper II: Genesis-DB: a database for autonomous laboratory systems

Problem

As discussed in Chapter 2, a robot scientist is a combination of laboratory robotics and AI that is capable of autonomous discovery. Currently in development is the robot scientist Genesis, capable of parallel experimentation in small-volume chemostats[1]. The goal for the project is to take the number of parallel experiments into the thousands. Automating biological research on this scale requires a robot scientist to have capability across different domains, including robotics, multiomic data analyses (see Chapter 3), laboratory experimentation and hypothesis formation. In order for machines to execute these tasks accurately, and to participate in collaborative science with human scientists, a controlled vocabulary is needed. Alongside the definition of ontology terms we need a database system that can handle both the volume and complexity of the data, both for storage and also for querying by an AI system for the design of experiments.

Approach and Contributions

For this project, we analysed the domains relevant for Genesis and identified three requirements for our database system: machine-interpretable data and metadata storage compatible with automated reasoning; easily deployable storage; and consistency into the future to ensure reproducibility of experiments and results. Using these requirements, we designed an ontology and a scalable database system (Genesis-DB) based on the open source semantic web framework Apache Jena. We tested Genesis-DB using our domain ontology on an example experiment design search for *S. cerevisiae.* In addition to these main contributions, we documented the project and published the code so it can be used by different domains, and other robot scientist projects, by supplying a relevant ontology.

Author contributions

The conceptualisation of the project was done by Ross D. King, Larisa N. Soldatova, Gabriel K. Reder, **Alexander H. Gower**, and Filip Kronström. The implementation of the database system was done by Vinay Mahamuni, Rushikesh Halle, Amit Patel, and Harshal Hayatnagarkar. The ontology was written by: R.H., V.M., F.K., and **A.H.G.** F.K., **A.H.G.**, G.K.R., V.M., and R.H. conducted the investigation into the database and ontology. The manuscript was written by **A.H.G.**, F.K., R.H., V.M., G.K.R., H.H., and

[1] **chemostat**—a liquid cultivation technique where the organisms (yeast) are held in suspension in a growth medium, with an input tube pumping fresh growth media in at a controlled rate, and another tube removing the culture mixture at the same rate. Thus the total volume remains the same, and by supplying a medium with a growth-limiting substrate, e.g. glucose, the culture can be held in a physiological steady state with a constant biomass specific growth rate equal to the rate of dilution (rate of input and output flow).

L.N.S. The figures and visualisations were realised by F.K., **A.H.G.**, R.H., V.M., and G.K.R. The project was supervised by R.D.K., L.N.S., G.K.R., A.P., and H.H. The data were prepared and curated by **A.H.G.**, F.K., G.K.R., V.M., and R.H. The project was administered by L.N.S., R.D.K., G.K.R., A.P. The funding for the project was acquired by R.D.K.

4.3 Paper III: RIMBO - An Ontology for Model Revision Databases

Problem

Knowledge in science, and particular in systems biology, is often stored in a machine-readable file that allows for computational modelling of the system. Improving these computational models through scientific enquiry means changing the content of these files. Commonly, a copy is made of the model and changes are incorporated into this copy. However as these models grow this presents challenges.

Changes to models are often bundled together to resolve the problem of multiple large files. However this breaks the relationship between a single hypothesis and a change to the model. This makes reasoning about changes to models quite difficult, especially for large-scale computational reasoners.

Approach and Contributions

We present here an ontology, RIMBO, for describing changes to a genome-scale metabolic model (GEM) written in RDF/XML. The ontology enables individual changes to be linked semantically to the reasons for the change and the details of the change to the model. RIMBO combines classes and relations from existing ontologies with new classes and relations. We demonstrate modelling example revisions to the GEM Yeast8.

Author contributions

The conceptualisation of the project was done by Ross D. King, Filip Kronström, and **Alexander H. Gower**. The ontology was designed and curated by F.K. Code to implement RIMBO was also developed and tested by F.K. The experiments to demonstrate revisions on the Yeast8 GEM were designed and executed by **A.H.G.** and F.K. The scalability experiments were designed and executed by F.K. The data were prepared and curated by F.K. and **A.H.G.** Figures were designed and prepared by F.K. The manuscript was written by F.K. The project was supervised by R.D.K., and Ievgeniia A. Tiukova. The funding for the project was acquired by R.D.K.

4.4 Paper IV: LGEM$^+$: Automated Improvement of Metabolic Network Models and Model-Driven Experimental Design through Abduction

This manuscript is an extension of a paper published in the proceedings of the 26th International Conference on Discovery Science (Gower et al. 2023, listed as paper [**b**] under "Other Publications").

Problem

Knowledge about yeast is highly structured due to community efforts to standardise, retain and distribute it. This is covered in more detail in Chapter 3. The models that are stored in these databases are therefore improved upon incrementally. Incremental improvements to these models are generally made through careful study of a particular entity, pathway or process in the organism. Making improvements to these models is a time-consuming process. This is primarily because of the human resource required to hypothesise and design experiments, as well as analyse results. The interface between different activities in this scientific method are sources of friction and delay.

With this paper we aimed to provide a framework for improvement of models of yeast metabolism that would reduce the time required of human scientists in the discovery process, and increase the quality of the hypotheses generated.

Approach and Contributions

The main contributions of LGEM$^+$ as presented in this paper are: (1) a compartmentalised first-order logic model of yeast metabolism; (2) a set of algorithms for the extraction and analysis of metabolic pathways from simulations; (3) a two-stage method for the abduction of novel hypotheses on improved models; (4) scalable methods for evaluating these models and hypotheses; (5) an algorithm to integrate FBA with abductive reasoning.

We use an ATP and express knowledge encoded in GEMs in first-order logic syntax, retaining the semantics of the original model (which is written using an ontology) and layering additional knowledge about general processes such as biochemical reactions, enzyme formation and catalysis.

We tested this framework using several community models on two separate tasks. Firstly, a genome-scale single-gene deletant viability screen. To evaluate this we used experimental data from previously published research to obtain a confusion matrix for the predictive task. Secondly we used the ATP to generate reaction pathway configurations for a defined growth medium. We compared these extracted pathways with the literature data, and also with simulations from a different deductive approach, FBA, but using the same community model (GEM) as a background knowledge source. The overall flow of this abductive process for improvement, including the initial construction of the

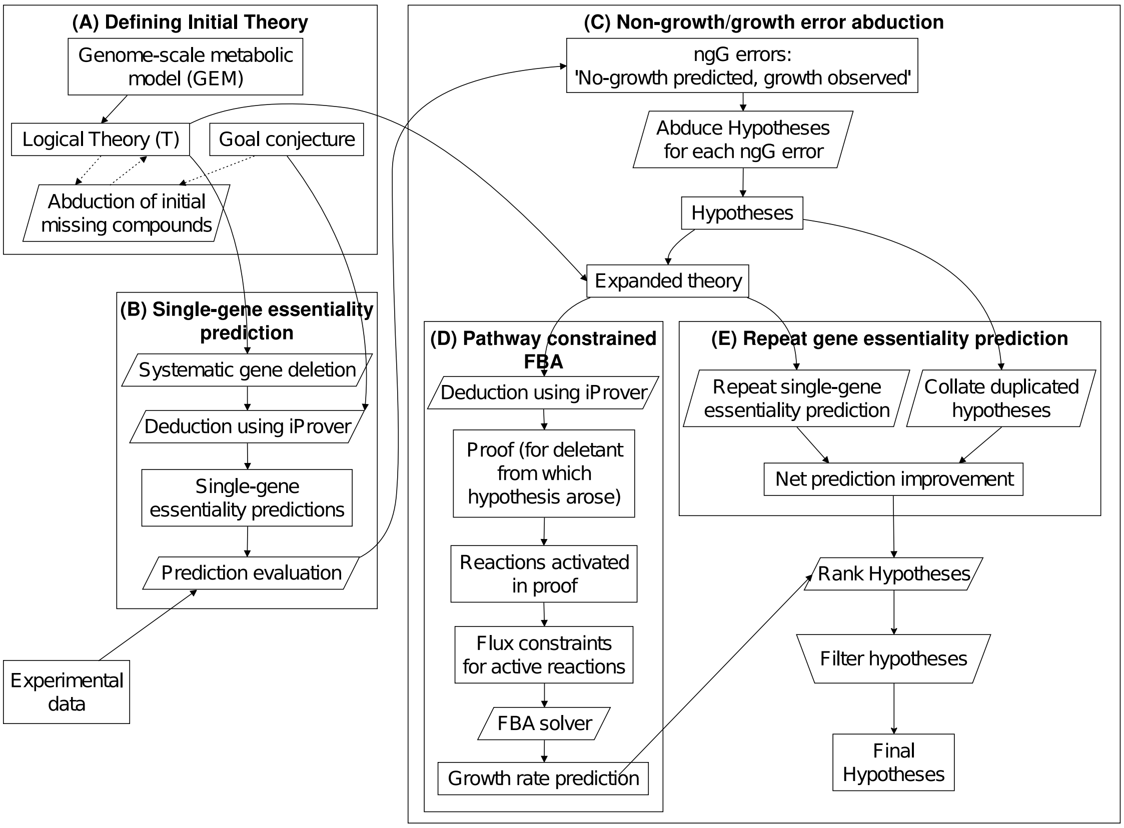
(A) Defining Initial Theory
Genome-scale metabolic model (GEM)
Logical Theory (T)
Goal conjecture
Abduction of initial missing compounds
(B) Single-gene essentiality prediction
Systematic gene deletion
Deduction using iProver
Single-gene essentiality predictions
Prediction evaluation
Experimental data
(C) Non-growth/growth error abduction
ngG errors: 'No-growth predicted, growth observed'
Abduce Hypotheses for each ngG error
Hypotheses
Expanded theory
(D) Pathway constrained FBA
Deduction using iProver
Proof (for deletant from which hypothesis arose)
Reactions activated in proof
Flux constraints for active reactions
FBA solver
Growth rate prediction
(E) Repeat gene essentiality prediction
Repeat single-gene essentiality prediction
Collate duplicated hypotheses
Net prediction improvement
Rank Hypotheses
Filter hypotheses
Final Hypotheses

Figure 4.1: Processes in LGEM$^+$. **(A)** defining the logical theory, including abduction of missing compounds to enable viability of base strain; **(B)** single-gene essentiality prediction; **(C)** abduction of hypotheses from *ngG* errors; **(D)** using FBA to assess viability of each hypothesis; and **(E)** repeating single-gene deletion to assess viability of each hypothesis.

files, is shown in Fig. 4.1. We repeated the above experiments for different defined media. We present a model-driven experimental design strategy, and demonstrate this with a differential expression study, and using the Δpfk2 mutant strain as a case study.

The programs used for these experiments were implemented primarily in Bash, Python and Perl, and of course using the command line interface (CLI) for the ATP, iProver. We containerised the project to improve replicability, and also to enable users to reuse the base functions for their own purposes. This will also help us to scale this software and integrate it with our robot scientist, Genesis.

Author contributions

The conceptualisation of the project was done by Ross D. King, Konstantin Korovin, and **Alexander H. Gower**. The logical predicate and clause structure were designed by K.K. and **A.H.G.** Code to generate logical theory structures from GEMs was developed and tested by **A.H.G.** Extensions to the ATP iProver to incorporate abduction were developed by K.K. The experiments were designed by R.D.K, Ievgeniia A. Tiukova, K.K., and **A.H.G.**, and executed by **A.H.G.** The microarray expression data analysis was conducted by Erik Y. Bjurström, Praphapan Lasin, and **A.H.G.** The data were prepared and curated by **A.H.G.** Figures were designed and prepared by **A.H.G.** The manuscript was written by **A.H.G.** and K.K, and edited by E.Y.B, Daniel Brunnsåker, and R.D.K. The project was supervised by R.D.K., K.K., and I.A.T. The funding for the project was acquired by R.D.K.

Chapter 5

Discussion and Future Work

In this thesis, we introduce tools designed to automate parts of the scientific discovery process in systems biology.

Paper I serves as an introduction to the core concepts of this thesis: the scientific method, automation thereof and robot scientists, and systems biology.

In **Paper II**, we present a database system (Genesis-DB) for the curation of experimental data from the robot scientist Genesis, the associated experiment designs, and data about the execution of experiments. In keeping with the requirements for automation, this system is scalable and is semantically rich so can be exploited by automated reasoners. This is essential for implementing active learning algorithms.

Paper III presents an ontology (RIMBO) for formalising hypotheses on genome-scale metabolic models (GEMs). We demonstrate the use of RIMBO for active learning using an example of selecting a valuable point in the experimental space. Similarly to Genesis-DB, RIMBO enables automated reasoners to exploit the content and context of hypotheses in systems biology, which in turn allows for scalable model comparison by a robot scientist.

In **Paper IV**, we present a first-order logic (FOL) model of *Saccharomyces cerevisiae*, and methods for the generation and evaluation of hypotheses for the improvement of GEMs. We show that an automated theorem prover (ATP) can be used for domain-specific reasoning tasks by remaining faithful to the semantics of domain ontologies when constructing the logical theories.

Overall the contributions of this thesis show that the reasoning tasks that are necessary for the formation of hypotheses, and the evaluation of scientific models, can be achieved with general-purpose tools when appropriate ontologies are employed, and that these approaches scale well. What remains to be researched is how well these tools integrate with the experimental platform and the wider knowledge base in systems biology, aiming for closed-loop automation in systems biology.

With this in mind, some potential future directions for this research are as listed on the following page.

- **Extending the first-order logic vocabulary to improve the power of LGEM$^+$.** We currently focus on biochemical pathways. There are effects from gene regulation and signalling that are important to predictions of phenotype, so omission of these limits the effectiveness of the active learning approach. We could, for example, include additional predicates and clause structures to provide more detail regarding enzyme availability, or gene regulation and signalling processes.

- **Align first-order logic model more closely with other domain ontologies,** for example RIMBO, or systems biology and chemistry ontologies. These will provide a broader range of knolwedge sources for abduction algorithms, and also decreased friction for automated reasoning techniques and model improvement.

- **Abstracting elements of LGEM$^+$ and RIMBO away from the domain of systems biology.** As explored in **Paper I**, there are aspects of the scientific method that are not domain specific. For example elements of reasoning about hypotheses and theories, or the confidence held in a hypothesis. Assessing how much of the algorithms and systems we develop for systems biology can be generalised to other scientific domains is an important research topic, both from an engineering perspective and a philosophical one.

- **Integrating LGEM$^+$ with other quantitative modelling techniques.** In **Paper IV**, we implemented one way to integrate logical reasoning over GEMs with flux balance analysis (FBA). As mentioned in Chapter 3, there are numerous other quantitative modelling paradigms used in systems biology, and these models can provide another source of information for the abduction tasks taken on by a robot scientist.

Bibliography

Baltimore, David (June 1970). 'Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses'. In: *Nature* 226.5252, pp. 1209–1211. DOI: 10.1038/2261209a0 (cit. on p. 10).

Ben-Ari, Mordechai (2012). *Mathematical Logic for Computer Science*. London: Springer. DOI: 10.1007/978-1-4471-4129-7 (cit. on pp. 6, 7).

Carthew, Richard W. (Apr. 2021). 'Gene Regulation and Cellular Metabolism: An Essential Partnership'. In: *Trends in genetics : TIG* 37.4, pp. 389–400. DOI: 10.1016/j.tig.2020.09.018 (cit. on p. 13).

Castiglione, Filippo et al. (2014). 'Modeling Biology Spanning Different Scales: An Open Challenge'. In: *BioMed Research International* 2014, p. 902545. DOI: 10.1155/2014/902545 (cit. on p. 12).

Crick, Francis (1958). *On Protein Synthesis. Symposium of the Society for Experimental Biology XII* (cit. on p. 9).

— (Aug. 1970). 'Central Dogma of Molecular Biology'. In: *Nature* 227.5258, pp. 561–563. DOI: 10.1038/227561a0 (cit. on p. 9).

Cruz, Daniel A. and Melissa L. Kemp (Oct. 2021). 'Hybrid Computational Modeling Methods for Systems Biology'. In: *Progress in Biomedical Engineering* 4.1, p. 012002. DOI: 10.1088/2516-1091/ac2cdf (cit. on p. 14).

García Sánchez, Carlos Eduardo and Rodrigo Gonzalo Torres Sáez (2014). 'Comparison and Analysis of Objective Functions in Flux Balance Analysis'. In: *Biotechnology Progress* 30.5, pp. 985–991. DOI: 10.1002/btpr.1949 (cit. on p. 13).

Georgiou, Pamina et al. (2022). 'The Rapid Software Verification Framework.' In: *FMCAD*, pp. 255–260 (cit. on p. 8).

Goel, Shilpi and Sandip Ray (2022). 'Microprocessor Assurance and the Role of Theorem Proving'. In: *Handbook of Computer Architecture*. Ed. by Anupam Chattopadhyay. Singapore: Springer Nature, pp. 1–43. DOI: 10.1007/978-981-15-6401-7_38-1 (cit. on p. 8).

Goffeau, A. et al. (Oct. 1996). 'Life with 6000 Genes'. In: *Science* 274.5287, pp. 546–567. DOI: 10.1126/science.274.5287.546 (cit. on p. 12).

Gower, Alexander H. et al. (2023). 'LGEM$^+$: A First-Order Logic Framework for Automated Improvement of Metabolic Network Models Through Abduction'. In: *Discovery Science*. Ed. by Albert Bifet et al. Vol. 14276. Cham: Springer Nature Switzerland, pp. 628–643. DOI: 10.1007/978-3-031-45275-8_42 (cit. on p. 20).

Greer, Eric L. and Anne Brunet (Feb. 2008). 'Signaling Networks in Aging'. In: *Journal of Cell Science* 121.4, pp. 407–412. DOI: 10.1242/jcs.021519 (cit. on p. 11).

Khasidashvili, Zurab, Konstantin Korovin and Dmitry Tsarkov (2015). 'EPR-based k-Induction with Counterexample Guided Abstraction Refinement'. In: *Global Conference on Artificial Intelligence, GCAI 2015, Tbilisi, Georgia, October 16-19, 2015*. Ed. by Georg Gottlob, Geoff Sutcliffe and Andrei Voronkov. Vol. 36. EPiC Series in Computing. EasyChair, pp. 137–150. DOI: 10.29007/SCV7 (cit. on p. 8).

King, Ross D. et al. (2009). 'The Automation of Science'. In: *Science* 324.5923. DOI: 10.1126/science.1165620 (cit. on pp. 5, 15).

Kohl, P et al. (2010). 'Systems Biology: An Approach'. In: *Clinical Pharmacology & Therapeutics* 88.1, pp. 25–33. DOI: 10.1038/clpt.2010.92 (cit. on p. 9).

Korovin, Konstantin (2008). 'iProver – An Instantiation-Based Theorem Prover for First-Order Logic (System Description)'. In: *Automated Reasoning*. Ed. by Alessandro Armando, Peter Baumgartner and Gilles Dowek. Vol. 5195. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 292–298. DOI: 10.1007/978-3-540-71070-7_24 (cit. on p. 8).

Krebs, Edwin G and Jonathan D Graves (June 2000). 'Interactions between Protein Kinases and Proteases in Cellular Signaling and Regulation'. In: *Advances in Enzyme Regulation* 40.1, pp. 441–470. DOI: 10.1016/S0065-2571(99)00030-8 (cit. on p. 10).

Mahadevan, R. and C.H. Schilling (Oct. 2003). 'The Effects of Alternate Optimal Solutions in Constraint-Based Genome-Scale Metabolic Models'. In: *Metabolic Engineering* 5.4, pp. 264–276. DOI: 10.1016/j.ymben.2003.09.002 (cit. on p. 13).

Mahadevan, Radhakrishnan, Jeremy S. Edwards and Francis J. Doyle (Sept. 2002). 'Dynamic Flux Balance Analysis of Diauxic Growth in Escherichia Coli'. In: *Biophysical Journal* 83.3, pp. 1331–1340. DOI: 10.1016/S0006-3495(02)73903-9 (cit. on p. 13).

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 (cit. on p. 13).

Österberg, Linnea et al. (Sept. 2020). 'A Novel Yeast Hybrid Modeling Framework Integrating Boolean and Enzyme-Constrained Networks Enables Exploration of the Interplay between Signaling and Metabolism'. In: *bioRxiv*, p. 2020.09.11.290817. DOI: 10.1101/2020.09.11.290817 (cit. on p. 14).

Resat, Haluk, Linda Petzold and Michel F. Pettigrew (2009). 'Kinetic Modeling of Biological Systems'. In: *Computational Systems Biology*. Ed. by Reneé Ireton et al. Totowa, NJ: Humana Press, pp. 311–335. DOI: 10.1007/978-1-59745-243-4_14 (cit. on p. 13).

Sánchez, Benjamín J et al. (Aug. 2017). 'Improving the Phenotype Predictions of a Yeast Genome-Scale Metabolic Model by Incorporating Enzymatic Constraints'. In: *Molecular Systems Biology* 13.8, p. 935. DOI: 10.15252/msb.20167411 (cit. on p. 13).

Schindler, Samuel (July 2022). ‘Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think?’ In: *Philosophy of Science* 89.3, pp. 542–564. DOI: `10.1017/psa.2021.40` (cit. on p. 5).

Segrè, Daniel, Dennis Vitkup and George M. Church (Nov. 2002). ‘Analysis of Optimality in Natural and Perturbed Metabolic Networks’. In: *Proceedings of the National Academy of Sciences* 99.23, pp. 15112–15117. DOI: `10.1073/pnas.232349399` (cit. on p. 13).

Shlomi, Tomer, Omer Berkman and Eytan Ruppin (May 2005). ‘Regulatory on/off Minimization of Metabolic Flux Changes after Genetic Perturbations’. In: *Proceedings of the National Academy of Sciences* 102.21, pp. 7695–7700. DOI: `10.1073/pnas.0406346102` (cit. on p. 13).

Southern, James et al. (2008). ‘Multi-Scale Computational Modelling in Biology and Physiology’. In: *Progress in Biophysics and Molecular Biology* 96.1, pp. 60–89. DOI: `10.1016/j.pbiomolbio.2007.07.019` (cit. on p. 12).

Temin, Howard M. and Satoshi Mizutani (June 1970). ‘Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus’. In: *Nature* 226.5252, pp. 1211–1213. DOI: `10.1038/2261211a0` (cit. on p. 10).

Urban, Josef and Jiří Vyskočil (2013). ‘Theorem Proving in Large Formal Mathematics as an Emerging AI Field’. In: *Automated Reasoning and Mathematics: Essays in Memory of William W. McCune*. Ed. by Maria Paola Bonacina and Mark E. Stickel. Berlin, Heidelberg: Springer, pp. 240–257. DOI: `10.1007/978-3-642-36675-8_13` (cit. on p. 8).

Weisberg, Ellen et al. (Feb. 2005). ‘Characterization of AMN107, a Selective Inhibitor of Native and Mutant Bcr-Abl’. In: *Cancer Cell* 7.2, pp. 129–141. DOI: `10.1016/j.ccr.2005.01.007` (cit. on p. 11).

Williams, Kevin et al. (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` (cit. on p. 15).