# The use of AI-robotic systems for scientific discovery

Alexander H. Gower[1(†)], Konstantin Korovin[2], Daniel Brunnsåker[1], Filip Kronström[1], Gabriel K. Reder[3], Ievgeniia A. Tiukova[1,3], Ronald S. Reiserer[4], John P. Wikswo[4], and Ross D. King[1,5,6]

[1] Chalmers University of Technology, Gothenburg, Sweden
{gower,danbru,filipkro,tiukova,rossk}@chalmers.se
[2] The University of Manchester, Manchester, United Kingdom
Konstantin.Korovin@manchester.ac.uk
[3] KTH Royal Institute of Technology, Stockholm, Sweden
reder@kth.se
[4] Vanderbilt University, Nashville, TN, USA
{ron.reiserer,john.p.wikswo}@vanderbilt.edu
[5] Cambridge University, Cambridge, United Kingdom
[6] Alan Turing Institute, London, United Kingdom

**Abstract.** The process of developing theories and models and testing them with experiments is fundamental to the scientific method. Automating the entire scientific method then requires not only automation of the induction of theories from data, but also experimentation from design to implementation. This is the idea behind a robot scientist—a coupled system of AI and laboratory robotics that has agency to test hypotheses with real-world experiments.

In this chapter we explore some of the fundamentals of robot scientists in the philosophy of science. We also map the activities of a robot scientist to machine learning paradigms, and argue that the scientific method shares an analogy with active learning.

We demonstrate these concepts using examples from previous robot scientists, and also from Genesis: a next generation robot scientist designed for research in systems biology, comprising a micro-fluidic system with 1000 computer-controlled micro-bioreactors and interpretable models based in controlled vocabularies and logic.

**Keywords:** Robot scientist · scientific discovery · active learning · laboratory robotics.

## 1 Introduction

The use of AI-robotic systems in scientific research has been demonstrated not only possible, but fruitful, in the past two decades, with several projects—from Adam [16] and Eve [34] to the Robot Chemist [6]—proving the value of coupling AI software agents with experimental platforms to give them real-world

agency. Research toward the goal of productive robot scientists is necessarily a multi-disciplinary endeavour. Fields that contribute include: artificial intelligence, robotics, nanotechnology and materials science. For robot scientist projects to be a success, researchers need a shared understanding of the process they are working to automate: the scientific method.

We begin this chapter by defining some of the core concepts in the philosophical ideas behind robot scientists in Section 2, beginning with theories and models, the fundamental entities in scientific research. We explore the concept of parsimony as it relates to the scientific method, something which is learned through scientific education in humans but requires explicit care when creating robot scientists. Then we cover methods of inference, and the difference scientific values which we use to assess the value of a theory or model. Section 3 aims to identify which machine learning paradigm is most apt for scientific discovery. Ultimately, we conclude that scientific discovery shares an analogy with active learning, and that this, rather than reinforcement learning, is the most useful paradigm to adopt when designing or analysing robot scientists. Section 4 gives an overview of systems biology, the domain in which the next generation robot scientist Genesis is applied. We argue that as biological systems are "complex systems" as described by [33], they are excellent targets for study by robot scientists with their superhuman abilities in reasoning and precision. In Section 5 we present a case study of Genesis, and one model that we have developed with the aim of automated scientific discovery.

## 2 The philosophy of robot scientists

A robot scientist is defined in [16] as

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

Part of the motivation for building robot scientists is to understand more about the nature of science by building a system that can replicate the scientific process [17].

The goal of science is to develop theories that explain and predict phenomena in the real world. To develop theories science uses models, which are representations of theories in some localised context. Models are a surrogate for the system being studied (object system), in the sense that they display characteristics or behaviour, of sufficient similarity to that of the object system, that allow for indirect study of the object system by studying its surrogate (the subject system). Models are particularly useful when direct study of a system is impossible, impractical or undesirable.

Two examples of models in biology that illustrate the diversity of desirable properties of a model are:

1. an illustrated diagram of the cross-section of a cell; and
2. a metabolic network model (MNM) representing the rates of biochemical reactions and chemical compound abundances using a system of ordinary differential equations (ODEs) with independent variable $t$, time.

In both of these cases, the object system is the same: a cell. However they have quite different qualities. Model 1 would be well-suited to teaching high school students how cells of, for example, the yeast *Saccharomyces cerevisiae* function. However, Model 2 is capable of quantitative predictions, enabling direct comparison with quantitative experimental data.

Models with deductive capacity using mathematics, such as Model 2, are useful for all forms of scientific discovery, but particularly when automated. The scientific discovery problem becomes

> an incremental process of refinement [of the model] strongly guided by the empirical observations [9].

Methods of scientific enquiry rely on: constructing a good starting model; inferring changes to the model; techniques to reason about which models are better; and the collection of relevant, high-quality empirical data to drive the process.

The purpose of a robot scientist is to provide software and hardware that together can achieve each each of these activities, and join them together in a "closed-loop" form of enquiry without human intervention [16]. Through the process of designing a system capable of independent scientific inquiry, we seek insight into the scientific method itself, as well as the system subject to enquiry.

### 2.1 Core aspects of scientific method

There are three components of the scientific method that we will consider here, namely: parsimony, logical inference, and statistical inference. A history and detailed treatment of each of these components is covered in [10].

**2.1.1 Parsimony** Parsimony is the concept that when choosing between theories that fit the data equally well, the theory that is simplest is preferable. Sometimes referred to as "Ockham's razor", parsimony is absolutely essential for science, and it is easy to take this principle as a common sense notion and end further rumination on the subject. However, a proper treatment of the motivations for adopting parsimony as a guiding principle is warranted when it comes to the automation of discovery and systems biology.

Firstly, it is worth distinguishing epistemological parsimony—which we alluded to above, that between theories that fit data equally well the simplest is preferable—from ontological parsimony—that nature itself prefers simplicity. Arguments for ontological parsimony suffer from many counter-examples, for example the yeast genome went through a duplication during its evolution resulting in numerous

genes with overlapping or identical function [15]. However all scientific argument must adopt some version of ontological parsimony, as it is the basis for generalisation of theories.

Concrete examples of ontological parsimony include: taking the assumption that there exist common properties between individuals of the same species. This could be a chemical species, that all glucose molecules react with water molecules in the same ways. Or it could be a species of organism, implying that we can apply the same computational model to two individual colonies of *S. cerevisiae*.

These ideas of ontological parsimony are so central to scientific enquiry that researchers may not consider them—they become necessary implicit biases.

One power of appealing to ontological parsimony is that it is the basis of factorial experimentation, described in [7] as a two-step process: first we identify relevant factors for the phenomenon to be studied; then we design experiments holding certain factors constant and varying over others. Determining relevant factors means stating that certain factors are irrelevant, that we do not expect them to affect the outcome of the experiment. Deciding which factors to include is not an easy process. Designing experiments in this way allows the controlled study of phenomena so that we can test hypotheses that relate to a restricted subset of factors. Factorial design of experiments requires an appeal to ontological parsimony, informally

> we expect empirical data from experiments of the same class to exhibit only random variation,

where the class of experiment depends only on the variable factors. This assumption allows the inference of empirical laws about the phenomenon.

In practice we cannot usually keep constant all the factors we would like, so the class of experiment does not only depend on relevant variables. The nature of a given experimental protocol will introduce systematic errors: variations in the empirical data not arising from the experimental variables or random noise. Many methods exist to mitigate and model systematic errors, using randomisation techniques, systematic design, and statistics. Modelling systematic errors allows us to create a map on to each class of experiments as defined above. Randomisation, or systematic designs like Latin squares, can be impractical when using robot scientists due to the limitations of the automation hardware. One example from biology is that liquid handling procedures for a Latin square are impossible on certain liquid handling robots, and can increase procedure times by orders of magnitude on those with such a capability.

On the other hand, robot scientists have advantages when it comes to relying on and examining this underlying bias. Firstly, that robots are capable of performing repeated tasks with a much higher accuracy and precision than human counterparts, as is shown by examples of adoption of laboratory automation in biology [12], physics [28], and chemistry [32]. Secondly, that we can evaluate the validity of this assumption by recording more data about the execution of

experiments than is usually recorded when humans complete experiments. This is a "natural by-product" [16], as robots and software frequently have automatic logging capability. Finally, that we are forced to encode the experimental protocol forces the scientist to specify which experimental factors will be constant, or that if they are not constant they are judged irrelevant, and therefore they do not break the validity of statement above by creating a new class of experiments (for example, the use of different individual glass flasks of the same brand, model, and age).

Epistemological parsimony is a concept much more familiar to human scientists, or at least one which is applied more explicitly. It has been formalised in information theory as the minimum message length (MML) by combining Shannon's formula for the information ($I$) of an event ($E$), given by $I(E) = -log_2(P(E))$, with Bayes' theorem, $P(E_1 \cap E_2) = P(E_1)P(E_2|E_1)$. MML states that the hypothesis $H$ that best explains data $D$—in other words it maximises $P(H \cap D)$—is that which minimises the information (message length): $I(H \cap D) = I(H) + I(D|H)$. In this form we can see there is a trade-off between the first and second term meaning information is only added to a hypothesis if we can use this information to explain the data, which is the principle of Ockham's razor [2,31,1].

Epistemological parsimony is the basis for several fundamental concepts in the modern scientific method. One example is the idea of a null hypothesis—we present a null hypothesis stating that any variance in empirical data can be explained by our extant model, to a degree of certainty due to random noise. To reject the null hypothesis is to accede a more complex model, and we only do so if we are certain our current model is insufficient to explain the empirical data.

**2.1.2 Logical inference** Formally, logics are mathematical languages that relate premises and conclusions. There are various logics that can be used for reasoning in science, including propositional logic and first-order logic. We use logics in science to represent facts (observations of the real world) and laws (parts of theories or models) [7,10].

There are three basic forms of logical inference. Firstly, deductive logic, where conclusions are derived from premises and laws. A valid deductive argument is one that guarantees the truth of its conclusions given the truth of its premises. Deductive logic is what enables deterministic simulations, and how we reason about a hypothesis in the scientific method.

Inductive logic and abductive logic cannot provide such guarantees. In both cases, we seek explanations for a set of observable facts. In the case of induction, we are seeking laws that can explain a general case, whereas in the case of abduction we are seeking facts that can explain specific cases. Both induction and abduction deal in uncertainty, and for that we need probabilistic and statistical inference.

**2.1.3 Statistical inference and probability** Certain laws in science may only be expressed to a degree of certainty. For such situations we require statements

that can deal with uncertainty. Carnap defines two types of probability [7]. Statistical probability is a frequentist idea about the relative frequencies of mass events. The defining characteristic of statements of statistical probability is that they cannot be decided by logic, but rest on empirical observations. Logical probability on the other hand is the probability on a logical relation between two propositions. Gauch, Jr prefers to describe the two types of probability as being about events and beliefs respectively [10].

Using statistics, we can reason to a model given observations. In the case of abduction, we are reasoning about facts; with induction, laws. The statistical element of this reasoning represents the uncertainty around the model, which can be viewed either as a belief in the law, or perhaps more commonly in empirical science, the relative frequencies of events. Either way, these processes is crucial to the scientific method because they allow us to form and improve upon models.

### 2.2 Comparing scientific models

The relative merit of competing scientific models is not a trivial assessment. This problem is referred to in the philosophy of science as the problem of *theory choice*. For a robot scientist to be effective we must incorporate values and an evaluation procedure, otherwise the scientific discovery process will depend on human evaluation.

In discussing whether scientists follow philosophical virtues in their methods, Schindler [30] presents six virtues following Kuhn [19].

- **Internal consistency**: defined as the absence of contradictions within a theory. This can be extended to include the various contexts in which the theory is applied; in biology this could mean growth of *S. cerevisiae* in different conditions.
- **External consistency**: then the absence of contradictions with other scientific theories. For example, that our model of the biochemistry of yeast is thermodynamically consistent.
- **Empirical accuracy**: otherwise referred to as predictive power, this is the degree to which deductions from the theory match observations. For example, predictions of growth rates for colonies of yeast.
- **Scope:** otherwise referred to as unifying power, is the quality that a theory explains concepts relevant to different phenomena. An example in biology is that we have a unifying theory of genetics for DNA-based life, rather than separate theories for different species or kingdoms.
- **Simplicity**: which comes in various forms, depending on the context. Kuhn relates this to tractability using the example of Ptolemy's and Copernicus' theory of astronomy and the number of calculations needed for prediction being equivalent in both systems. In some theoretic sense, Copernicus' theory was simpler than Ptolemy's, having a simpler mathematical formulation.
- **Fruitfulness (or fertility)**: which has various interpretations. But can in one way be understood to be: how well does this theory lead to "more

science"? When combined with scope we obtain models that generalise well to new applications or other domains. Fruitfulness also has one clear implication for closed-loop discovery.

Importantly, Schindler [30] found that scientists do not agree on the relative importance of these characteristics, although there were some common views. And Kuhn argued that even if there were a common order and weighting, that two individual scientists may honestly differ in their assessment of a better theory because of the ways they evaluate them. This presents a difficulty for robot scientists as well. The difficulty of differences in evaluation is common to human scientists, so we can accept this as part of science. The unique difficulty is imbuing these values into the software used to evaluate theories.

Models trained on human knowledge will pick up some of the implicit biases in these data, which could be seen as a way to learn these values implicitly. This is particularly true of using foundation models such as large language models (LLMs, covered in Section 3) to evaluate theories. However, this can present risks, due to the lack of knowledge we have on their training and our inability to interrogate them.

Thankfully, the problem is not as difficult as it may initially appear, as machine learning research is itself informed by similar values. The task then for designing a robot scientist is to align mechanisms from AI research with scientific values in the relevant domain. A domain-specific cost function ensures accuracy is considered; likewise regularisation terms cover simplicity; and there are active research areas in enforcing external consistency on machine learning models, for example imposing symmetry constraints from physics on to neural networks.

## 3 Scientific discovery as machine learning

The problem of automating scientific discovery can be considered a machine learning problem. We cover how we can view the components of scientific discovery, as presented in Section 2, as components of machine learning techniques. There are many machine learning paradigms, so we also investigate which are the most appropriate for application to robot scientists.

Common to all machine learning techniques is that given some data to train with, we seek to learn a function that will perform in a desirable manner, and generalise beyond the given data.

In terms of the logical reasoning components of scientific discovery, forming a model ($\hat{h}$) from data is an induction problem. In machine learning settings we seek to provide good fit with our data; this could mean that we define a loss function for an optimizer, or in the case of inductive logic programming that we seek to explain positive examples and avoid inconsistencies with negative examples. To evaluate a candidate model, $\hat{h}$, we first obtain predictions from the model (deduction) and then perform some test of fit to the data, which is often a form of statistical reasoning. And finally, it is usually the case that there are

many possible models that may have similar performance in the fitness tests. In which case, most machine learning techniques appeal to parsimony to choose the model which is simpler, in some defined sense. This could be the inclusion of a regularisation term in a loss function, or selecting the simplest logic program that covers the examples.

Machine learning techniques can be placed into three broad categories: supervised learning, unsupervised learning, and reinforcement learning [29]. Semi-supervised learning is sometimes also included as a fourth category, being a hybrid of supervised and unsupervised learning.

- **Supervised learning** covers machine learning techniques that work from labelled data. These are input-output pairs, and the machine learning task is to learn a function that maps input to output.
- **Unsupervised learning** techniques seek to find structure and extract information from unlabelled data.
- **Reinforcement learning** techniques require feedback from the environment in which the learning agent is embedded. This feedback is used to evaluate actions taken by the agent, which information is used to optimise a strategy for the agent.

### 3.1 Reinforcement learning is an unsatisfactory paradigm for robot scientists

Recalling the definition of a robot scientist given in the introduction, the fact that there is an agent (the laboratory automation system) embedded in an environment (the physical surroundings of the lab) lends us to consider that discovery using a robot scientist is a reinforcement learning problem.

Analysing the system from the reinforcement learning perspective we must identify: (A) the agent; (B) the environment in which the agent is embedded; and (C) the reward function that evaluates the agent's actions in its environment.

We see that for pairs (A, B) we have several choices. Here we consider a few.

1. (laboratory robotic system, physical lab surroundings)
2. (experiment selector, experimental design space)
3. (model improvement algorithm, model space)

In the first pair, our reward function could include measures of how closely the robot followed the protocol and whether it dropped equipment or created hazards such as spills. The second pair we discussed in Section 2. For third pair, we would have to construct a reward function composed of the values enumerated in Section 2. This is a daunting task to try and find a single function that incorporates each of the values. And one that goes against the advice of Kuhn who said that such a comparison would be exceedingly difficult and subjective.

In contrast to other applications of reinforcement learning, for example autonomous vehicles or chess engines, the goal of a robot scientist is not in and of itself to take action. Actions taken by the agents that compose a robot scientist are in service of the broader aim of generating new scientific knowledge.

Typically in reinforcement learning, the action to feedback time is short and the evaluation of the reward function is cheap. This is not generally the case for robot scientists, as physical experiments have high cost, and usually take a significant amount of time (hours or days).

Consequently, techniques within reinforcement learning are less likely to be applicable to the scientific discovery aspect of the robot scientist. We conclude that reinforcement learning is unsatisfactory as a paradigm around which to design scientific discovery algorithms for a robot scientist. (It may well be that reinforcement learning algorithms can be of great use in optimising the laboratory automation, where the goal is to take actions in an optimal way.)

### 3.2 Supervised learning is a more appropriate paradigm

The crucial process to which we are applying machine learning is that of model improvement: given a model of an object system, how we make changes to the model such that it is more faithful to the object system. According to the scientific values presented in Section 2, there are numerous ways to evaluate this. However, in [30] "accuracy" was consistently ranked second by scientists in order of preference, only beaten by the "internal consistency" of a theory.

We argue that it is appropriate to consider scientific discovery as a supervised learning problem, both in observational and controlled experimentation. In the case of observational experiments, Medawar's "Baconian" experiments [21], the input-output pairs will be a partition of the overall observational data. Equally, in controlled experiments, Medawar's "Galilean experiments", our input-output pairs will be the experimental factors and the empirical data. We can exploit various mechanisms from the field of supervised learning to obtain theories which align with the scientific values stated in Section 2, with accuracy captured in a relevant loss function. This aligns with how previous robot scientists have operated, which we cover in more detail in Section 4.

### 3.3 Active learning integrates agency into supervised learning

While supervised learning is the appropriate paradigm for scientific discovery, when designing robot scientists we need to integrate agency. Active learning is a specific category of supervised (or semi-supervised) learning where the learning agent chooses the next data to inform the learning. This selection policy is not the focus of the learning, and there is no requirement that the selection policy be learned. This differs from reinforcement learning, where the objective is to learn a good policy. (We could perhaps use reinforcement learning to design this agent's policy, or alternatively use some pre-determined policy, or a combination of the

two approaches.) Active learning approaches select a point in the input space, usually because the model has high uncertainty around that point. Uncertainty can arise from poor exploration of, or shallow exploitation in, the neighbourhood of that point. The agent then requests a corresponding output from an oracle. In many applications of active learning this means choosing an input-output pair from a large, extant dataset. In scientific discovery, this could be the case (querying a database of previous experimental data), or in the case of a robot scientist the oracle in question can be the experimental platform. We see that active learning shares an analogy to the scientific method, and therefore is an appropriate paradigm to choose for the design and analysis of robot scientists.

### 3.4 Foundation models and their use in scientific discovery

Recent developments in machine learning, driven by industrial applications and enabled by new technologies and vast amounts of data, have resulted in widespread adoption of foundation models. Foundation models are defined in [4] as

> any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.

Within the scope of this definition, most current manifestations of foundation models are large language models (LLMs)—e.g. BERT [8], GPT-3 [5]—or large multi-modal models (LMMs)—e.g. GPT-4 [23]. However, under this definition, we could also include models such as AlphaFold [14] or Evo [22], transformer-based models trained on protein structure data and genomic data respectively.

Foundation models show promise in scientific discovery applications. Besides AlphaFold there is Coscientist, a system built around GPT-4 that could autonomously design, plan and execute experiments in chemistry[3]. Coscientist exploited the general purpose nature of the foundation model to combine information from various sources and to execute code and instructions on machines to achieve its goal. In many applications there is some concern for so-called "hallucinations" of LLMs and LMMs, i.e. claims made by the model with little or no justification or evidence. This is not a problem for a robot scientist provided the use it to generate hypotheses, as the resultant hypothesis will be tested via experiment. Hallucinations could cause problems in the experiment design phase; Coscientist mitigated the impact of hallucinations by grounding the LLM with database search, and ultimately evaluated the system's performance using explicit criteria rather than use the LLM.

Hallucinations could also cause problems if the foundation model is applied to the evaluation and assessment of competing theories. It is a distinct possibility that a robot scientist might justify a theory choice based on fabricated data or through faulty logical or statistical inference. And these models' black box nature means we cannot interrogate them about their reasoning. The best we can do currently is prompt the model for a post-hoc rationalisation of its reasoning, and it is not at all clear that this is of equal value [24].

These properties of LLMs and LMMs are a problem if we want to use them for "closed-loop" discovery in a robot scientist. And we also want scientific models to be interpretable and usable by other scientists, displaying the values of fruitfulness and simplicity. Foundation models often have hundreds of millions of parameters, and methods for interpreting their internal reasoning require further research before these models can be considered broadly suitable for automated discovery.

From a practical and economic perspective, LLMs and LMMs frequently require huge resources throughout the lifecycle of their development. Much focus is rightly directed to the immense electricity demands during training, but further resources are needed during research and development, data collection and storage, the construction and commissioning of hardware, and in the implementation and maintenance of LLMs and LMMs. [13].

Because of these economic demands, foundation models are often developed by large private enterprises rather than public science bodies or universities. This introduces risks to any scientific project dependent on these foundation models. Code can be closed-source, the training data and regimes are often held as trade secrets, and the models are provisioned on third-party hardware. Each project must weigh these risks against the clear benefits of using foundation models in scientific discovery work.

Having covered some important areas of theory behind computational scientific discovery and robot scientists in a domain-agnostic manner, we proceed to examine applications in a specific domain, biology.

## 4 Biological systems are good target for scientific discovery automation

Of the scientific challenges this 21st century, understanding the biology of eukaryotic organisms ranks among the most consequential. And despite fantastic advances during the 20th century of our understanding of the fundamental components and processes in biology, and in the application of this knowledge to medicine, engineering, agriculture, etc., we are still some way off an accurate predictive model of the physiology of one organism, let alone a system of broadly applicable theories, such as those we have developed 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.

The cellular physiology of eukaryotes is a complex system, in the spirit of the definition given by Simon [33] that

> given the properties of the parts [of the system] and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.

A reductionist approach to biology (breaking down a system and studying its components) has resulted in great advances in our understanding of the fundamental "parts and laws", for example the discovery of the double-helix structure of DNA molecules, or that all known proteins are composed from the same set of 21 amino acids via translation from RNA.

To understand complex systems it is necessary yet wholly insufficient to take a reductionist approach. Systems biology is an integrationist approach to studying biological systems. That is to say that we aim to understand how the parts and their interaction lead to the resultant behaviour of the system, be that system a cell, an organ or an entire organism.

By way of two examples of robot scientists, Adam and Eve, that were applied to the domain of biology, we will show how why this is a suitable domain for automated scientific discovery, and the importance of parsimony.

### 4.1 Superhuman logic and probabilistic reasoning

The first robot scientist was Adam [16], and was the first machine to autonomously discover new scientific knowledge. Adam was designed to cultivate bacteria and yeast in batch under varying conditions and measure phenotype, in this case the growth rates of the cultures over time. Adam used logic programming to analyse the theory of *S. cerevisiae* to identify hypotheses, which it then evaluated using quantitative analysis of the growth data. Adam hypothesised that three genes encoded for the enzyme 2-aminoadipate:2-oxoglutarate aminotransferase (2A2OA), previously an orphan enzyme. Even the reduced system that Adam studied resulted in a vast logical theory. To form hypotheses at scale in the manner that Adam did would be beyond human capabilities. Biology is a good domain for robot scientists to reason over because of the very large number of facts and entities involved.

Eve was a robot scientist designed to automate aspects of early stage drug development [34]. Eve had several modes of operation, but we focus here on the learning of quantitative structure activity relationships (QSARs). As the name suggests, they take as input the structure of a compound and predict the activity on the assay for a particular disease. Eve used a least-squares regression to learn the QSARs, which in turn were used to guide synthesis of new compounds, and further refine the QSAR. These steps are all dependent on large-scale statistical inference. Part of the reason why we need statistical inference is that the data collected from biological systems is often noisy. There is a degree of stochasticity to biological processes that are hard for humans to understand intuitively, but that machines are apt at modelling.

### 4.2 Applying parsimony

Adam was able to achieve autonomous discovery in part because of the parsimony of the logical theory used to generate hypotheses and evaluate them. There were many more facts and relations that could have been included in the theory, but by omitting these and building a theory that was focused on the scientific discovery task the theory was tractable and resulted only in hypotheses that were testable by the experimental apparatus available to the robot.

## 5 Case study: Genesis and LGEM$^+$

At Chalmers University in Sweden we are building a next generation robot scientist "Genesis". Our goal is to demonstrate that the robot scientist Genesis can investigate an important area of science a thousand times more efficiently (in terms of cost and money) than human scientists.

This is an extreme challenge for AI as the number of experiments to plan and coordinate is several orders of magnitude more than the previous case studies. Achieving this goal will involve advances in automated hypothesis formation (how best to utilise background biological knowledge and models in ML, etc.), automated experiment generation (how best to optimise gain of information with cost and time constraints), laboratory robotic control, and scientific data analysis.

The scientific discovery goal of Genesis is

> to develop a systems biology model of *Saccharomyces cerevisiae*, that is both more detailed and more accurate at predicting experimental results than any in existence.

The foundation of Genesis is a micro-fluidic system with 1000 computer-controlled micro-bioreactors (co-developed in Vanderbilt University, USA). Achieving this will be a step-change in laboratory automation as most biological labs have fewer than 10 chemostats. These micro-bioreactors are being integrated with ion-flow mass-spectroscopy (to measure metabolites at speed) and RNA-seq (to measure RNA expression levels).

To design the experiments that the robot scientist conducts, and to create and improve on the model of *S. cerevisiae*, we designed a modelling framework, which we present briefly now.

### 5.1 LGEM$^+$: a first-order logic model

The task of scientific discovery is described by Langley [20] in generic terms that given: (a) scientific data; (b) prior knowledge about the domain; and (c) a space of candidate categories, theories, laws, or models, we seek candidates that describe or explain the data.

In Genesis' domain, the scientific data are in the form of controlled experiments using *S. cerevisiae*, and empirical data arising thereof. There are many types of empirical data one could collect from such experiments. In our discovery application we focus on growth profiling and gene expression data.

Prior knowledge on *S. cerevisiae* is well-curated in genome-scale metabolic network models (GEMs). These are community developed models that follow a controlled vocabulary, so form a rich prior for automated scientific discovery. We chose to express the mechanisms of the biochemical pathways using first-order logic (FOL), an approach first proposed in 2001 [27]. We use a FOL structure that is grounded in the controlled vocabulary of the GEMs to express our knowledge about how entities are known to interact, for example that each reaction has reactants, products, and possible an enzyme annotation. We call this framework LGEM$^+$; below is a brief description to illustrate the case and a more detailed explanation of the methods is in [11].

LGEM$^+$ has five predicates: met\2, gn\1, pro\1, enz\1, and rxn\1. These are given specific semantic meaning, which is shown in Table 1. Here a cellular "compartment" refers to a component of the cellular anatomy, e.g. mitochondrion, nucleus or cytoplasm. There are seven types of clause that we included that encode the implications needed to describe phenomena such as reaction activity and gene expression. More detail on the specification of the logical theories is given in [11].

Finally, the space of candidate theories is those able to be constructed from the predicate symbols and constants relating to the cellular compartments, genes, compounds, reactions, and enzymes that could be present in yeast.

**Table 1.** Predicates used in the logical theory of yeast metabolism, LGEM$^+$.

| Predicate | Arguments | Natural language interpretation |
|---|---|---|
| met\2 | compound, compartment | "Compound X is present in cellular compartment Y." |
| gn\1 | gene identifier | "Gene X is expressed." |
| pro\1 | protein complex identifier | "Protein complex X is available (in every cellular compartment)." |
| enz\1 | enzyme category identifier | "Enzyme category X is available." |
| rxn\1 | reaction | "There is positive flux through reaction X." |

We use automated theorem provers for first-order logic (ATPs) to perform logical inference. In comparison to previous approaches using bespoke algorithmic methods, such as MENECO [25], using an ATP removes a large part of the burden of algorithm design and simulation, particularly when it comes to abductive inference. For the reasoning tasks we use the ATP iProver [18], which was chosen due to its performance and scalability as well as completeness for first-order theorem finding. We extended iProver to include abduction inference. ATPs

are designed to provide explanations, and have many tools to simplify theories to which they are applied. All these properties make ATPs a good choice for applications in scientific discovery.

From the logical theory we can deduce testable facts, e.g. production of compounds. This allows us to generate input-output pairs that can be tested against truth data. The first truth data that we tested the model predictions against was single gene essentiality data for *S. cerevisiae*. Essential genes are those genes whose removal from the genome leads to a loss of viability for the organism. We predicted single-gene essentiality for *S. cerevisiae* by providing: as input (the theory $T$) the yeast genotype (including the deletion), compounds that were present in the growth medium, and compounds assumed to be ubiquitous in the cell, along with the rest of the theory containing rules for activation of reactions and formation of enzymes; and deducing the output (the goal $G$) as a binary outcome of whether every compound assessed to be essential for growth was produced.

We compared these predicted pairs against empirical data for single-gene essentiality. The F1 score on the prediction task was 0.266, which was state-of-the-art for a qualitative method on that task, but still quite far away from the best quantitative models. In the case that a particular mutant is falsely predicted essential, this shows that there is room for improvement to the model, that the robot scientist needs to come up with hypotheses $H_i$ such that combined with the theory this hypothesis entails the goal($T \wedge H_i \vDash G$). The ATP achieves this through reverse consequence finding, using a rearrangement of the previous statement: $T \wedge \neg G \vdash \neg H_i$. It is also possible to steer iProver to find specific forms of H_i, though we did not do this in this case. We did filter the hypotheses found by LGEM$^+$ to just those that might be testable via an experiment. For more details of these processes, along with further examples of deductions and abduction of hypotheses can be found in [11].

The hypotheses that LGEM$^+$ generates may well be very close together in the experimental space. The interactions between the components of such a complex system mean that a small difference in input could have a large effect on the outcome of the experiment. This is one of the benefits of using a robot scientist to execute the experiments. As we mentioned in Section 2 we can achieve higher precision with robots and also log more data that could help explain systematic errors. Yeast cultivations typically stretch over a few days, so by using robots tiredness becomes less of an issue for the human scientists involved in the project. Genesis is also equipped with a microformulator allowing for fine adjustment of input media for the cultures, to a degree that would be onerous to replicate with handheld pipettes or even a traditional liquid handling robot.

### 5.2 Interpretation from ML paradigms

As we can see from the stated goal in the application domain, the criteria for success of the resulting model is to measure its predictive performance against

empirical results. Mapping between input and output pairs is non-trivial in a lot of cases. The information obtained through experimental measurements does not always align directly with what LGEM$^+$ can predict, particularly with relation to current metabolomics methods, However, despite the imperfect mapping of predicted to observed outputs, this follows the supervised learning paradigm.

With regards to active learning, we recall that exploration of the experimental space was due to uncertainty of the model in particular neighbourhoods. Due to the complexity of the system and the stochastic factors, certain phenomena might require hundreds of experiments which are very close together in the experimental space. This requires precision, and Genesis has been designed to provide this precision.

## 6 Conclusions and future directions

In this chapter, we covered the core concepts of the scientific method as they relate to robot scientists. We provided some tools for the analysis of robot scientist projects, namely to consider them through an active learning paradigm, and to map the values of scientific models onto techniques from machine learning. We discussed an example application domain, systems biology and a next-generation robot scientist, Genesis. We concluded by showing LGEM$^+$, a first order logic (FOL) model for Genesis.

One motivation for choosing FOL to model yeast metabolism was grounded in epistemic parsimony; FOL frameworks are easily extensible. There are many phenomena that could be relevant for experimentation, that LGEM$^+$ does not currently model. But we can update the logical framework without affecting the underlying infrastructure. We saw from our results on the single-gene essentiality prediction that there is a gap in explanation. We need to accede a more complex model, and the background knowledge suggests that we should incorporate new mechanisms such as gene regulation to capture some of the higher order behaviours of the system.

Our next steps will be working toward full integration of LGEM$^+$ with the Genesis hardware, relying on controlled vocabularies [26] to specify experiments and results so that we can map inputs to outputs for the machine learning algorithms. These are necessary steps so that we can employ LGEM$^+$ and Genesis for closed-loop experimentation.

**Acknowledgements**

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Alice Wallenberg Foundation. Funding was also provided by the Chalmers AI Research Centre and the UK Engineering and Physical Sciences Research Council (EPSRC) grant nos: EP/R022925/2 and EP/W004801/1, as well as the Swedish Research Council Formas (2020-01690).

## References


1. Allison, L.: Coding Ockham's Razor. Springer Science+Business Media, New York, NY (2018)
2. Barnard, G.A., Bayes, T.: Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances. Biometrika **45**(3/4), 293 (Dec 1958). https://doi.org/10.2307/2333180
3. Boiko, D.A., MacKnight, R., Kline, B., Gomes, G.: Autonomous chemical research with large language models. Nature **624**(7992), 570–578 (Dec 2023). https://doi.org/10.1038/s41586-023-06792-0
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al.: On the Opportunities and Risks of Foundation Models (Jul 2022). https://doi.org/10.48550/arXiv.2108.07258
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language Models are Few-Shot Learners (Jul 2020). https://doi.org/10.48550/arXiv.2005.14165
6. Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R.S., Cooper, A.I.: A mobile robotic chemist. Nature **583**(7815), 237–241 (Jul 2020). https://doi.org/10.1038/s41586-020-2442-2
7. Carnap, R.: An Introduction to the Philosophy of Science. Basic Books, New York (1974)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
9. Flach, P., Kakas, A., Ray, O.: Abduction, induction, and the logic of scientific knowledge development. In: Workshop on Abduction and Induction in AI and Scientific Modelling. pp. 21–23 (2006)
10. Gauch, Jr, H.G.: Scientific Method in Brief. Cambridge University Press, 1 edn. (Sep 2012). https://doi.org/10.1017/CBO9781139095082
11. Gower, A.H., Korovin, K., Brunnsåker, D., Tiukova, I.A., King, R.D.: LGEM+: A First-Order Logic Framework for Automated Improvement of Metabolic Network Models Through Abduction. In: Bifet, A., Lorena, A.C., Ribeiro, R.P., Gama, J., Abreu, P.H. (eds.) Discovery Science, vol. 14276, pp. 628–643. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-45275-8_42
12. Holland, I., Davies, J.A.: Automation in the Life Science Research Laboratory. Frontiers in Bioengineering and Biotechnology **8**, 571777 (Nov 2020). https://doi.org/10.3389/fbioe.2020.571777
13. Jiang, P., Sonne, C., Li, W., You, F., You, S.: Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. Engineering (Apr 2024). https://doi.org/10.1016/j.eng.2024.04.002
14. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (Aug 2021). https://doi.org/10.1038/s41586-021-03819-2
15. Kellis, M., Birren, B.W., Lander, E.S.: Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**(6983), 617–624 (Apr 2004). https://doi.org/10.1038/nature02424

16. King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E., Clare, A.: The automation of science. Science **324**(5923) (2009). https://doi.org/10.1126/science.1165620
17. King, R.D., Schuler Costa, V., Mellingwood, C., Soldatova, L.N.: Automating sciences: Philosophical and social dimensions. IEEE Technology and Society Magazine **37**(1) (2018). https://doi.org/10.1109/MTS.2018.2795097
18. Korovin, K.: iProver – An Instantiation-Based Theorem Prover for First-Order Logic (System Description). In: Armando, A., Baumgartner, P., Dowek, G. (eds.) Automated Reasoning, vol. 5195, pp. 292–298. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-71070-7_24
19. Kuhn, T.S.: The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press (1977). https://doi.org/10.7208/chicago/9780226217239.001.0001
20. Langley, P.: Integrated Systems for Computational Scientific Discovery. Proceedings of the AAAI Conference on Artificial Intelligence **38**(20), 22598–22606 (Mar 2024). https://doi.org/10.1609/aaai.v38i20.30269
21. Medawar, P.B.: Advice to a Young Scientist. The Alfred P. Sloan Foundation Series, Harper & Row, New York, 1st ed edn. (1979)
22. Nguyen, E., Poli, M., Durrant, M.G., Thomas, A.W., Kang, B., Sullivan, J., Ng, M.Y., Lewis, A., Patel, A., Lou, A., Ermon, S., Baccus, S.A., Hernandez-Boussard, T., Ré, C., Hsu, P.D., Hie, B.L.: Sequence modeling and design from molecular to genome scale with Evo (Feb 2024). https://doi.org/10.1101/2024.02.27.582234
23. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al.: GPT-4 Technical Report (Mar 2024). https://doi.org/10.48550/arXiv.2303.08774
24. Park, P.S., Schoenegger, P., Zhu, C.: Diminished diversity-of-thought in a standard large language model. Behavior Research Methods (Jan 2024). https://doi.org/10.3758/s13428-023-02307-x
25. Prigent, S., Frioux, C., Dittami, S.M., Thiele, S., Larhlimi, A., Collet, G., Gutknecht, F., Got, J., Eveillard, D., Bourdon, J., Plewniak, F., Tonon, T., Siegel, A.: Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. PLOS Computational Biology **13**(1), e1005276 (Jan 2017). https://doi.org/10.1371/journal.pcbi.1005276
26. Reder, G.K., Gower, A.H., Kronström, F., Halle, R., Mahamuni, V., Patel, A., Hayatnagarkar, H., Soldatova, L.N., King, R.D.: Genesis-DB: A database for autonomous laboratory systems. Bioinformatics Advances **3**(1), vbad102 (Jan 2023). https://doi.org/10.1093/bioadv/vbad102
27. Reiser, P.G.K., King, R.D., Muggleton, S.H., Bryant, C.H., Oliver, S.G., Kell, D.B.: Developing a logical model of yeast metabolism. Electronic Transactions in Artificial Intelligence **5**(B), 223–244 (2001)
28. Roccapriore, K.M., Dyck, O., Oxley, M.P., Ziatdinov, M., Kalinin, S.V.: Automated Experiment in 4D-STEM: Exploring Emergent Physics and Structural Behaviors. ACS Nano **16**(5), 7605–7614 (May 2022). https://doi.org/10.1021/acsnano.1c11118
29. Sarker, I.H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science **2**(3), 160 (Mar 2021). https://doi.org/10.1007/s42979-021-00592-x
30. Schindler, S.: Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think? Philosophy of Science **89**(3), 542–564 (Jul 2022). https://doi.org/10.1017/psa.2021.40

31. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal **27**(3), 379–423 (Jul 1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
32. Shi, Y., Prieto, P.L., Zepel, T., Grunert, S., Hein, J.E.: Automated Experimentation Powers Data Science in Chemistry. Accounts of Chemical Research **54**(3), 546–555 (Feb 2021). https://doi.org/10.1021/acs.accounts.0c00736
33. Simon, H.A.: The Architecture of Complexity. Proceedings of the American Philosophical Society **106**(6), 467–482 (1962)
34. Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L.N., De Grave, K., Ramon, J., de Clare, M., Sirawaraporn, W., Oliver, S.G., King, R.D.: Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. Journal of the Royal Society, Interface **12**(104), 20141289 (Mar 2015). https://doi.org/10.1098/rsif.2014.1289