The Use of AI-Robotic Systems for Scientific Discovery

Alexander H. Gower[1][0000−0002−8358−0842], Konstantin Korovin[2][0000−0002−0740−621X], Daniel Brunnsåker[1][0000−0002−5167−0536], Filip Kronström[1][0000−0002−3011−5541], Gabriel K. Reder[5][0000−0001−8918−0789], Ievgeniia A. Tiukova[1,3][0000−0002−0408−3515], Ronald S. Reiserer[4][0000−0002−3786−7893], John P. Wikswo[4][0000−0003−2790−1530], and Ross D. King[1,5][0000−0001−7208−4387]

[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
[4] Vanderbilt University, Nashville, TN, USA
{ron.reiserer,john.p.wikswo}@vanderbilt.edu
[5] University of Cambridge, Cambridge, United Kingdom
gr513@cam.ac.uk

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 relate these general guiding principles for designing robot scientists to examples from the domain of system biology: Adam, Eve, and Genesis. We present a case study of Genesis, a next-generation robot scientist designed for research in systems biology, comprising a micro-fluidic system with 1000 computer-controlled micro-bioreactors. We discuss LGEM$^+$, a logic-based model that is used in Genesis, and related them to discussed general principles.

Keywords: Robot scientist · Scientific discovery · Active learning · Laboratory robotics

1 Introduction

In the past two decades, the use of AI-robotic systems in scientific research has been demonstrated not only possible, but fruitful. Several projects—from

Adam [20] and Eve [40] to the Robot Chemist [8]—have proved the value of coupling AI software agents with experimental platforms to give them real-world agency. A commonly used term to describe such systems is *robot scientist*, defined in King et al. [20] as:

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

Robot scientists have the potential to broaden and deepen the capability of the scientific community, enabling high-throughput science and new modes of research, as well as helping to address problems with reproducibility, and human resource bottlenecks [41]. 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. This chapter aims to provide researchers with knowledge and tools helpful in analysing and designing robot scientists.

Section 2 first defines some of the core concepts in the philosophical ideas behind robot scientists, beginning with theories and model—the fundamental entities in scientific research. Then, methods of inference are introduced and how these should be considered when designing robot scientists. 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. The section concludes with a description of the different scientific values used 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 because biological systems are "complex systems" as described by Simon [38], 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 computational model developed with the aim of automated scientific discovery.

2 The Philosophy of Robot Scientists

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 [21].

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). This means they have characteristics or behaviour sufficiently similar to that of the object system to allow indirect study of the object system by studying its surrogate. 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 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, as defined in Flach [11]:

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

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. Some of these activities are domain-specific, but there are elements that are common among sciences, particularly inference (Section 2.1) and model evaluation (Sections 2.1 and 2.2).

The purpose of a robot scientist is to provide software and hardware that together can achieve each of these activities, and join them together in a "closed-loop" form of enquiry without human intervention [20]. 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 Elements of scientific method

There are three components of the scientific method considered here: logical inference, statistical inference, and parsimony. A history and detailed treatment of each of these components is covered in Gauch [12]; here we provide a brief introduction and highlight how these concepts can be used to analyse robot scientists and how they shape decisions during their design.

Logical inference Logics are mathematical languages that relate premises and conclusions. Logics are used in science to represent facts (observations of the real world) and laws (parts of theories or models) [9, 12].

There are three basic forms of logical inference: deduction, induction and abduction. With deduction, 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. Deduction is what enables deterministic simulations, and how we reason about a hypothesis in the scientific method.

Induction and abduction cannot provide guarantees of truth. Both seek explanations for a set of observable facts. Induction seeks laws that can explain a general case, whereas abduction seeks facts that can explain specific cases.

A simple example of induction is: having observed 10 yeast cultures that thrived in a sugar solution, and another 10 that died in plain water, one *induces the law* that yeast need sugar to thrive. While this law fits the observations, it turns out to be false in general; some yeasts, including *S. cerevisiae* can grow on ethanol.

An example of abduction is: given the law induced above, and an observation that a yeast culture is thriving in a liquid of unknown composition, one *abduces the fact* that there is sugar in the liquid.

Robot scientists require concretely defined induction or abduction problems to make contributions to scientific knowledge. In the case of the first robot scientist, Adam, the scientific problem was to identify links between *S. cerevisiae* genes and biochemical reactions [39]. By using Enzyme Commission (E.C.) numbers to refer to classes of enzymes which catalyzed a given reaction, the problem was transformed to one of abduction; Adam was to find facts of a given form.

> "A single hypothesis is the mapping of one *S. cerevisiae* ORF [gene] to one E.C. class - e.g. YER152C $\longrightarrow$ 2.6.1.39."

The discovery algorithm therefore had a well-defined output form. This was possible because the laws applying to E.C. classes were defined in Adam's logical model. In other cases, when the laws are unknown or not well enough defined, induction will be necessary, and to make these problems concrete one should either use techniques for induction that provide explanations—e.g. inductive logic programming—or ones that do not—e.g. learning neural network models.

Both induction and abduction deal in uncertainty, and therefore need probabilistic and statistical inference.

Statistical inference and probability Certain laws in science may only be expressed to a degree of certainty, and empirical data are subject to random effects. These situations require statements that can deal with uncertainty, or probability, of which two types are defined by Carnap [9]. *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 prefers to describe the two types of probability as being about events and beliefs respectively [12].

Statistics can be used to obtain a model by reasoning about observations. In the case of abduction, one is 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 are crucial to the scientific method because they allow us to form and improve upon models.

Statistical reasoning can be done when evaluating hypotheses against empirical data, or during the hypothesis generation stage. Adam, for example, used statistical measures of gene sequence similarity (PSI-BLAST, FASTA) to identify candidate genes in *S. cerevisiae* from genes in other organisms [39]. And when designing the most recent discovery framework for the robot scientist Eve, Brunnsåker et al. used inductive logic programming to identify candidate phenotypes [7].

Parsimony Parsimony is a term that refers to two different concepts. *Epistemological 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". *Ontological parsimony* is the idea that nature itself prefers simplicity. Both are both absolutely essential for science, and it is easy to just take parsimony as a common sense notion. However, a proper treatment of the motivations for adopting parsimony as a guiding principle is warranted when it comes to systems biology and the automation of scientific discovery.

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 the same computational model can be applied to two individual colonies of *S. cerevisiae*.

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 [19]. However all scientific argument must adopt some version of ontological parsimony, as it is the basis for generalisation of theories. 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 [9] as a two-step process: firstly, identify relevant factors for the phenomenon to be studied; then design experiments holding certain factors constant and varying over others. Determining relevant factors means stating that certain factors are irrelevant, i.e. that they are not expected 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 to test hypotheses relating to a restricted sub-

set of factors. Factorial design of experiments requires an appeal to ontological parsimony, which we can informally express as:

> empirical data from experiments of the same class are expected to exhibit only random variation, where the class of experiments depends only on the variable factors. (A)

This assumption allows the inference of empirical laws about the phenomenon.

In practice one cannot usually keep constant all the factors one would like, so the class of experiments 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. 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 Latin square designs 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 Statement (A). 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 [16], physics [33], and chemistry [37]. Secondly, that the validity of Statement (A) can be evaluated by recording more data about the execution of experiments than usually recorded when humans complete experiments. This is a "natural by-product" [20], as robots and software frequently have automatic logging capability. Finally, that those who design robot scientists encode the experimental protocol forces them 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): that the hypothesis that best explains the data is that which minimises the total information. The total information has two components, the *a priori* information contained in the hypothesis, and the information of the data given the hypothesis[6]. There is a trade-off between these two quantities. One can make a hypothesis so specific as to explain all the data, meaning the information content of the data given the

[6] Combine 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): $I(H \cap D) = I(H) + I(D|H)$.

hypothesis is zero. Or one can have no hypothesis at all. In practice, MML ensures that information is only added to a hypothesis if this information explains the data, which is the principle of Ockham's razor [2, 3, 36].

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

For robot scientists, epistemological parsimony should also be applied in the choice of model as it relates to the experimental hardware. More complex models that suggest hypotheses outside of set of possible experiments result in inaction. 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.

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, its design 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 [35] presents six virtues following Kuhn [23].

- **Internal consistency** is 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** is the absence of contradictions with other scientific theories. For example, that a model of the biochemistry of yeast is thermodynamically consistent.
- **Empirical accuracy**, otherwise referred to as predictive power, 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 the existence of 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, fruitfulness leads to models that generalise well to new applications or other domains. Fruitfulness also has one clear implication for closed-loop discovery.

Schindler [35] found that scientists do not agree on the relative importance of these characteristics, although there were some common views. And Kuhn [23] 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 humans, so we can accept this as part of science. The unique difficulty for robot scientists 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[1].

3 Scientific Discovery as Machine Learning

Satisfied that automating scientific discovery can be considered a machine learning problem, the question remains: how best should methods from the field of machine learning be applied to scientific discovery? This section covers how to view the components of scientific discovery, as presented in Section 2, as components of machine learning techniques, and investigates which of the machine learning paradigms are the most appropriate for application to robot scientists.

Common to all machine learning techniques is that given some data to train with, the goal is 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. Machine learning seeks to form a model using information from a relevant signal to assess performance against

a goal; this could mean defining a loss function for an optimiser, or in the case of inductive logic programming explaining positive examples and avoid inconsistencies with negative examples. To evaluate a candidate model, $\hat{h}$, predictions or consequences are obtained from the model (deduction) and then evaluated against the signal, often using a form of statistical reasoning. And finally, it is usually the case that there are many possible models that may have similar performance against the goal. 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 [34]; these paradigms are described below and in Figure 1. 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. The signal received is an intrinsic objective measure, for example data likelihood.
- **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.

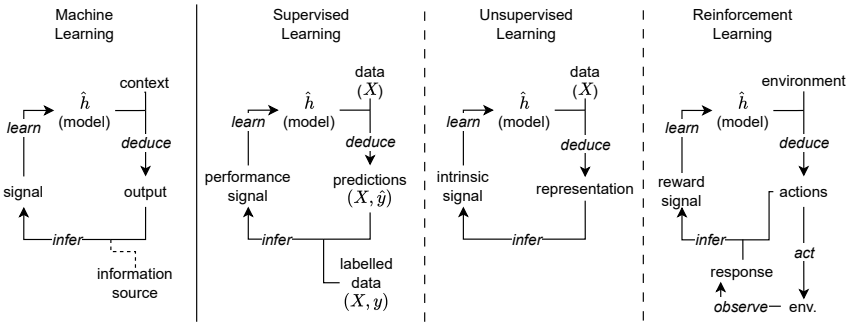

Fig. 1. Flowcharts representing high level design for: generic machine learning; supervised learning; unsupervised learning; and reinforcement learning. Unitalicised text represents inputs and outputs; italicised labels for connectors represent processes.

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.

For pairs (A, B) there are 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, the 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 useful paradigm

The crucial process machine learning is applied to in scientific discovery is that of model improvement: given a model of an object system, how to 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 Schindler [35] "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 useful 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 [26], the input-output pairs will be a partition of the overall observational data. Equally, in controlled experiments, Medawar's "Galilean experiments", the input-output pairs will be the experimental factors and the empirical data. When designing robot scientists, various mechanisms from the field of supervised learning can be exploited 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 Semi-supervised learning

Unsupervised learning techniques, such as embedding and clustering, are used in scientific discovery applications during data representation. For example, unsupervised techniques have been used to cluster molecular dynamics data in materials science [24], and transcriptomic data in biology [17]. In each of these cases, human scientists analysed the output of the unsupervised learning to draw conclusions.

As discussed above, a supervised learning design allows robot scientists to close the discovery loop and not require human scientists to analyse results. However, semi-supervised learning—integrating unsupervised techniques into supervised learning—allows for contextual information and the structure of theories to be exploited in the discovery task, as well as addressing the issue that the experimental space is sparsely annotated.

Semi-supervised techniques can exploit the background knowledge that is available to the robot scientist to make better predictions and learn more efficiently. Gleaves et al. [13] used a semi-supervised framework to improve synthesisability models in materials science, giving evidence that a semi-supervised approach to learning could provide the most promise for scientific discovery applications.

3.4 Active learning integrates agency into supervised learning

While supervised learning is the appropriate paradigm for scientific discovery, the design of a robot scientist must integrate agency. Active learning is a specific category of supervised learning where the learning agent chooses the next data point (input) for which a label (output) has not been observed. 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. (Reinforcement learning could be used 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 that has not been observed. Points are selected, usually either because the model has high uncertainty around that point, or to increase the diversity of the dataset. Uncertainty can arise from poor exploration of, or shallow exploitation in, the neighbourhood of that point. The agent then requests a corresponding output. In many applications of active learning this means asking for a human or expert annotation. In scientific discovery, this could be the case, or in the case of a robot scientist it can use its experimental platform to perform an experiment to get output data. We see that active learning shares an analogy to the scientific method, and therefore is an appropriate and useful paradigm to choose for the design and analysis of robot scientists. Both Adam and Eve used active learning by searching the hypothesis space and executing experiments to then improve the scientific model. This is covered in more detail in Section 4.1.

3.5 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 [5] 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 [10], GPT-3 [6]—or large multi-modal models (LMMs)—e.g. GPT-4 [28]. However, this definition also encompasses models such as Evo [27] or ESM3 [15], transformer-based models trained on genomic data and protein 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 [4]. 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 they 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 [29].

These properties of LLMs and LMMs are a problem for use in "closed-loop" discovery in a robot scientist. Scientific models should 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 life cycle 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 [18].

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 a Good Target for Scientific Discovery Automation

Of the scientific challenges in this 21st century, understanding the biology of eukaryotic organisms ranks among the most consequential. 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 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. 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 [38] 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. The aim is 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 why this is a suitable domain for automated scientific discovery.

4.1 Discovery through superhuman logic and probabilistic reasoning

The first robot scientist was Adam [20], 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 [40]. 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, partly because data collected from biological systems are 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.

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 design (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 *S. cerevisiae*, that is both more detailed and more accurate at predicting experimental results than any in existence. An outline of Genesis' discovery process is shown in Figure 2.

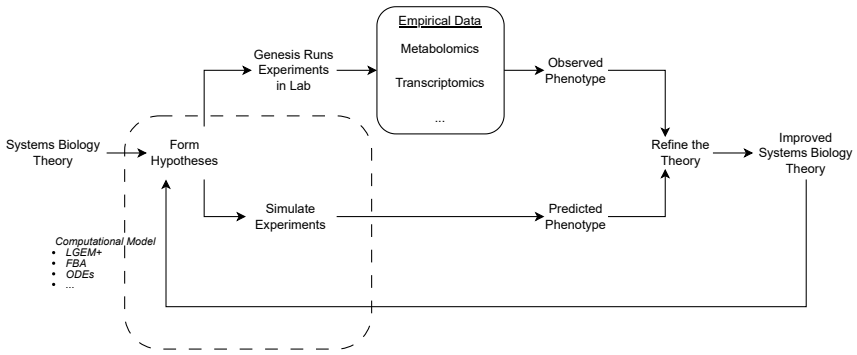

Fig. 2. Flowcharts representing the scientific discovery process of Genesis. The robot scientist starts with a systems biology theory, contructed from community knowledge. After using a computational model (e.g. LGEM$^+$) to form hypotheses, Genesis will design and run lab experiments using its hardware, and will take measurements of phenotype using automated procedures, e.g. for metabolomics. Simulated phenotype will be compared against the observed phenotype to generate information used to refine the theory, and the cycle will begin again with the improved theory.

The foundation of Genesis is a micro-fluidic system with 1000 computer-controlled micro-bioreactors (or chemostats) co-developed in Vanderbilt University. 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 next.

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

The task of scientific discovery is described by Langley [25] 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, scientists seek the candidates that describe or explain the data.

In Genesis' domain, the scientific data are in the form of controlled experiments using *S. cerevisiae* and resultant empirical data. There are many types of empirical data one could collect from such experiments. In our discovery application we focus on metabolomics 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 [32]. We use a FOL structure that is grounded in the controlled vocabulary of the GEMs to express knowledge about how entities are known to interact, for example that each reaction has reactants, products, and possibly 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 [14].

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 [14].

Finally, the space of candidate theories is those able to be constructed from the predicate symbols and constants relating to the cellular compartments, genes, metabolites, 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	metabolite, 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. ATPs are software that can automatically prove or disprove logical statements, by applying logical inference to a set of axioms. In comparison to previous approaches using bespoke algorithmic methods, such as MENECO [30], 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 [22], 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, the ATP can deduce testable facts, e.g. production of metabolites. This allows us to generate input-output pairs that can be tested against truth data. The first truth data the model predictions were tested against were 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. Single-gene essentiality was predicted for *S. cerevisiae* by providing: as input (the theory T) the yeast genotype (including the deletion), metabolites that were present in the growth medium, and metabolites 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 metabolite assessed to be essential for growth was produced.

Predictions were compared 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 ($T \wedge H_i \models G$). 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 filter the hypotheses found by LGEM$^+$ to just those that might be testable via an experiment, and used flux balance analysis to select plausible candidates. Hypotheses that correct errors with minimal effect on the current model (produce as few as possible new metabolites) are selected. This is in accordance with the principle of parsimony as discussed in Section 2.1. For more details of these processes, along with further examples of deductions and abduction of hypotheses can be found in [14].

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.

Models like LGEM$^+$ provide plausible hypothesis candidates for testing and feeding back to the discovery loop, which is aligned with the active learning paradigm and targeted exploration. When exploring the experimental space, due to the complexity of the system and the stochastic factors, certain phenomena might require many experiments that are very close together in the experimental space. This requires precision, and Genesis has been designed to provide this precision. The other selection criterion was to increase diversity in the dataset. A general purpose microformulator and modular design were deliberately chosen to increase options for experimentation.

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 the logical framework can be updated without affecting the underlying infrastructure. We saw from the results on the single-gene essentiality prediction that there is a gap in explanation. It is necessary to 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 the computational models with the Genesis hardware, relying on controlled vocabularies [31] 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 computational models like LGEM$^+$ with Genesis for closed-loop experimentation.

Acknowledgments. 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. Akhound-Sadegh, T., Perreault-Levasseur, L., Brandstetter, J., Welling, M., Ravanbakhsh, S.: Lie Point Symmetry and Physics-Informed Networks. Advances in Neural Information Processing Systems **36**, 42468–42481 (Dec 2023)
2. Allison, L.: Coding Ockham's Razor. Springer Science+Business Media, New York, NY (2018)
3. 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
4. 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
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models (Jul 2022). https://doi.org/10.48550/arXiv.2108.07258
6. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner,

C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (Jul 2020). https://doi.org/10.48550/arXiv.2005.14165
7. Brunnsåker, D., Gower, A.H., Naval, P., Bjurström, E.Y., Kronström, F., Tiukova, I.A., King, R.D.: Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology (Aug 2025). https://doi.org/10.1101/2025.06.24.661378
8. 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
9. Carnap, R.: An Introduction to the Philosophy of Science. Basic Books, New York (1974)
10. 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
11. 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)
12. Gauch, Jr, H.G.: Scientific Method in Brief. Cambridge University Press, 1 edn. (Sep 2012). https://doi.org/10.1017/CBO9781139095082
13. Gleaves, D., Fu, N., Siriwardane, E.M.D., Zhao, Y., Hu, J.: Materials synthesizability and stability prediction using a semi-supervised teacher-student dual neural network. Digital Discovery **2**(2), 377–391 (Apr 2023). https://doi.org/10.1039/D2DD00098A
14. 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
15. Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R.S., Thomas, N., Khan, Y.A., Mishra, C., Kim, C., Bartie, L.J., Nemeth, M., Hsu, P.D., Sercu, T., Candido, S., Rives, A.: Simulating 500 million years of evolution with a language model. Science **387**(6736), 850–858 (Feb 2025). https://doi.org/10.1126/science.ads0018
16. 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
17. Hozumi, Y., Tanemura, K.A., Wei, G.W.: Preprocessing of Single Cell RNA Sequencing Data Using Correlated Clustering and Projection. Journal of Chemical Information and Modeling **64**(7), 2829–2838 (Apr 2024). https://doi.org/10.1021/acs.jcim.3c00674
18. 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
19. 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

20. 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
21. 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
22. 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
23. 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
24. Kývala, L., Montero de Hijes, P., Dellago, C.: Unsupervised identification of crystal defects from atomistic potential descriptors. npj Computational Materials **11**(1), 50 (Feb 2025). https://doi.org/10.1038/s41524-025-01544-2
25. 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
26. Medawar, P.B.: Advice to a Young Scientist. The Alfred P. Sloan Foundation Series, Harper & Row, New York, 1st ed edn. (1979)
27. 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
28. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa,

E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F.d.A.B., Petrov, M., Pinto, H.P.d.O., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C.J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (Mar 2024). https://doi.org/10.48550/arXiv.2303.08774

29. 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
30. 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
31. 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
32. 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)
33. 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
34. 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
35. 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
36. 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
37. 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

38. Simon, H.A.: The Architecture of Complexity. Proceedings of the American Philosophical Society **106**(6), 467–482 (1962)
39. Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M.N., Liakata, M., Markham, M., Rowland, J., Soldatova, L.N., Whelan, K.E., Young, M., King, R.D.: Towards Robot Scientists for autonomous scientific discovery. Automated Experimentation **2**(1) (2010). https://doi.org/10.1186/1759-4499-2-1
40. 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
41. Zenil, H., Tegnér, J., Abrahão, F.S., Lavin, A., Kumar, V., Frey, J.G., Weller, A., Soldatova, L., Bundy, A.R., Jennings, N.R., Takahashi, K., Hunter, L., Dzeroski, S., Briggs, A., Gregory, F.D., Gomes, C.P., Rowe, J., Evans, J., Kitano, H., King, R.: The Future of Fundamental Science Led by Generative Closed-Loop Artificial Intelligence (Aug 2023). https://doi.org/10.48550/arXiv.2307.07522