Thesis for The Degree of Licentiate of Engineering

Knowledge Representations for Scientific Discovery

Integrating Ontologies and Neurosymbolic models for Yeast Systems Biology Research

Filip Kronström

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

Knowledge Representations for Scientific Discovery
Integrating Ontologies and Neurosymbolic models for Yeast Systems Biology Research

Filip Kronström

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 2025.

"Science is magic that works."
- Kurt Vonnegut

Knowledge Representations for Scientific Discovery

Integrating Ontologies and Neurosymbolic models for Yeast Systems Biology Research

Filip Kronström

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

Abstract

Scientific discovery has a lot to gain from advances in artificial intelligence and machine learning, and where progress can have major societal impact. These techniques can help scientists uncover patterns in large datasets, generate new hypotheses, and guide experimental design. Combined with robotic systems they can also greatly increase the number of performed experiments. This thesis investigates how structured knowledge representations can support scientific discovery in systems biology, with a particular focus on the model organism *Saccharomyces cerevisiae* (baker's yeast).

The work introduces two ontologies designed as semantic schemas for research databases. The first captures metadata and results from μ-chemostat experiments, accommodating multiple measurement modalities. The second ontology formalises revisions to computational models, with a focus on domains where mechanistic models are updated iteratively and it is important to record what was changed and why.

In the third contribution, information about *S. cerevisiae* from public databases is integrated into a knowledge graph with well-defined class hierarchies. Graph neural networks, in combination with box embeddings representing the hierarchical structure, are used to predict growth outcomes of double gene deletions. Furthermore, explainability techniques are applied to identify candidate biological interactions, forming hypotheses about traits in *S. cerevisiae*. One such hypothesis is experimentally validated, illustrating how structured representations can aid data-driven discovery from publicly available resources.

Taken together, this work introduces knowledge representations for emerging domains, designed as tools to support scientific discovery, while also demonstrating how rich, structured representations can enhance the interpretation of existing data.

Keywords

Knowledge Represenations, Ontologies, Knowledge Graphs, Neurosymbolic AI, Graph Neural Networks, Semantic Web, Systems Biology

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] 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.

[**Paper II**] **F. Kronström**, A. H. Gower, I. A. Tiukova, R. D. King, *RIMBO - An Ontology for Model Revision Databases.* Lecture Notes in Computer Science, Volume 14276, 26th International Conference on Discovery Science, Porto, October 2023.

[**Paper III**] **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, 19th International Conference on Neurosymbolic Learning and Reasoning, Santa Cruz, September 2025.

[†] Equal contribution.

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

[**b**] 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 metabolomic fingerprinting.* Journal of the American Society for Mass Spectrometry, Volume 35, Issue 3, February 2024.

[**c**] A. H. Hower, 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 review.

[**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.

[**e**] 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.* Under review.

Acknowledgment

I want to begin by thanking my supervisor, Ross, for giving me this opportunity. I have also learnt so much from working and discussing with you, both about conducting research, science in general, life in academia, but also topics which, at first glimpse (and sometimes if you keep looking), might seem completely unrelated. Thank you also to my co-supervisor, Ievgeniia, for your support, for introducing lots of biological terms and concepts to me, and for sometimes pushing me out of my comfort zone. I also want to thank my examiner Nir, I have always felt encouraged after talking to you.

Next, I want to thank everyone in the King Group at Chalmers. Thank you Gabe for welcoming me when I started, for giving me lots of advice about staying sane as a PhD student, and for all the good times we had during your time here. Thank you also Beera and Prajakta, it has been a pleasure sharing office with you. My fellow PhD students, Alec, Daniel, and Erik, without your help and discussions, none of this would probably have been completed. However, more importantly, I want to thank you for being such good friends!

Thank you to everyone at the Data Science and AI division, especially the PhD students, unfortunately I can't name you all. Ever since we moved here I have felt very welcome and I have made many new friends.

I also want to thank Agnieszka Ławrynowicz and Filip Cornell for fruitful discussions about my research.

Outside of academia I want to thank all of my friends. Whether through pub quizes, book clubs, concerts, travels, or anything else, it has been fantastic to have you help me think if other things than research. To my family, Mum and Dad, Hanna and Alice, thank you for all your support, you are my biggest inspirations. Lastly, Anna, you have given me new perspectives about everything and you make me smile every day.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, as well as the Chalmers AI Research Centre. Computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

Part I

Summary

Chapter 1

Introduction

The scientific process has been of immense importance for our understanding of the world and the technological advancements which shape the way we live. Science involves the formulation of hypotheses, the design and execution of experiments, the analysis of results, and the iterative refinement of theories. This cycle, while powerful, is limited by human cognitive and perceptual capacities, as well as by the time and resources available to researchers. As scientific disciplines grow increasingly data-rich and complex, interest is rising in computational systems that can augment or automate elements of scientific reasoning and discovery.

This thesis focuses on the scientific domain of yeast (*Saccharomyces cerevisiae*) systems biology. The field aims to understand and predict the behaviour of *S. cerevisiae* as an integrated biological system, both to inform our understanding of higher eukaryotes, such as humans, [1] and to support its many applications in biotechnology, including biofuel production [2].

Within this context, the thesis explores the role knowledge representations play for automated scientific discovery in systems biology. Since the early days of artificial intelligence how to represent knowledge has been a central problem. According to Davis et al. (1993),

> a knowledge representation is most fundamentally a *surrogate*, a substitute for the thing itself, that is used to enable an entity to determine consequences by thinking rather than acting, that is, by reasoning about the world rather than taking action in it [3].

In other words, knowledge representations provide a structured way of encoding information so that machines can use and reason about it. They can be expressed in formal languages, based on logic. These symbolic representations make facts verifiable and can provide a common grounding for human users and computer systems, providing interpretability to the system.

Representations can also be learned from data, for example in the form of trained model weights or dense vector representations. These sub-symbolic or neural representations have proved successful for tasks such as computer vision and natural language processing where vast amounts of data are available.

However, providing guarantees or explanations regarding their behaviour is generally difficult. The field of neurosymbolic AI tries to combine neural and symbolic techniques in various ways. This thesis contributes to the branch of neurosymbolic AI covering the integration of symbolic knowledge representations with neural vector representations and models.

The papers appended to this thesis highlight the important role of formal knowledge representations in enabling automated scientific discovery. Fig. 1.1 shows which parts of the scientific cycle in systems biology are covered in each paper. **Papers I** and **II** present ontologies designed as semantic schemas for describing μ-chemostat (microscale continuous cell cultivation systems) experiments and revisions to computational models, respectively. These representations provide structured, machine-readable descriptions that ground experimental and modelling data in shared conceptual frameworks. The semantic grounding supports both human understanding and automated reasoning, facilitating reproducibility, integration, and interpretation of data across research contexts. This grounding proves especially valuable for data sharing and reuse, a theme that culminates in **Paper III**. Here, data from public biological databases is integrated in a knowledge graph. This representation of knowledge allows for modelling and prediction of phenotypic outcomes of gene knockouts in *S. cerevisiae*. The constructed model is used to generate a testable hypothesis, which is tested and validated in a wet-lab experiment. This work demonstrates how neurosymbolic methods can integrate both symbolic structures and quantitative measurements, as well as guide discoveries. Together, the three papers demonstrate how knowledge representations are essential tools for enabling automated scientific discovery, promoting the sharing of findings, and supporting the interpretation of data.

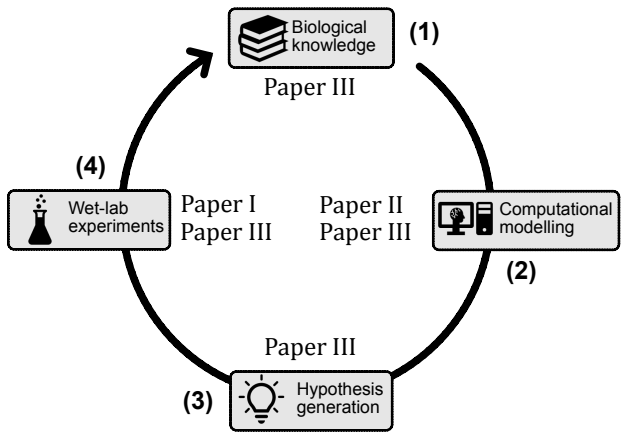


Figure 1.1: An illustration of the scientific cycle in systems biology. Available knowledge (1) is used to construct computational models (2). The models inform the generation of hypotheses about new knowledge (3). These hypotheses are tested experimentally (4). The resulting information is integrated with existing knowledge. The cycle is adopted from [4]. Indicated is which parts of the cycle are covered in each of the appended papers.

This thesis has the following structure. Chapter 2 gives a brief background to relevant topics, Chapter 3 gives a summary of the appended papers, and in Chapter 4 the findings in these papers and future research directions are discussed. The three papers are also appended to this thesis in Part II.

Chapter 2

Background

2.1 Automated Scientific Discovery

Automated scientific discovery refers to the use of algorithms and artificial intelligence (AI) to perform tasks traditionally associated with human scientists, including hypothesis generation, experimental design, and knowledge synthesis. Early work in this area, such as DENDRAL [5], analysed mass spectrometry data, while the BACON system [6] demonstrated that computers could rediscover physical laws. More recently, the most successful application of AI in natural sciences has likely been the protein structure predictions by AlphaFold [7], awarded the 2024 Nobel Prize in Chemistry. However, a central aspect to scientific discovery not addressed in the aforementioned systems is the generation of new experimental data from the physical world.

Robot Scientists

A robot scientist is a physically implemented AI system, combining computational and experimental capabilities. Through this they are capable of performing automated cycles of experiments, by iteratively generating hypotheses, designing and executing experiments, and analysing the results [8]. The first such system, Adam, studied functional genomics in *S. cerevisiae* [9]. The next iteration, Eve, was designed for early-stage drug development [10], and is the laboratory automation hardware used to perform the experiments in **Paper III**. Similar systems have also been applied in other domains, such as a mobile robotic chemist searching for photocatalysts for hydrogen production [11].

One of the main advantages of robot scientists is their ability to dramatically scale up the number of experiments and hypotheses that can be tested. Traditional experimental approaches are often constrained by human time and labour, whereas robot scientists can operate continuously, perform experiments with high precision, and systematically explore large experimental spaces. This scalability is a key focus of the next-generation robot scientist developed in our lab, Genesis, intended for yeast systems biology. Genesis integrates a large

number of μ-chemostats[1], enabling the parallel execution of many experiments while measuring the metabolome through mass spectrometry, the transcriptome via RNA sequencing, and organism growth by monitoring optical density [12]. **Paper I** presents an ontology for a graph database designed to store both the conditions and results of such μ-chemostat experiments.

One of the intended use cases for Genesis is closed-loop model improvement in systems biology, where computational models are used to guide experiment selection and the experimental results are used to improve the models, without (or with minimal) human intervention. Improving systems biology models has been described as a 'Grand Challenge' in science [13] and robot scientists performing closed-loop model improvements have previously proved to be a promising approach [14]. **Paper II** presents an ontology describing revisions to computational models, allowing large numbers of iterations to be saved.

The scientific community has increasingly acknowledged a reproducibility crisis, both regarding experiments in natural sciences [15], and computational sciences [16]. Robot scientists record experimental procedures, parameters, and results in a precise and structured manner, for example using methods suggested in **Paper I**. This reduces ambiguity and facilitates accurate replication of experiments. While scientific communication through natural language remains essential for human understanding, it is often too imprecise and incomplete to fully capture complex experimental workflows. By contrast, formal, machine-readable representations, such as structured databases, ontologies, and executable protocols, enable findings to be recorded and communicated in a way that supports both reproducibility and automation [8]. As a result, robot scientists not only accelerate the pace of discovery but can also contribute to making scientific knowledge more reliable and systematically accessible. Similarly, structured recording of computational model development, as is proposed in **Paper II**, has the potential to increase reproducibility and foster community trust by improving transparency in the development process.

2.2 Systems Biology

Systems biology is an interdisciplinary field that aims to understand how interactions among genes, proteins, metabolites, and other cellular components give rise to the emergent behaviour of biological systems. [17]. In contrast to reductionist approaches that examine isolated parts of a system, systems biology emphasises a holistic perspective, integrating data from multiple levels of biological organisation to build comprehensive models of living systems [18].

One of the central goals of systems biology is to connect genotype (the genetic makeup of an organism) to phenotype (the observable characteristics of an organism). This idea is rooted in Crick's "Central Dogma" of molecular biology, which states that genetic information flows from DNA to RNA to proteins, which in turn are responsible for most cellular functions [19]. This flow

[1] A chemostat is a bioreactor with a constant flow of growth media, allowing to carefully control the conditions under which an organism grows. The prefix μ indicates the small volume of these cultivation chambers.

of information is illustrated in Fig. 2.1. Modern biology recognises that cellular processes are highly complex, and that information flow is often neither strictly unidirectional nor isolated. These processes are influenced by intricate feedback loops, post-translational modifications, and additional molecular actors such as metabolites and signalling molecules [20].

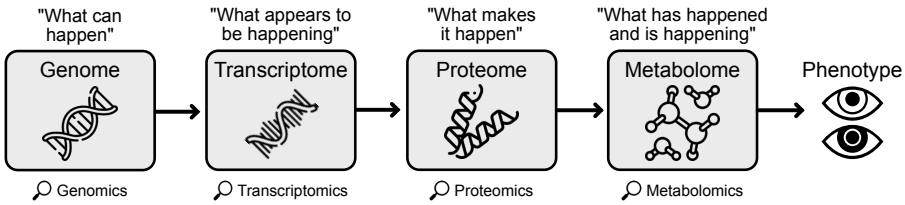


Figure 2.1: The "Omics-cascade" depicts the flow of information across different biological layers in a cell. Genetic variation influences RNA expression (transcriptomics), which affects protein abundance and activity (proteomics), and subsequently alters metabolite levels (metabolomics). These molecular changes collectively shape observable cellular traits and behaviors (phenotypes). The cascade highlights how data from multiple omics layers can be integrated to understand complex biological systems. Adopted from [21].

To investigate these complexities, systems biology integrates experimental and computational approaches. Large-scale biological datasets are generated using high-throughput technologies such as genomics (study of DNA), transcriptomics (RNA expression), proteomics (proteins), and metabolomics (small-molecules). These datasets capture complementary layers of cellular information, collectively referred to as the "Omics-cascade", which again can be seen in Fig. 2.1. By linking information across these levels, the "Omics-cascade" provides a framework for understanding how changes at the genetic level propagate to proteins, metabolites, and ultimately phenotypes. These data are used to construct and refine computational models that aim to both describe and predict system behaviour under various perturbations, such as gene knockouts or drug treatments, which can help for example personalised healthcare or drug-development [22].

Among the models used in systems biology, genome-scale metabolic models (GEMs) are particularly important for linking different omics data to metabolism. They provide a holistic representation of an organism's metabolism as a network of biochemical reactions. Genes are linked to the enzymes and reactions they encode through gene-protein-reaction (GPR) rules: genes code for enzymes (proteins), which in turn catalyse particular reactions. In the model, these relationships are often expressed using Boolean logic, for example indicating whether alternative gene products can catalyse the same reaction (an "OR" relationship), or whether multiple gene products must act together as subunits of a protein complex (an "AND" relationship). The reactions themselves are organised into a stoichiometric network, where each reaction is described by a balanced equation that specifies the number of molecules of each metabolite consumed and produced. This structure ensures that mass

and energy are conserved and allows the full metabolic capacity of the cell to be represented as a single, interconnected system [23]. Such a network can be analysed using computational methods like flux balance analysis (FBA), which applies linear optimisation under the assumption of steady state to predict the distribution of fluxes through the network under given conditions [24]. **Paper II** demonstrates the proposed data model on revisions of such a GEM.

Over the years, increasingly refined GEMs of *S. cerevisiae* have been developed, with Yeast9 representing the most recent consensus model, integrating updates in reaction coverage, gene-reaction associations, and extensive experimental validation [25].

Important for the developments in systems biology, and many other fields, has been the sharing of scientific discoveries and data. In addition to sharing findings in literature, there are several databases containing curated information. Reactome [26], KEGG [27], and BioCyc [28] are examples of databases aggregating information about reactions and pathways in different organisms. MetaboLights [29] and PRIDE [30] are databases containing mass-spectrometry based metabolomics and proteomics data. There are also databases which aggregate gene annotations from publications about various organisms, for example Saccharomyces Genome Database (SGD) [31] for yeast (*S. cerevisiae*). Information from such public databases provides curated data which can be used by machine learning systems in these domains. Annotations of *S. cerevisiae* genes from SGD have for example been used to predict protein abundances [32]. In **Paper III** data from BioCyc and SGD are combined in a knowledge graph.

Systems biology has played an important role both for our understanding of biology and for industrial applications [33]. Robot scientists show great promise in further advancing the field as they can both increase experiment throughput, while also providing superhuman capabilities when analysing experimental results [34]. An example of this is how Coutant et al. performed closed-loop cycles of experiments on an autonomous lab robotic system to improve a GEM of yeast [14].

The model organism *Saccharomyces cerevisiae*

When the direct study of the organism of interest is impractical, for example due to ethical, financial, or logistical reasons, a model organism can be used instead. A model organism shares key biological characteristics with the target species, making it a suitable proxy for investigation while being more amenable to experimental manipulation. Common examples include the fruit fly, mouse, and yeast [35].

Baker's yeast (*S. cerevisiae*) is a single-celled eukaryote which has been used for food production and beverage fermentation by humans for thousands of years [36]. In addition, *S. cerevisiae* is interesting for the scientific community due to its ease of cultivation, robustness, and shared fundamental biological processes with higher organisms. For example, Roger Kornberg received the 2006 Nobel Prize in Chemistry for discoveries about eukaryotic transcription, found by studying yeast [37]. Additionally, its biology is highly amenable to

genetic modification, for example through homologous recombination (a DNA exchange process) [38], and it was the first eukaryote to be fully sequenced [39]. Although years of research on *S. cerevisiae* have produced a wealth of knowledge, the organism is still far from fully understood [40]. One way to accelerate further discoveries is to make greater use of computational resources and robot scientists. To fully exploit existing knowledge in this context, it must first be represented in a form accessible to computers.

2.3 Formal Knowledge Representations

Propositional and First-order logic

Propositional logic is a formal system where statements (propositions) are either true or false and are combined using logical connectives such as conjunction ("AND", $\wedge$), disjunction ("OR", $\vee$), and negation ("NOT", $\neg$). It allows reasoning about the truth of whole statements but does not handle, meaning each statement must refer to a specific, fully specified case rather than a general or unknown element.

First-order logic (FOL) extends propositional logic by introducing quantifiers, such as for all ($\forall$) and exists ($\exists$), constants, and predicates, which can have different arity (numbers of arguments). The unary predicate Person(alice) states that the constant alice is a person and the binary predicate ParentOf(alice, bob) indicates a relationship, where the alice is the parent of bob. FOL also allows for predicates with more arguments, for example Gives(alice, catch22, bob) stating that alice gives catch22 to bob. These extensions make FOL more expressive and capable of representing complex and generalised relations between sets.

Description logic

Description logics (DLs) are a family of formal knowledge representation languages, and (in most cases) decidable fragments of first-order logic, for example only allowing unary and binary predicates, balancing expressivity and computational tractability. In DL, knowledge about a domain is represented using concepts, roles, and individuals. Concepts correspond approximately to unary predicates in first-order logic, roles to binary predicates, and individuals to constants. A DL knowledge base typically consists of two parts: the TBox and the ABox[2]. The TBox (terminological box) contains general knowledge about the domain using concepts and roles, while the ABox (assertional box) contains facts about specific individuals. In a database setting, the TBox can be viewed as the schema and the ABox corresponds to the instances [41]. A term often used interchangeably with the TBox and ABox is ontology, defined by Gruber as "an explicit specification of a conceptualisation" [42].

The following is a simple example of a knowledge base in DL:

[2]For more more expressive DLs there can also be a third part, the RBox (role box), with axioms about roles, for example defining symmetric roles or role inclusion.

$$
\left.\begin{array}{l}
\texttt{Parent} \sqsubseteq \texttt{Person} \\
\texttt{Child} \sqsubseteq \texttt{Person} \\
\texttt{Parent} \equiv \texttt{Person} \sqcap \exists \texttt{hasChild.Person} \\
\texttt{Child} \equiv \texttt{Person} \sqcap \exists \texttt{hasParent.Person}
\end{array}\right\} \text{TBox}
$$

$$
\left.\begin{array}{l}
\texttt{alice : Person} \\
\texttt{bob : Person} \\
\texttt{(alice, bob) : hasChild} \\
\texttt{(bob, alice) : hasParent}
\end{array}\right\} \text{ABox}
$$

The symbol $\sqsubseteq$ denotes concept inclusion (i.e., subclassing), while $\equiv$ and $\sqcap$ represent concept equivalence and conjunction, respectively. The existential restriction, e.g., $\exists$`hasChild.Person`, states that there exists a `Person` filling the `hasChild`-role. This example also demonstrates how we can reason about knowledge expressed in DL. The definitions of `Parent` and `Child`, together with the asserted facts about `alice` and `bob`, entails that `alice` is a `Parent` and `bob` is a `Child`. Typical reasoning tasks are checking for consistency between TBox and ABox, subsumption (determining whether one class is more specific than another), and query answering [41].

An important feature in the DL semantics is the *open world assumption* (OWA) under which, the absence of a fact in the knowledge base does not imply that the fact is false, it simply means that this information is not known yet. This contrasts with the Closed World Assumption (CWA) commonly used in databases, where any fact not present is assumed to be false [43]. The OWA aligns well with the view on knowledge in science, where unknown facts are yet to be discovered.

Basic description logics operate on a symbolic level, describing relationships between concepts. However, in many applications we want to be able to refer to concrete domains, such as numbers or strings. Considering the example above we might want to be able to record and reason about someone's age. To represent this we can introduce `age` as a concrete feature of the concrete domain non-negative numbers, and for example define `Adult` $\equiv$ `Person` $\sqcap \exists$`hasAge.`$\geq_{18}$, with `hasAge` being a data type role [44]. The possibility of recording data types, through concrete domains, allows DL based databases also for quantitative data. This is important for the ontologies developed in **Paper I** and **Paper II**, which are intended to record data about biological experiments and metadata about computational models.

The Semantic Web was launched as an extension to the World Wide Web to give its content semantic meaning and make it machine-readable. DLs and ontologies are defining concepts at the core of the semantic web [45]. Even if the original vision is not realised [46], the Semantic Web has still been important for the development of, for example, Resource Description Framework (RDF) to represent directed edge-labelled graphs, the query language SPARQL to query RDF graphs, and the Web Ontology Language (OWL). The second version of OWL, OWL 2, is the World Wide Web Consortium (W3C) standard language

for ontologies based on DL [41].

Different languages from the description logics family allow for different balance between expressivity and computational efficiency, and are denoted by combinations of letters. The DL $\mathcal{AL}$ can be seen as a base language allowing top and bottom classes, concept intersection, universal restrictions, limited existential quantification, and atomic negation. This can be extended to allow for example inverse roles ($\mathcal{I}$), role hierarchies ($\mathcal{H}$), concept union ($\mathcal{U}$), or data types ($^{(\mathcal{D})}$) [47]. The standard OWL 2 language corresponds to the expressive DL $\mathcal{SROIQ}^{(\mathcal{D})}$. $\mathcal{S}$ extends $\mathcal{AL}$ with concept negation, disjunction, existential restriction, and transitive roles. $\mathcal{ROIQ}$ allows for an RBox with for example role inclusions and reflexivity, nominals (enumerated individuals in the TBox), inverse roles, and number restrictions. To allow for efficient reasoning, OWL 2 has three profiles: EL, QL, and RL. OWL 2 EL is based on the more restricted $\mathcal{EL}^{++}$ description logic. $\mathcal{EL}^{++}$ supports conjunction, existential restriction, and the top concept as concept constructors, as well as transitive roles and nominals, while still allowing for reasoning in polynomial time [48]. As a result, several large-scale and widely used ontologies in systems biology, such as GO [49] and ChEBI [50], are fully or to a large extent compatible with the OWL 2 EL profile. The knowledge graph in **Paper III** is expressed in $\mathcal{EL}^{++}$, as are the ontologies in **Paper I** and **II**, extended with support for data types.

Ontologies

There exist a large number of ontologies describing different aspects of the biomedical domain. Here the ones most relevant for this work will be introduced.

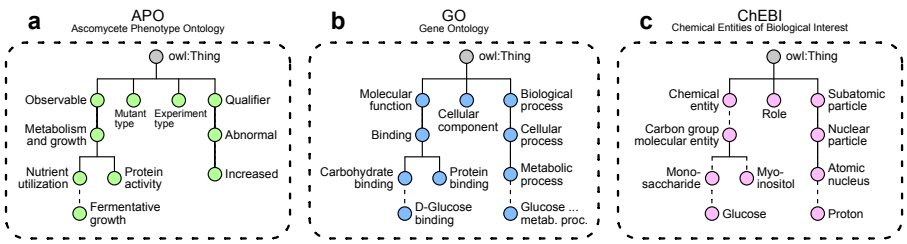


Figure 2.2: Simplified overview of the class structure in APO, GO, and ChEBI.

Ascomycete Phenotype Ontology (APO)

The Ascomycete Phenotype Ontology [51] is an ontology describing mutant phenotypes in *S. cerevisiae* and was developed for representing this information in SGD. It provides a structured vocabulary for describing `observable` characteristics such as *growth* and *stress responses*, together with `qualifiers` that specifies how these traits are affected, for example decreased, *delayed*, or *absent*. In addition, APO includes classes for `mutant type` (e.g., *null mutant*) and `experiment type` (e.g., *competitive growth*), which give context to the observed phenotype. This structured representation enables consistent annotation and comparison of phenotypes across studies. In this thesis, APO is used

in **Paper II** and **III**. The overall class structure, with selected examples, is shown in Fig. 2.2a.

Gene Ontology (GO)

The Gene Ontology [49] provides a controlled vocabulary to describe the roles of gene products across species in terms of their `molecular functions` (what a protein does, such as *ATP binding*), `biological processes` (the larger tasks it contributes to, such as *cell division*), and `cellular components` (where in the cell it acts, such as *mitochondrion*). GO enables functional annotation of genes in a standardised way, making it possible to compare results across experiments and organisms. In high-throughput transcriptomics or proteomics studies, GO terms are often used to identify groups of genes that are over-represented in a condition. For example, they can reveal that stress-related pathways are activated after heat shock, or that metabolic processes are downregulated during nutrient limitation. GO is used in **Paper II** and **III**, and an overview of parts of the ontology can be seen in Fig. 2.2b.

Chemical Entities of Biological Interest (ChEBI)

ChEBI [50] provides a controlled vocabulary for molecular entities, with a focus on small chemical compounds. It organises compounds in a class hierarchy based on their chemical structures and further describes their biological roles, such as *metabolite* or *inhibitor*, through defined relations. ChEBI enables standardised annotation of chemical information, supporting applications such as metabolic modelling, chemical annotation of biological assays, and semantic integration across databases like SGD. This consistent representation makes it possible to compare and integrate chemical knowledge across studies. In this thesis, ChEBI is used in **Paper I**, **II**, and **III**. Parts of the ontology are shown in Fig. 2.2c.

Ontology for Biomedical Investigations (OBI)

OBI [52] provides a formal framework for describing experimental investigations, including study design, protocols, and used materials. It has been used to standardise metadata and annotations in databases, for example by specifying how terms from different ontologies relate to each other. OBI is used to describe terms in **Paper I**.

COmputational MOdels DIffer (COMODI)

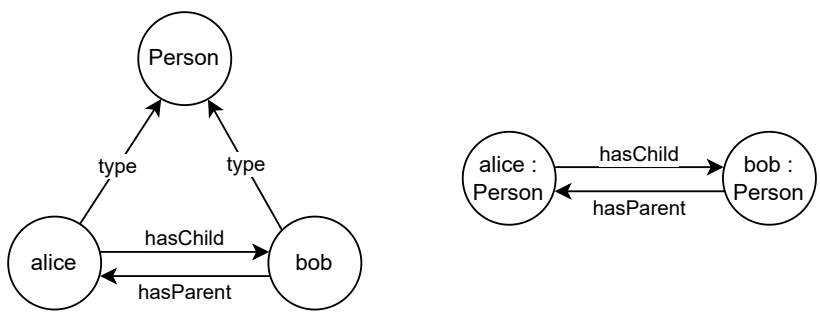
COMODI [53] describes updates to biological models expressed in XML format by annotating changes in the XML tree, along with `Changes` and `Reasons`. By doing this the traceability and reusability in computational biology can be improved. **Paper II** was inspired by, and uses terms from, COMODI.

2.4 Knowledge graphs

A knowledge graph (KG) is a flexible and semantically rich knowledge representation, defined by Hogan et al. (2021) as

> a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities [54].

KGs can be seen as graph implementations of description logic knowledge bases. One of the distinguishing features of KGs is their grounding in ontological schemas which defines the entities in the graph and allows for symbolic reasoning. The underlying graph model of a KG can vary depending on the use case and implementation. One common representation is a directed edge-labelled graph, where nodes represent entities and directed, labelled edges represent binary relations between them. The simple description logic knowledge base introduced in Section 2.3 can be expressed in this form, as illustrated in Fig. 2.3(a). This model underlies RDF, where knowledge is represented as triples of the form (*subject*, *predicate*, *object*). Another widely used model is the heterogeneous graph, which extends the directed edge-labelled graph by associating each node, in addition to the edges, with a specific type, illustrated in Fig. 2.3(b). A further extension is the property graph, which allows properties to be assigned to triples [54], and is the graph model is used in many graph databases, such as Neo4j[3].



(a) Directed edge-labelled graph (b) Heterogeneous graph

Figure 2.3: The knowledge base presented in Section 2.3 as knowledge graphs, in *a* as a directed edge-labelled graph and in *b* as a heterogeneous graph.

Learning and reasoning

Typical knowledge graph reasoning tasks can often be categorised as deductive or inductive. Deductive reasoning involves drawing logically valid conclusions from known facts or axioms, while inductive reasoning aims to infer generalisations from observed data and patterns. The example involving `alice` being

[3] `https://neo4j.com/`

a `Parent` and `bob` a `Child` in Section 2.3 illustrates an example of deductive reasoning. By contrast, had there been a link, `isSupporterOf`, between `alice` and `bob`, and the football team `bkHäcken`, an example of inductive reasoning would be inferring a rule saying that children support the same football teams as their parents,

$$\forall A \forall B \forall C(\texttt{hasChild}(A, B) \wedge \texttt{isSupporterOf}(A, C)) \rightarrow \texttt{isSupporterOf}(B, C).$$

Symbolic approaches based on for example inductive logic programming (ILP) [55] can be used for pattern mining in KGs, however special care might have to be taken to the open world assumption [56]. Connectionist or neural approaches have been applied to learn embeddings; dense, low-dimensional vector representations of KGs, that capture the graph's structural and semantic properties. These embeddings can be used for downstream tasks such as link prediction, entity classification, or as features in other machine learning models. Studies have shown that embeddings of symbolic features can be useful, especially when the knowledge used to generate the features is deemed to be insufficient for the task [57].

Knowledge graph embedding (KGE) methods often interpret individuals as points and relations as geometric operations on the involved entities. TransE is an example of this, where edges are interpreted as translations of points in embedding space [58]. Concepts are often interpreted as convex sets, e.g., spheres [59], cones [60], or boxes [61–63]. Box embeddings, where entities are represented as Cartesian products of closed intervals,

$$\text{Box} = \prod_{i=1}^{n} [z_i, Z_i], \tag{2.1}$$

have grown increasingly popular due to their trade-off between expressiveness and computational efficiency [64]. These representations enable the natural interpretation of transitive relations, such as `subClassOf` relation, by encoding hierarchical structure through set containment. Fig. 2.4 illustrates class hierarchies can be represented by box embeddings. It should be noted that box embeddings are simplified views of ontologies, and even simple examples of some $\mathcal{EL}$++ ontologies cannot be properly represented [64].

In addition to the embedding-based models mentioned, graph neural networks (GNNs) have emerged as a powerful class of models to learn representations of nodes in knowledge graphs. GNNs operate by "passing messages" between neighbouring nodes, allowing each node to aggregate information and update its features based on its local neighbourhood. By stacking multiple message-passing layers, GNNs can learn node representation that capture multi-hop relational dependencies across the graph. For example, GraphSAGE [65], used in **Paper III**, calculates node embedding $h_v^{(k)}$ for node v at depth k in the network via

$$h_v^{(k)} = f^{(k)}\big(W^{(k)} \underset{u \in \mathcal{N}(v)}{\text{AGG}} [h_u^{(k-1)}] + B^{(k)} h_v^{(k-1)}\big), \tag{2.2}$$

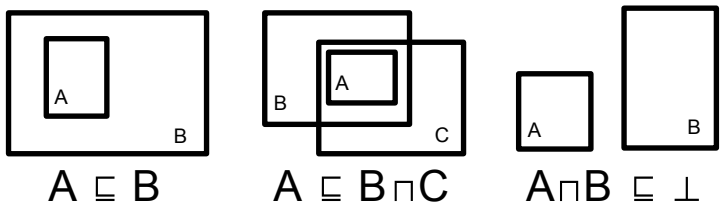


Figure 2.4: Illustration of how different class constructions are represented as boxes in box embeddings. A $\sqsubseteq$ B is saying that A is a subclass of B, A $\sqsubseteq$ B $\sqcap$ C means that A is a subclass of the intersection of B and C, and A $\sqcap$ B $\sqsubseteq \bot$ states that A and B are disjoint.

where f is a learnable and possibly non-linear function, W and B are learnable weights, and AGG is an aggregation function applied to the neighbours of v. The aggregation function can for example be fixed (e.g. mean) or learnt.

In **Paper III** box embeddings are used as node features, as a way of encoding the hierarchical class structure, used together with a GNN to make node predictions from a KG.

Chapter 3

Summary of Included Papers

3.1 Paper I: Genesis-DB: a database for autonomous laboratory systems

In this paper we present an ontology for data and metadata from small-volume chemostat cultivation experiments, as well as an Apache Jena based RDF-store.

Problem

As described in Section 2.1 a robot scientist platform, Genesis, with small-volume chemostats and multiple measurement modalities, able of performing a large number of parallel experiments, is under development. One central aspect for autonomous agents using such experimental hardware is a data storage system capable of handling the amounts of data generated, as well as representing it such that machines can reason about it.

Approach and Contributions

We develop an ontology for an RDF database describing the μ-chemostat cultivation conditions, as well as the measurement modalities; growth, metabolomics, and transcriptomics. The ontology follows the OBI [52] structure and reuses terms from OBO Foundry [66] ontologies where possible, to simplify integration with other knowledge bases.

An experiment consists of a sequence of `cell culturing regimes` specifying the controllable experimental conditions for that part of the experiment, together with the yeast strain used. Beside these conditions, measurements from the different modalities are recorded, along with sampling meta-data, such as the sample volume and sampling times.

A simple experiment selection demonstration was used to show how this database could be used. Previously published experimental data, including

both experimental conditions and results, were saved. Having this in the same database simplified the analysis and suggestion of interesting experiments to perform. Along with the ontology we also provide an easily deployed Apache Jena based database implementation.

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 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.

3.2 Paper II: RIMBO - An Ontology for Model Revision Databases

In this paper we present RIMBO (Revisions for Improvements of Models in Biology Ontology), an ontology to systematically describe changes to computational biology models, intended as the foundation for a database containing iteratively revised models.

Problem

Computational models are fundamental across several scientific domains, including biology. Such models are not static, instead they evolve as new knowledge is discovered and this evolution is typically iterative. This process of refining models is often poorly documented and multiple smaller changes are bundled together in larger releases, making it impossible to reason about the impact of individual changes.

As discussed in Section 2.1, autonomous agents with access to knowledge and physical laboratories are a promising way forward for scientific discovery. In the domain of systems biology a common representation of knowledge is computational models, and thus, improving them is of great scientific interest. If this is done autonomously, the recording of the revisions, along with descriptions and reasons for them, is central, since it facilitates reasoning about previous changes and can help in communicating the results.

Furthermore, the difficulty of reproducing results across various scientific disciplines is a well-recognised challenge. One contributing factor is the lack of structured documentation of the research process. Maintaining a detailed record of a model's origin, the modifications made to it, and the rationale behind those changes could significantly improve transparency and traceability. This holds true even when models are developed through traditional means by human researchers, rather than through automated systems.

Approach and Contributions

RIMBO combines classes from several ontologies to describe changes made to a model, as can be seen in Fig. 3.1. Classes from the Gene Ontology [49] describe what is modelled, and PROV-O [67] and REPRODUCE-ME [68] specifies important metadata for the update. Classes from COMODI [53] and SBO [69] describes the mechanistic part of the model being revised, and COMODI together with APO [51] gives the reason for a revision.

As models in biology often are expressed in text- or XML-based formats RIMBO allows to specify the revision as the diff-patch to reduce the storage footprint, allowing for large numbers of revisions to be saved.

We demonstrate how RIMBO can be used by modelling chains of revisions to the genome-scale metabolic model (GEM) Yeast 8 [70]. Initially we model parts of the community update from v8.4.1 to v8.4.2. We also include an update suggested by an autonomous agent through abductive reasoning [71], as well as thousands of random revisions.

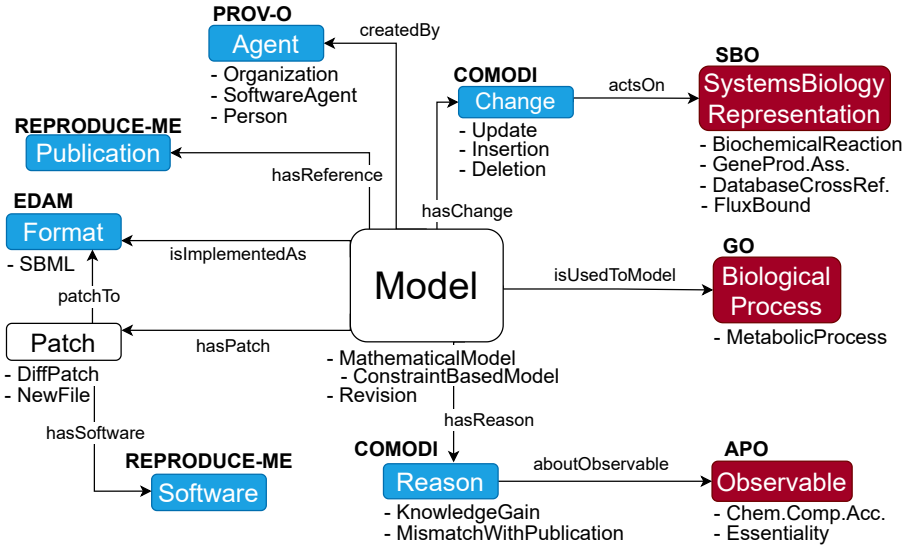


Figure 3.1: Overview of RIMBO showing classes, how they are connected, and which ontologies they are from. Under the boxes are examples of subclasses to describe model revisions. Red boxes denotes domain-specific classes that would need replacing if the ontology is applied to another domain. Blue denotes classes from other foundational scientific ontologies and the white boxes are classes introduced in RIMBO.

Conclusion

With RIMBO we propose a method to systematically describe revisions of computational models in a semantically meaningful way. We believe this adds value to both traditional and automated labs through the improved traceability and transparency in model development.

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 and the funding for the project was acquired by R.D.K

3.3 Paper III - Ontology-based box embeddings and knowledge graphs for predicting phenotypic traits in *Saccharomyces cerevisiae*

In this paper we construct a KG describing genes in *S. cerevisiae* using terms from several ontologies. Box embeddings are found for the class hierarchies and used together with GNNs to predict digenic gene deletion fitness from the KG. Using interpretability techniques we identify combinations of edges important for fitness. This finding is validated by a biological experiment showing an association between inositol utilization and osmotic stress resistance.

Problem

S. cerevisiae is very well-studied, serving as a model eukaryote helping us understand other organisms. Decades of research have generated a vast amount of knowledge, some of which is systematically curated in databases like the Saccharomyces Genome Database (SGD) [31] and BioCyc [28]. These resources contain detailed information about genes, phenotypes, biochemical pathways, and genetic interactions, providing a rich foundation for computational models.

Much of this knowledge is structured using ontologies, such as the Gene Ontology (GO) [49] for molecular functions, Ascomycete Phenotype Ontology (APO) [51] for phenotypic traits, and ChEBI [50] for chemical compounds. These ontologies capture hierarchical relationships and domain knowledge, making them valuable for machine learning approaches.

Despite this wealth of information, many gene functions and interactions remain unknown. Computational methods that leverage existing knowledge can help prioritise hypotheses and guide experimental efforts, reducing the need for exhaustive laboratory testing. By integrating structured knowledge from ontologies into predictive models, we can better utilise existing biological data to uncover new insights.

Approach and Contributions

Information from SGD and BioCyc is integrated in a KG describing genes in *S. cerevisiae*, using terms defined in various ontologies. The KG contains information about for example processes, functions, and phenotypes the genes are associated with, reactions catalysed by genes, and where in the cell the gene product is located. An overview of the graph can be seen in Fig. 3.2a.

To demonstrate the utility of the knowledge graph, we use a heterogeneous graph neural network to predict digenic deletion fitness, i.e., the impact of pairwise gene deletions on growth, based on data from Costanzo et al. [72]. We represent the nodes in the graph by Gumbel box embeddings [73] of the class hierarchies in the ontologies. The nodes are divided into the eight domains seen in Fig. 3.2b for which separate embeddings are found.

The performance for the GNN with box embeddings is compared to a GNN with shallow node embeddings learnt for the prediction task, as well as a

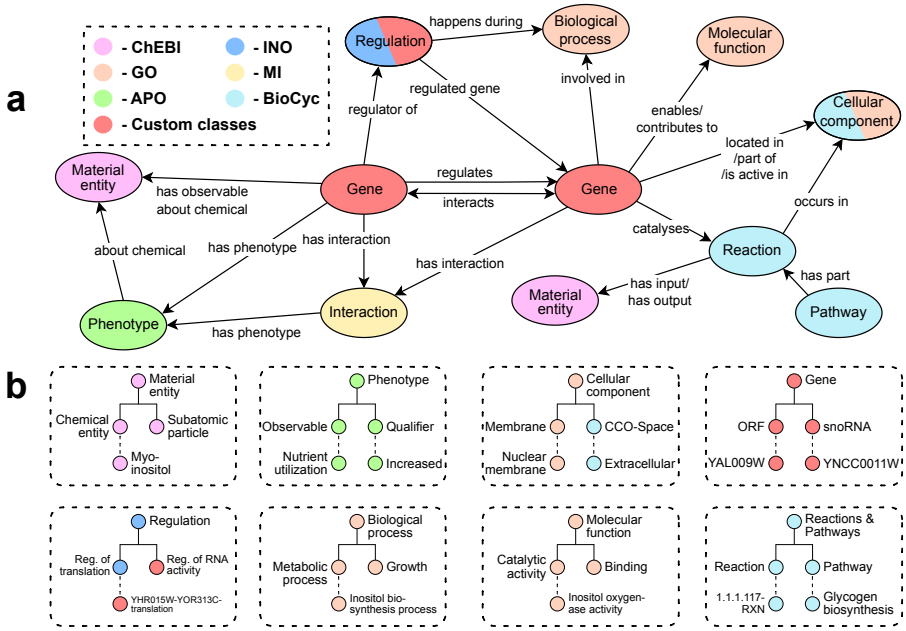


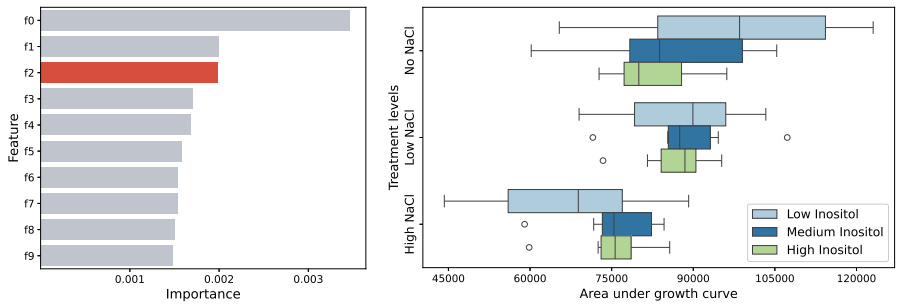
Figure 3.2: An overview of the different types of classes and how they are connected in the knowledge graph is shown in **a**. The color of the nodes specifies in which ontology the classes are defined. **b** shows examples of how the classes in each domain is organised according to hierarchies from the ontologies.

LightGBM model predicting from instantiations of the phenotype information in the graph. This comparison can be seen in Table 3.1.

The training of the GNN can be viewed as a way of distilling the biologically relevant information in the KG. Applying explainability techniques (input×gradient [74]) to the predictions assigned importance values to individual edges, as well as co-occurring edge pairs important across all predictions. These pairs are interpreted as hypotheses about potentially interacting traits. Wanting to verify this in a biological experiment, we filter the edge-pairs for edge-types corresponding to experiments we feasibly and safely can perform given our lab setup. This resulted in nutrient utilisation of inositol (vitamin B8) and stress resistance to NaCl (table salt) being the highest ranked pair and the hypothesis being that these traits have an interacting effect. The hypothesis was tested by cultivating an Δ*ino1* mutant, which cannot synthesise inositol, in varying inositol and NaCl concentrations. This experiment supported our hypothesis, showing a significant interaction effect. A summary of the feature selection and experimental results can be seen in Fig. 3.3. The literature suggests a likely mechanistic explanation: inositol, being involved in the biosynthesis and maintenance of cell membrane integrity [75], may exert a mitigating effect on the osmotic stress induced by NaCl. Importantly, this information was not present in the KG and thereby not "known" to the model when suggesting the hypothesis.

Table 3.1: Results from 10-fold cross-validation of digenic deletion fitness. Both GNN models share the same architecture but differ in class representations: one uses box embeddings for ontology hierarchies, while the other employs task-specific shallow embeddings. The instantiation model uses a sparse feature matrix with non-zero entries for phenotype annotations from the KG.

Model	Mean R^2	SD
GNN with box embeddings	**0.360**	0.043
GNN without box embeddings	0.329	0.043
Instantiations + LightGBM	0.211	0.022



(a) Feature importances

(b) AUC growth boxplots

Figure 3.3: An overview of the selection and results of the experiment we performed. (*a*) shows the highest ranked importances of edge-pairs and the pair selected for the experiment, nutrient utilisation of inositol and stress resistance to NaCl, is highlighted in red. *f0* and *f1*, which have a higher assigned weight, are discarded due to safety and lab constraints as it involves the chemical bleomycin. (*b*) Box plot showing the distribution of AUC for all of the experimental conditions tested. Inositol supplementation significantly impacts growth dynamics in high doses ($p < 0.05$). NaCl stress changes the impact of inositol in a dose dependent manner, suggesting an interactive effect ($p < 0.05$).

Conclusion

With this work, we demonstrate how KGs can be used to represent heterogeneous, qualitative information in a way that enables the prediction of quantitative data. Furthermore, we show that ontologies contain valuable domain knowledge that can be effectively captured through box embeddings. Finally, we leverage the model not only for prediction but also for hypothesis generation, identifying a potential interaction between phenotypic traits that was subsequently supported by a biological experiment.

Author contributions

The conceptualisation of the project was done by **Filip Kronström**, Ross D. King, and Daniel Brunnsåker. **F.K.** designed the knowledge graph and the machine learning models. **F.K.** and D.B. decided on the biological experiment. D.B. performed and analysed the biological experiment. The project was supervised by R.D.K., and Ievgeniia A. Tiukova and the funding for the project was acquired by R.D.K

Chapter 4

Discussion and Future Work

This thesis presents three papers related to knowledge representation and scientific discovery. **Papers I** and **II** present ontologies to describe μ-chemostat experiments and revisions to biological models. **Paper III** illustrates how ontologies can be used together with publicly available data for prediction of traits in *S. cerevisiae*, as well as generation of hypothesis about new knowledge. This demonstrates that task-specific ontologies may need to be developed for certain applications, but existing ontologies can also provide valuable information that can be effectively leveraged.

Paper I is particularly relevant for the robot scientist Genesis, as it provides a structured experiment record. Such a record is essential not only for enabling the robot scientist to reason about past experimental conditions and outcomes but also for effectively communicating experimental results. The Genesis platform is unique due to its combination of cultivation capabilities and diverse measurement modalities. As a result, a custom ontology was developed, both a key contribution and a limitation of the work. While the ontology is tailored to this rather specific experimental setup, the semantic annotation of data facilitates data sharing, even if reproducing the experiments exactly may require specialised hardware. Although semantic annotation introduces a slight storage overhead, it enables integration with other knowledge sources, which is valuable in the context of a robot scientist system. It is also worth noting that the Genesis hardware is still under development; therefore, the ontology has not yet been fully validated for its intended use and may undergo further refinement.

The ontology in **Paper II** is developed for closed-loop model improvement and any research context where (particularly mechanistic) models undergo iterative revision. It provides a structured record of model changes, which is essential for autonomous agents engaged in reasoning and decision making, and it also enhances transparency and reproducibility for human researchers. The ontology can be interpreted as a formal representation of a specific type of hypothesis: proposed updates to computational models. However, the current framework does not capture the testing or validation of these hypotheses.

For example, it lacks mechanisms to associate model changes with statistical evaluations or comparisons to experimental data, an essential component in determining the impact of a revision. In a closed-loop modelling setup, such evaluation steps should be explicitly tracked. While the paper focuses on systems biology applications, the ontology itself is largely domain agnostic. With only minor adaptations, such as substituting a few domain specific terms, it can be applied to model revision workflows in other scientific fields.

Paper III demonstrates how combining hierarchical knowledge from ontologies with factual data from biological databases can enhance the prediction of quantitative traits in *S. cerevisiae*. By incorporating symbolic knowledge into neural models, specifically through box embeddings of ontology hierarchies, the study shows improved predictive performance compared to models that learn task-specific class representations from scratch. This is particularly noteworthy because hierarchical embeddings provide a rather simple representation of a domain, and they are more feasible for domain experts to construct than more complex, high-dimensional representations.

The model exhibited promising generalisation: although trained on digenic gene deletion fitness data, it also performed reasonably well on predicting trigenic deletion fitness, with the caveat that this was tested on a smaller and more limited dataset. This raises the question of whether the gene embeddings learned from one task can be effectively transferred to other, more distinct tasks.

Beyond trait prediction, the model was used to generate hypotheses about potential interactions between traits by identifying important co-occurring edges in the knowledge graph. One such hypothesis, suggesting an interaction between inositol utilisation and NaCl stress resistance, was supported by follow-up experiments in an autonomous laboratory. While this interaction may not be novel from a biological perspective, it was not explicitly encoded in the knowledge graph, and therefore represented a discovery from the model's point of view. Importantly, the process of selecting testable hypotheses involved filtering based on the constraints of the available laboratory setup. Modifying this filtering process could lead to the generation of more scientifically interesting or unexpected hypotheses. In the context of robot scientist systems this approach provides a promising strategy for incorporating background knowledge into automated hypothesis generation.

Together, the three presented papers illustrate how semantically precise knowledge representations can help at different stages of scientific discovery. Both for organisation and to facilitate sharing of generated data, as well as to make sense of and make new discoveries from already known information.

Future work

Looking ahead, several promising directions for future work emerge, particularly in relation to **Paper III**. First, as discussed above, the hypotheses generated by the current system are not always of high biological interest. Exploring ways to improve the scientific relevance of these hypotheses would be valuable. One possible avenue is the incorporation of large language models (LLMs),

which could assist in ranking or refining hypotheses based on broader biological knowledge or literature context.

Another extension concerns the embeddings generated by the model. While the input to the network includes box embeddings that capture hierarchical information, this structure is not preserved in the embeddings generated by the GNN. Introducing mechanisms to propagate the hierarchical structure, such as using a semantic loss function [76], could help ensure that the final representations remain grounded in domain knowledge.

Finally, a natural continuation of the work in **Paper II** involves the representation of more general forms of hypotheses. By integrating known information, e.g., through embedding-based approaches like those explored in **Paper III**, this could allow for estimation of the prior support for hypotheses. This capability is particularly important for scaling up hypothesis generation and testing in autonomous systems, where the space of possible hypotheses far exceeds what can be practically tested in the laboratory.

Bibliography

[1] D. Botstein, S. A. Chervitz and M. Cherry, 'Yeast as a model organism,' *Science*, vol. 277, no. 5330, pp. 1259–1260, 29th Aug. 1997, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.277.5330.1259 (cit. on p. 3).

[2] E. Nevoigt, 'Progress in metabolic engineering of saccharomyces cerevisiae,' *Microbiology and Molecular Biology Reviews*, vol. 72, no. 3, pp. 379–412, Sep. 2008, Publisher: American Society for Microbiology. DOI: 10.1128/mmbr.00025-07 (cit. on p. 3).

[3] R. Davis, H. Shrobe and P. Szolovits, 'What is a knowledge representation?' *AI Mag.*, vol. 14, no. 1, pp. 17–33, 1st Mar. 1993, ISSN: 0738-4602. DOI: 10.1609/aimag.v14i1.1029. Accessed: 24th Apr. 2025. [Online]. Available: https://doi.org/10.1609/aimag.v14i1.1029 (cit. on p. 3).

[4] D. Brunnsåker, 'Automating hypothesis generation and testing: Towards self-driving biology,' ISBN: 9789181032970, Ph.D. dissertation, Chalmers University of Technology, 2025. DOI: 10.63959/chalmers.dt/5755 (cit. on p. 4).

[5] B. G. Buchanan and E. A. Feigenbaum, 'Dendral and meta-dendral: Their applications dimension,' *Artificial Intelligence*, Applications to the Sciences and Medicine, vol. 11, no. 1, pp. 5–24, 1st Aug. 1978, ISSN: 0004-3702. DOI: 10.1016/0004-3702(78)90010-3 (cit. on p. 7).

[6] P. W. Langley, 'BACON: A production system that discovers empirical laws,' in *Proceedings of the 5th international joint conference on Artificial intelligence - Volume 1*, ser. IJCAI'77, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 22nd Aug. 1977, p. 344 (cit. on p. 7).

[7] J. Jumper et al., 'Highly accurate protein structure prediction with AlphaFold,' *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 7).

[8] R. D. King et al., 'The automation of science,' *Science*, vol. 324, no. 5923, pp. 85–89, 3rd Apr. 2009, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.1165620 (cit. on pp. 7, 8).

[9] R. D. King et al., 'Functional genomic hypothesis generation and experimentation by a robot scientist,' *Nature*, vol. 427, no. 6971, pp. 247–252, Jan. 2004, Number: 6971 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature02236 (cit. on p. 7).

[10] K. Williams et al., 'Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases,' *Journal of The Royal Society Interface*, vol. 12, no. 104, p. 20 141 289, 6th Mar. 2015, Publisher: Royal Society. DOI: 10.1098/rsif.2014.1289 (cit. on p. 7).

[11] B. Burger et al., 'A mobile robotic chemist,' *Nature*, vol. 583, no. 7815, pp. 237–241, Jul. 2020, Number: 7815 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2442-2. Accessed: 7th Jun. 2023 (cit. on p. 7).

[12] I. A. Tiukova et al., *Genesis: Towards the automation of systems biology research*, 4th Sep. 2024. DOI: 10.48550/arXiv.2408.10689. arXiv: 2408.10689[cs] (cit. on p. 8).

[13] G. S. Omenn, 'Grand challenges and great opportunities in science, technology, and public policy,' *Science*, vol. 314, no. 5806, pp. 1696–1704, 15th Dec. 2006, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.1135003 (cit. on p. 8).

[14] A. Coutant et al., 'Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast,' *Proceedings of the National Academy of Sciences*, vol. 116, no. 36, pp. 18 142–18 147, 3rd Sep. 2019, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1900548116 (cit. on pp. 8, 10).

[15] K. Roper et al., 'Testing the reproducibility and robustness of the cancer biology literature by robot,' *Journal of The Royal Society Interface*, vol. 19, no. 189, p. 20 210 821, 6th Apr. 2022, Publisher: Royal Society. DOI: 10.1098/rsif.2021.0821. Accessed: 26th Feb. 2025 (cit. on p. 8).

[16] P. Ivie and D. Thain, 'Reproducibility in scientific computing,' *ACM Comput. Surv.*, vol. 51, no. 3, 63:1–63:36, 16th Jul. 2018, ISSN: 0360-0300. DOI: 10.1145/3186266 (cit. on p. 8).

[17] P Kohl, E. J. Crampin, T. A. Quinn and D Noble, 'Systems biology: An approach,' *Clinical Pharmacology & Therapeutics*, vol. 88, no. 1, pp. 25–33, 2010, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1038/clpt.2010.92, ISSN: 1532-6535. DOI: 10.1038/clpt.2010.92 (cit. on p. 8).

[18] M. H. V. V. Regenmortel, 'Reductionism and complexity in molecular biology,' *EMBO reports*, vol. 5, no. 11, pp. 1016–1020, Nov. 2004, Num Pages: 1020 Publisher: John Wiley & Sons, Ltd, ISSN: 1469-221X. DOI: 10.1038/sj.embor.7400284 (cit. on p. 8).

[19] F. Crick, 'Central dogma of molecular biology,' *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/227561a0 (cit. on p. 8).

[20] H. Kitano, 'Systems biology: A brief overview,' *Science*, vol. 295, no. 5560, pp. 1662–1664, Mar. 2002, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.1069492 (cit. on p. 9).

[21] K. Dettmer, P. A. Aronov and B. D. Hammock, 'Mass spectrometry-based metabolomics,' *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007, ISSN: 1098-2787. DOI: 10.1002/mas.20108. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.20108 (cit. on p. 9).

[22] I. Tavassoly, J. Goldfarb and R. Iyengar, 'Systems biology primer: The basic methods and approaches,' *Essays in Biochemistry*, vol. 62, no. 4, pp. 487–500, 4th Oct. 2018, ISSN: 0071-1365. DOI: 10.1042/EBC20180003. Accessed: 16th Apr. 2025 (cit. on p. 9).

[23] M. L. Mo, B. Ø. Palsson and M. J. Herrgård, 'Connecting extracellular metabolomic measurements to intracellular flux states in yeast,' *BMC Systems Biology*, vol. 3, no. 1, p. 37, 25th Mar. 2009, ISSN: 1752-0509. DOI: 10.1186/1752-0509-3-37 (cit. on p. 10).

[24] J. D. Orth, I. Thiele and B. Ø. Palsson, 'What is flux balance analysis?' *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, Mar. 2010, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.1614 (cit. on p. 10).

[25] C. Zhang et al., 'Yeast9: A consensus genome-scale metabolic model for s. cerevisiae curated by the community,' *Molecular Systems Biology*, vol. 20, no. 10, pp. 1134–1150, 12th Aug. 2024, ISSN: 1744-4292. DOI: 10.1038/s44320-024-00060-7 (cit. on p. 10).

[26] M. Gillespie et al., 'The reactome pathway knowledgebase 2022,' *Nucleic Acids Research*, vol. 50, pp. D687–D692, D1 7th Jan. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkab1028 (cit. on p. 10).

[27] M. Kanehisa and S. Goto, 'KEGG: Kyoto encyclopedia of genes and genomes,' *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 1st Jan. 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27 (cit. on p. 10).

[28] P. D. Karp et al., 'The BioCyc collection of microbial genomes and metabolic pathways,' *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1085–1093, 19th Jul. 2019, ISSN: 1477-4054. DOI: 10.1093/bib/bbx085 (cit. on pp. 10, 23).

[29] O. Yurekten et al., 'MetaboLights: Open data repository for metabolomics,' *Nucleic Acids Research*, vol. 52, pp. D640–D646, D1 5th Jan. 2024, ISSN: 0305-1048. DOI: 10.1093/nar/gkad1045 (cit. on p. 10).

[30] Y. Perez-Riverol et al., 'The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences,' *Nucleic Acids Research*, vol. 50, pp. D543–D552, D1 7th Jan. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkab1038 (cit. on p. 10).

[31] S. R. Engel et al., 'Saccharomyces genome database: Advances in genome annotation, expanded biochemical pathways, and other key enhancements,' *Genetics*, iyae185, 12th Nov. 2024, ISSN: 1943-2631. DOI: 10.1093/genetics/iyae185 (cit. on pp. 10, 23).

[32] D. Brunnsåker, F. Kronström, I. A. Tiukova and R. D. King, 'Interpreting protein abundance in saccharomyces cerevisiae through relational learning,' *Bioinformatics*, vol. 40, no. 2, btae050, 1st Feb. 2024, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae050. Accessed: 17th Feb. 2025 (cit. on p. 10).

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

[34] A. H. Gower et al., *The use of AI-robotic systems for scientific discovery*, 25th Jun. 2024. DOI: 10.48550/arXiv.2406.17835. arXiv: 2406.17835[cs]. Accessed: 16th Apr. 2025 (cit. on p. 10).

[35] M. R. Dietrich, R. A. Ankeny and P. M. Chen, 'Publication trends in model organism research,' *Genetics*, vol. 198, no. 3, pp. 787–794, Nov. 2014, ISSN: 0016-6731. DOI: 10.1534/genetics.114.169714 (cit. on p. 10).

[36] S.-F. Duan et al., 'The origin and adaptive evolution of domesticated populations of yeast from far east asia,' *Nature Communications*, vol. 9, no. 1, p. 2690, 12th Jul. 2018, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-018-05106-7 (cit. on p. 10).

[37] L. Vanderwaeren, R. Dok, K. Voordeckers, S. Nuyts and K. J. Verstrepen, 'Saccharomyces cerevisiae as a model system for eukaryotic cell biology, from cell cycle control to DNA damage response,' *International Journal of Molecular Sciences*, vol. 23, no. 19, p. 11 665, 1st Oct. 2022, ISSN: 1422-0067. DOI: 10.3390/ijms231911665 (cit. on p. 10).

[38] Z. Yang and M. Blenner, 'Genome editing systems across yeast species,' *Current Opinion in Biotechnology*, vol. 66, pp. 255–266, Dec. 2020, ISSN: 1879-0429. DOI: 10.1016/j.copbio.2020.08.011 (cit. on p. 11).

[39] A. Goffeau et al., 'Life with 6000 genes,' *Science*, vol. 274, no. 5287, pp. 546–567, 25th Oct. 1996, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.274.5287.546 (cit. on p. 11).

[40] V. Wood, A. Lock, M. A. Harris, K. Rutherford, J. Bähler and S. G. Oliver, 'Hidden in plain sight: What remains to be discovered in the eukaryotic proteome?' *Open Biology*, vol. 9, no. 2, p. 180 241, 20th Feb. 2019, ISSN: 2046-2441. DOI: 10.1098/rsob.180241. Accessed: 11th Feb. 2025 (cit. on p. 11).

[41] F. Baader, I. Horrocks, C. Lutz and U. Sattler, *An Introduction to Description Logic*. Cambridge: Cambridge University Press, 2017, ISBN: 978-0-521-87361-1. DOI: 10.1017/9781139025355. Accessed: 8th Apr. 2025 (cit. on pp. 11–13).

[42] T. R. Gruber, 'A translation approach to portable ontology specifications,' *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1st Jun. 1993, ISSN: 1042-8143. DOI: 10.1006/knac.1993.1008 (cit. on p. 11).

[43] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi and P. F. Patel-Schneider, Eds., *The Description Logic Handbook: Theory, Implementation and Applications*, 2nd ed. Cambridge: Cambridge University Press, 2007, ISBN: 978-0-521-15011-8. DOI: 10.1017/CBO9780511711787. Accessed: 11th Apr. 2025 (cit. on p. 12).

[44] F. Baader and P. Hanschke, 'A scheme for integrating concrete domains into concept languages,' in *Proceedings of the 12th international joint conference on Artificial intelligence - Volume 1*, ser. IJCAI'91, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 24th Aug. 1991, pp. 452–457, ISBN: 978-1-55860-160-4. Accessed: 14th Apr. 2025 (cit. on p. 12).

[45] I. Horrocks, 'Ontologies and the semantic web,' *Commun. ACM*, vol. 51, no. 12, pp. 58–67, 1st Dec. 2008, ISSN: 0001-0782. DOI: 10.1145/1409360.1409377. Accessed: 14th Apr. 2025 (cit. on p. 12).

[46] A. Hogan, 'The semantic web: Two decades on,' *Semantic Web*, vol. 11, no. 1, pp. 169–185, 31st Jan. 2020, Publisher: SAGE Publications, ISSN: 1570-0844. DOI: 10.3233/SW-190387. Accessed: 14th Apr. 2025 (cit. on p. 12).

[47] L. F. Sikos, 'Description logics: Formal foundation for web ontology engineering,' in *Description Logics in Multimedia Reasoning*, L. F. Sikos, Ed., Cham: Springer International Publishing, 2017, pp. 67–120, ISBN: 978-3-319-54066-5. DOI: 10.1007/978-3-319-54066-5_4. Accessed: 15th Apr. 2025 (cit. on p. 13).

[48] F. Baader, S. Brandt and C. Lutz, 'Pushing the EL envelope,' in *Proceedings of the 19th international joint conference on Artificial intelligence*, ser. IJCAI'05, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 30th Jul. 2005, pp. 364–369 (cit. on p. 13).

[49] M. Ashburner et al., 'Gene ontology: Tool for the unification of biology,' *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000, Number: 1 Publisher: Nature Publishing Group, ISSN: 1546-1718. DOI: 10.1038/75556 (cit. on pp. 13, 14, 21, 23).

[50] J. Hastings et al., 'ChEBI in 2016: Improved services and an expanding collection of metabolites,' *Nucleic acids research*, vol. 44, pp. D1214–9, D1 1st Jan. 2016, ISSN: 1362-4962. DOI: 10.1093/nar/gkv1031. Accessed: 29th Jan. 2025 (cit. on pp. 13, 14, 23).

[51] M. C. Costanzo et al., 'New mutant phenotype data curation system in the saccharomyces genome database,' *Database: The Journal of Biological Databases and Curation*, vol. 2009, bap001, 2009, ISSN: 1758-0463. DOI: 10.1093/database/bap001 (cit. on pp. 13, 21, 23).

[52] A. Bandrowski et al., 'The ontology for biomedical investigations,' *PLoS ONE*, vol. 11, no. 4, e0154556, 29th Apr. 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0154556. Accessed: 4th Apr. 2025 (cit. on pp. 14, 19).

[53] M. Scharm, D. Waltemath, P. Mendes and O. Wolkenhauer, 'COMODI: An ontology to characterise differences in versions of computational models in biology,' *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 46, 11th Jul. 2016, ISSN: 2041-1480. DOI: 10.1186/s13326-016-0080-2 (cit. on pp. 14, 21).

[54] A. Hogan et al., *Knowledge Graphs* (Synthesis Lectures on Data, Semantics, and Knowledge 22), English. Springer, 2021, ISBN: 9783031007903. DOI: 10.2200/S01125ED1V01Y202109DSK022. [Online]. Available: https://kgbook.org/ (cit. on p. 15).

[55] L. D. Raedt, Ed., *Logical and Relational Learning*, Cognitive Technologies, ISSN: 1611-2482, Berlin, Heidelberg: Springer, 2008, ISBN: 978-3-540-20040-6 978-3-540-68856-3. DOI: 10.1007/978-3-540-68856-3 (cit. on p. 16).

[56] L. A. Galárraga, C. Teflioudi, K. Hose and F. Suchanek, 'AMIE: Association rule mining under incomplete evidence in ontological knowledge bases,' in *Proceedings of the 22nd international conference on World Wide Web*, ser. WWW '13, New York, NY, USA: Association for Computing Machinery, 13th May 2013, pp. 413–422, ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488425 (cit. on p. 16).

[57] L. Vig, A. Srinivasan, M. Bain and A. Verma, 'An investigation into the role of domain-knowledge on the use of embeddings,' in *Inductive Logic Programming*, N. Lachiche and C. Vrain, Eds., Cham: Springer International Publishing, 2018, pp. 169–183, ISBN: 978-3-319-78090-0. DOI: 10.1007/978-3-319-78090-0_12 (cit. on p. 16).

[58] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston and O. Yakhnenko, 'Translating embeddings for modeling multi-relational data,' in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13, vol. 2, Red Hook, NY, USA: Curran Associates Inc., 5th Dec. 2013, pp. 2787–2795. Accessed: 17th Feb. 2025 (cit. on p. 16).

[59] M. Kulmanov, W. Liu-Wei, Y. Yan and R. Hoehndorf, 'EL embeddings: Geometric construction of models for the description logic EL++,' in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 6103–6109, ISBN: 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/845 (cit. on p. 16).

[60] Özçep, Özgür Lütfü, M. Leemhuis and D. Wolter, 'Cone semantics for logics with negation,' presented at the Twenty-Ninth International Joint Conference on Artificial Intelligence, vol. 2, 9th Jul. 2020. DOI: 10.24963/ijcai.2020/252 (cit. on p. 16).

[61] B. Xiong, N. Potyka, T.-K. Tran, M. Nayyeri and S. Staab, 'Faithful embeddings for EL++ knowledge bases,' in *The Semantic Web – ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*, Berlin, Heidelberg: Springer-Verlag, 23rd Oct. 2022. DOI: 10.1007/978-3-031-19433-7_2 (cit. on p. 16).

[62] X. Peng, Z. Tang, M. Kulmanov, K. Niu and R. Hoehndorf, *Description logic EL++ embeddings with intersectional closure*, 28th Feb. 2022. DOI: 10.48550/arXiv.2202.14018. arXiv: 2202.14018[cs]. Accessed: 17th Feb. 2025 (cit. on p. 16).

[63] M. Jackermeier, J. Chen and I. Horrocks, 'Dual box embeddings for the description logic EL++,' in *Proceedings of the ACM Web Conference 2024*, ser. WWW '24, New York, NY, USA: Association for Computing Machinery, 13th May 2024. DOI: 10.1145/3589334.3645648 (cit. on p. 16).

[64] M. Leemhuis and O. Kutz, 'Understanding the expressive capabilities of knowledge base embeddings under box semantics,' in *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning*, 2025 (cit. on p. 16).

[65] W. L. Hamilton, R. Ying and J. Leskovec, 'Inductive representation learning on large graphs,' in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA: Curran Associates Inc., 4th Dec. 2017, pp. 1025–1035, ISBN: 978-1-5108-6096-4 (cit. on p. 16).

[66] R. Jackson et al., 'OBO foundry in 2021: Operationalizing open data principles to evaluate ontologies,' *Database*, vol. 2021, baab069, 29th Sep. 2021, ISSN: 1758-0463. DOI: 10.1093/database/baab069. Accessed: 4th Apr. 2025 (cit. on p. 19).

[67] T. Lebo et al., 'PROV-o: The PROV ontology,' World Wide Web Consortium, Tech. Rep., 30th Apr. 2013, Publisher: World Wide Web Consortium. Accessed: 10th May 2023. [Online]. Available: https://research.manchester.ac.uk/en/publications/prov-o-the-prov-ontology (cit. on p. 21).

[68] S. Samuel and B. König-Ries, 'End-to-end provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach,' *Journal of Biomedical Semantics*, vol. 13, no. 1, p. 1, 6th Jan. 2022, ISSN: 2041-1480. DOI: 10.1186/s13326-021-00253-1 (cit. on p. 21).

[69] N. Juty and N. le Novère, 'Systems biology ontology,' in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho and H. Yokota, Eds., New York, NY: Springer, 2013, pp. 2063–2063, ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_1287 (cit. on p. 21).

[70] H. Lu et al., 'A consensus s. cerevisiae metabolic model yeast8 and its ecosystem for comprehensively probing cellular metabolism,' *Nature Communications*, vol. 10, no. 1, p. 3586, 8th Aug. 2019, Number: 1, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-019-11581-3 (cit. on p. 21).

[71] A. H. Gower, K. Korovin, D. Brunnsåker, I. A. Tiukova and R. D. King, 'LGEM+: A first-order logic framework for automated improvement of metabolic network models through abduction,' in *Discovery Science*, A. Bifet, A. C. Lorena, R. P. Ribeiro, J. Gama and P. H. Abreu, Eds., Cham: Springer Nature Switzerland, 2023, pp. 628–643, ISBN: 978-3-031-45275-8. DOI: 10.1007/978-3-031-45275-8_42 (cit. on p. 21).

[72] M. Costanzo et al., 'A global genetic interaction network maps a wiring diagram of cellular function,' *Science*, vol. 353, no. 6306, aaf1420, 23rd Sep. 2016, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.aaf1420 (cit. on p. 23).

[73] S. S. Dasgupta, M. Boratko, D. Zhang, L. Vilnis, X. L. Li and A. McCallum, 'Improving local identifiability in probabilistic box embeddings,' in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Red Hook, NY, USA: Curran Associates Inc., 6th Dec. 2020, pp. 182–192, ISBN: 978-1-7138-2954-6 (cit. on p. 23).

[74] A. Shrikumar, P. Greenside, A. Shcherbina and A. Kundaje, *Not just a black box: Learning important features through propagating activation differences*, 11th Apr. 2017. DOI: 10.48550/arXiv.1605.01713. arXiv: 1605.01713[cs]. Accessed: 26th Jan. 2025 (cit. on p. 24).

[75] M. R. Culbertson and S. A. Henry, 'INOSITOL-REQUIRING MUTANTS OF SACCHAROMYCES CEREVISIAE,' *Genetics*, vol. 80, no. 1, pp. 23–40, 1st May 1975, ISSN: 1943-2631. DOI: 10.1093/genetics/80.1.23. Accessed: 4th Mar. 2025 (cit. on p. 24).

[76] J. Xu, Z. Zhang, T. Friedman, Y. Liang and G. Van den Broeck, 'A semantic loss function for deep learning with symbolic knowledge,' in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 5502–5511 (cit. on p. 29).