ARTIFICIAL INTELLIGENCE FOR SCIENCE: THE EASY AND HARD PROBLEMS

Ruairidh M. Battleday and Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University
{battleday, gershman}@g.harvard.edu

ABSTRACT

A suite of impressive scientific discoveries have been driven by recent advances in artificial intelligence (AI). These almost all result from training flexible algorithms to solve difficult optimization problems specified in advance by teams of domain scientists and engineers with access to large amounts of data. Although extremely useful, this kind of problem solving only corresponds to one part of science—the "easy problem." The other part of scientific research is coming up with the problem itself—the "hard problem." Solving the hard problem is beyond the capacities of current algorithms for scientific discovery because it requires continual conceptual revision based on poorly defined constraints. We can make progress on understanding how humans solve the hard problem by studying the cognitive science of scientists, and then use the results to design new computational agents that automatically infer and update their scientific paradigms.

Keywords Scientific discovery · Artificial intelligence · Cognitive science

The easy problem

Most work applying AI to science has focused on what might be called the "easy problem." This is a relative term, since the easy problem is actually quite hard. A scientist specifies a function that they want to optimize (e.g., a function that generates a protein's structure given its amino acid sequence). Included in the specification is the input for the function (e.g., the amino acid sequence), the output (e.g., the 3D structure), and a way to compare the function's output with the ground truth (e.g., the average 3D distance of an amino acid residue from where it should be). The scientist then finds or collects a dataset, usually very large, with examples of the ground truth; or, designs some other way of assessing the model's output (e.g., turbulence parameters in plasma flow). AI optimization tools can then be applied to this problem. So far, this kind of application has been highly successful, with new discoveries of tertiary protein structures, antibiotics, and nuclear fusion reactor designs (see [1] for a recent review).

What makes this problem "easy" is not the form of the solution (which may require a great deal of engineering work) but rather the form of the problem. It is clear from the beginning what needs to be optimized, and what kinds of tools can be brought to bear on this problem. The engineering breakthrough comes from building much better versions of these tools. In other words, the problem is relatively easy because it does not require any conceptual breakthroughs of the sort involved in the discovery of relativity theory, genetics, or the periodic table.

Are these conceptual breakthroughs just patterns that can be discovered with a sufficiently powerful pattern recognition system? In a sense yes, but before that can happen, something has to tell the pattern recognition system what kind of patterns are interesting, important, and useful. What problem is the pattern-recognition system designed to solve, and where does this come from?

The hard problem

The fundamental barrier to automating science is conceptual. Great scientists are not simply extraordinary optimizers of ordinary optimization problems. It is not like Einstein had a better function approximator in his brain than his peers did; or Mendeleev's brain had a better version of backprop. More commonly, great scientists are ordinary optimizers of extraordinary optimization problems. It is the formulation of the problem, not its solution, that is the truly hard problem: The hard problem is the "problem problem."

One might be tempted to relegate the hard problem to the fringes of "revolutionary science," which rarely erupt into mainstream scientific practice, whereas the easy problem occupies the focus of the "normal science" that scientists spend most of their time on [2]. However, normal science is not simply optimization. This is obvious to any first-year graduate student trying to figure out what to work on. Normal science isn't a catalog of optimization problems waiting to be solved by a queue of grad students. Their fundamental barrier is the same one facing AI scientists: It is the conceptual problem of formulating an optimization problem. This encompasses both major conceptual breakthroughs, like relativity theory, and the more modest ones achieved by graduate students on a regular basis, which nonetheless remain out of reach for existing AI systems.

The problem with optimization

Much of the classic work on AI science (mainly by Simon, Langley, and their collaborators [3, 4, 5], but also more recently by Schmidt & Lipson [6], Udrescu & Tegmark [7], and others [8]) focused on the easy problem. For Simon and Langley, this approach was premised on the psychological thesis that scientific cognition was essentially the same as regular problem solving, only applied to a different (and sometimes more challenging) set of problems. Consequently, they developed algorithms that emulated human problem solving, and applied these to scientific discovery.

Existing AI scientists have had some success at the easy problem. Simon, Langley, and others were able to solve a range of seminal science problems [9], including the (re-)discovery of oxygen with STAHLp [10]; more modern methods for automated physics have inferred many existing and novel laws, including classical and quantum problems with AI Feynman and non-linear dynamical systems with SINDy [7, 8]; and, discovery algorithms in biology have advanced our ability to solve many difficult problems, including AlphaFold2 for protein folding [11].

This success is analogous to the earliest use of computers, in which they were used to complete calculations too laborious for any human (such as in the Enigma cryptography project in World War II; or, to prove all edge cases of a complicated theorem [12]). Algorithms that solve the easy problems of science are useful, even essential, to progress. For example, there is an increasing discrepancy between the number of amino-acid sequences that are discovered in biology and the recovery of the 3D protein structures they correspond to using the experimental method.

While recognizing the importance of such algorithms, we should also recognize their limitations. Several decades ago, Chalmers, French, & Hofstadter (focusing on the models of Simon, Langley, and their collaborators) challenged the idea that this kind of optimization was a complete model of scientific discovery and investigation [13]. Systems like STAHLp are only able to solve scientific problems and make discoveries, they argued, because the modelers have *represented the inputs and outputs to the problem in hindsight*; only relevant data have been included and those data are already organized such that the proposed heuristics will be able to easily extract the right solution. In other words, they have been provided a representation of the scientific problem that already includes the basic primitives needed for the final theory, but skirt the central problem of representation itself: Where do the primitives come from, and how do we know if we have discovered the right ones?

Simon insisted (contra Popper [14]) that there *was* a logic of scientific discovery, but Simon's proposal was really a logic of scientific problem solving—how to sequentially search through hypotheses given a problem statement and primitive representations [3]. This is not discovery in the sense of problem creation. The latter involves representation learning in service of the problem, but also something deeper: Identification of the goal or objective function itself.

In machine learning terms, these systems might be extremely good at interpolation, and they may become better at extrapolation to new data, but they will never automatically generate or choose to investigate new scientific problems. This is because neither the inferential nor the learned components of the algorithm contain the knowledge

necessary to do so. Instead, that knowledge came from the team that specified the problem by choosing how to represent the inputs, outputs, and objective function.

Problem representation

The representation of input data involves two fundamental choices—which are the right primitive variables and which datapoints to include. In trying to emulate the investigative processes of scientists, it is important to consider the primitives they would have begun with. The representation chosen for the inputs cannot be too permissive—it cannot use concepts or data that scientists came up with in the course of solving the problem; nor too restrictive—it cannot exclude concepts or data that the would have originally affected the problem-solving process.

The output for a scientific problem comprises the scientific theory and any predictions it generates. This choice of representation determines which theories are considered "well-formed" and therefore valid solutions for the problem. The output representation can be defined either explicitly, in the form of a set of symbols and operations, or implicitly, through the space of operations that can be applied to the input variables. Modeling the full scientific process requires specifying a generative system for theories that has sufficient flexibility for conceptual change, which in turn might affect the space of theories that are considered well-formed.

The third component of a problem representation is the goal, expressed in the language of optimization as a loss function that assesses the adequacy of a solution compared to the "ground-truth." The choice of loss function corresponds closely to the way the modeler has chosen to represent the structure of the natural domain. For example, when providing deterministic physical models for cosmically short distances, loss functions based on Euclidean distance might be suitable; for classification models, some assessment of decision accuracy; for probabilistic models, a loss based on relative entropy. The choice of loss is also influenced by the cognitive biases of scientists themselves. A classic finding from cognitive science is that for perceptual stimuli with separable feature dimensions participants' generalizations are better captured with a Manhattan loss, in contrast to a Euclidean loss for stimuli with integral dimensions [15].

Solving the hard problem

In contemplating how to build AI systems that solve the hard problem, it is instructive to look at how human scientists do it. The high-level objective in science is clear: We would like to account for more data with our theories. At this level, human scientists break the hard problem down into several sub-problems:

- Domain specification. What are the relevant phenomena that need to be explained by a theory?
- Constraint specification. What kinds of constraints need to be imposed on a theory based on existing knowledge (both domain-specific and domain-general)?

Once the domain and constraints have been specified, we can define an optimization problem (theory search); hence, we have converted the hard problem into the easy problem. For most current AI scientists, the modeling team conducts domain specification in advance in the representation and selection of data, and constraint specification in the representational scheme for potential scientific theories (outputs) and the objective function that assesses them. However, it is uncommon for real scientists to do a single pass from hard to easy, because they often realize that the problem they are solving is the wrong one. This may happen for several reasons. One is the realization that a theory is internally inconsistent or paradoxical. Another is the realization that the theory may (with suitable modification) be able to explain a broader range of phenomena, prompting a respecification of the domain. Conversely, phenomena which were previously included in a domain may need to be excluded if no adequate unifying theory is found for all the phenomena. Respecification can also happen when new empirical phenomena are reported. In a related vein, constraint respecification can happen when domains are merged, split, expanded, or shrunk. The key point is that problem creation and problem solving are cyclically coupled in scientific practice.

In the following sections, we motivate the distinction between the easy and hard problems with three case studies from the birth of modern chemistry, physics, and molecular biology. For each case study, we summarize the elements of the problem, the historical setting, and modern computational systems that have tried to recapture some aspects of these discoveries. We will argue that none of these modern systems offers a complete solution to the hard problem.

Case study 1: The discovery of oxygen

In the 18th century, it had been observed that lead increased in weight when it was slowly heated (which today we call "oxidation," but at that time was called "calcination"). This was difficult to explain with contemporary chemical theories, because they posited that something *left* a metal when it was heated (a type of inflammable earth called "phlogiston"). In 1774, the English chemist Joseph Priestley collected and identified a particularly inflammable and respirable form of air following the thermal reduction of calx-of-mercury (mercury-oxide) [16]. The French chemist Antoine Lavoisier eventually called this air "oxygen," and posited that it went *into* the metal during calcination instead, causing the weight change [17]. Lavoisier's course of investigations were so successful that he has been credited as having started the Chemical Revolution and introduced the principled application of the conservation of mass into the quantitative sciences.

STAHLp

Rose and Langley proposed a computational model called STAHLp to account for the discovery of the role of oxygen in calcination reactions [10] (see Box 1).

Box 1: STAHLp

Beliefs

The inputs and outputs of STAHLp are two types of belief:

Componential model: "Substance X is COMPOSED OF {Y, Z})";
X = Y, Z

Reaction: "{W, X} REACT to produce {U, Z})";
W, X → U, Z

- - - - - - - - - -

Production rules

STAHLp applies a set of production rules to generate further beliefs:

Substitute:	**Reduce**:	**Infer-components**:
W, X → U, Z	W, Y, $\cancel{Z}$ → U, $\cancel{Z}$	U = W, Y
X = Y, Z		
∴ W, Y, Z → U, Z		

- - - - - - - - - -

Objective

The objective function STAHLp uses to assess the consistency of a theory:

$$\begin{cases} \text{Fail if } nil \in \text{production rules(beliefs)} \\ \text{Continue otherwise.} \end{cases}$$

- - - - - - - - - -

Belief revision rules

If an inconsistent belief is generated, STAHLp applies a different set of belief revision rules to the beliefs upstream of the problematic statement.

The input to STAHLp is a set of interconnected beliefs about 1) which substances are present before and after a particular reaction or 2) the chemical composition of each substance. These inputs are encoded using two types of variable: The functions (or programs) REACTS and COMPOSED OF, which operate on an unbounded space of discrete chemical names.

STAHLp's desired output is a coherent and consistent "theory"—a set of beliefs that entail the inputs and do not contradict each other. STAHLp uses a hard objective function to enforce this: The theory cannot contain any inconsistent equations containing "nil" (the empty set).

STAHLp solves this problem by applying a set of "production rules" to its beliefs at each step, generating further beliefs. If the system generates an inconsistent belief, STAHLp throws an error. At that point, a second set of "belief revision" heuristics is applied to try to identify the source of the inconsistency and correct it. After the lowest-cost correction is made, STAHLp applies its production rules to generate the updated theory entailed by
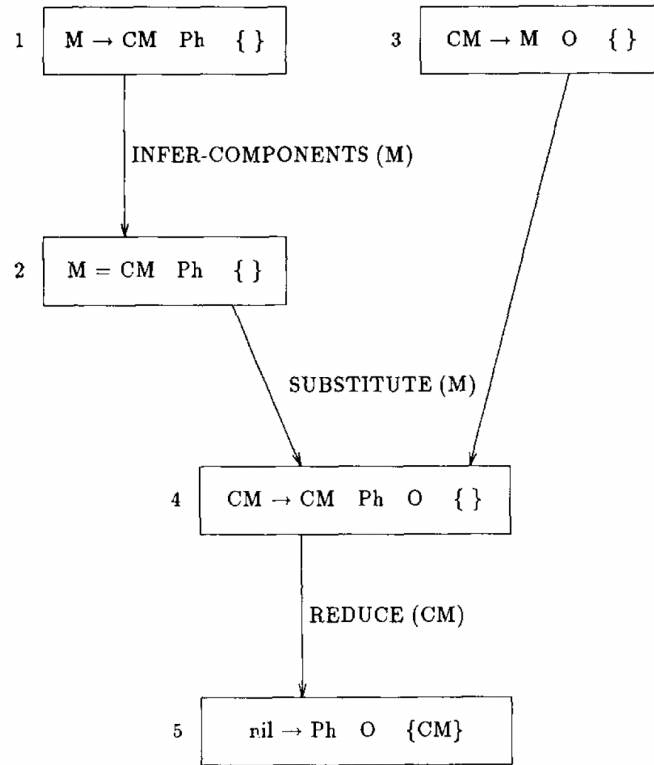
Figure 1: The discovery of oxygen by STAHLp. The input system of two beliefs about the composition of mercury (M) and calx-of-mercury (CM), reflecting the knowledge in the early 1770s. Ph refers to phlogiston, and O to the air observed by Priestley that Lavoisier eventually named oxygen. Reproduced from [10].

the new starting beliefs. The algorithm keeps running until no remaining beliefs are or an inconsistency has been detected; this is what Simon meant by scientific problem-solving as search through a hypothesis space [3].

Rose and Langley showed that for a particular pair of beliefs STAHLp "discovers" oxygen (see Figure 1). In particular, the first belief, inherited from Georg Ernst Stahl, states that mercury is composed of calx-of-mercury and phlogiston. The second, reflecting an empirical observation by Joseph Priestley, states that calx-of-mercury is composed of mercury and a colorless gas.

When STAHLp's production rules are applied to these observations, they produce an inconsistent belief—mercury-calx can be decomposed into itself, phlogiston, and oxygen (a circularity present in the two starting observations), which is then reduced to a statement containing nil on the left-hand side.

The update-belief rules are triggered, generating a set of "effect hypotheses" that "balance" the inconsistent belief (belief **4** in Figure 1):

- (EH1) CM [Ph O] → CM Ph O; left: missing Ph and O
- (EH2) CM [Ph] → CM Ph (O); left: missing Ph: right: extra O
- (EH3) CM [O] → CM (Ph) O; left: missing O; right: extra Ph
- (EH4) CM → CM (Ph O); right: extra Ph and O

By back-tracing where the left and right sides come from, STAHLp can generate "cause-hypotheses" about how the initials beliefs should be updated. The cause-hypothesis that affect the least downstream statements is chosen—in this case the addition of "oxygen" to the left hand side of belief **1**.

5

STAHLp: Analysis

It is hard to argue with the assumption that reactions and compositions of substances were the central concepts in chemistry—indeed, this was how Stahl himself defined its scope [18]. However, leading up to the chemical revolution, chemists had a different way of thinking about the internal structure of substances, in which observable substances arose from mixtures of the latent primitives of earth, water, and fire. A new name could not be added arbitrarily, and had to be placed within the existing ontological structure. Second, they would not have considered air—what we now call gas—to have chemical properties and enter into chemical combinations: The discrete name representation is too permissive, and "oxygen" is simply not a valid entry [19].

It also excludes relevant data. Most of the scientific work leading up to the discovery of oxygen was concerned with the sensory properties of substances,[1] where they came from,[2] and their weight. For example, there was actually a great deal of inconclusive or even negative evidence that metals apart from lead increased weight on calcination, detracting from the general statement that an air entered into metals. Similar arguments can be put forward for using a single function to represent a number of different reactions.

There is also the question of data selection. The creators of STAHLp include only two facts from the many heterogenous and often inconsistent observations and beliefs in 18th century chemistry. If they took into account others—for example, that when nitrous acid was poured on mercury colored vapors and fumes were given off—the model's conclusions may well have changed.

Finally, the objective function for STAHLp is based on the detection of nil statements. Once again, this is a retrospective assumption that relies on the application of the conservation of mass. By contrast, in the 18th century it was widely held that substances could dissipate away to nothing—from diamond [19], to phlogiston itself [20]. Lavoisier had to create the right conceptual framework needed to support the use of equations in his investigations before using them.

Historical perspective: Lavoisier and the discovery of oxygen

Instead of being bound by a fixed type structure, Lavoisier made a number of conceptual innovations that were closely informed by the ontological structure of chemical knowledge [21]. The first was that "air" (what we would now call gas) could be involved in chemical reactions at all. Robert Boyle and others had offered a physical interpretation of air and derived various laws. But, as surprising as it sounds today, in Continental Europe air was not thought to enter into chemical combinations—it was not a chemical type. By including air in the "definitions of chemistry," Lavoisier respecified the domain to include gross changes in air volume, and in turn explained the gross weight changes in calcination by the chemical fixation of air.

Next, Lavoisier broadened his scope to include all operations that fix or release air [22], with the aim of tracing the flow of air and water through different coupled reactions in order to infer the chemical composition of a more complex substance (like chalk). This richer set of data, including previous analyses by the Scottish chemist Joseph Black, led to the development of quantitative models based on equations. Prior to Lavoisier, chemists categorized and weighed solids and liquids before and after reactions, but did not routinely measure the air surrounding these materials. So, their "equations" seldom balanced, and the conservation of mass was used more as a post-hoc and abstract principle rather than a tool for quantitative purposes. Lavoisier developed the conceptual machinery to represent a reaction in terms of the total weight of materials at the start and end, and in doing so established the loss function to be optimized—the inference of a consistent and useful set of equations. In other words, he constructed the right representation of the problem. This placed emphasis on the use of a density constant to relate changes in air volume to changes in weight.

Discrepancies in subsequent experiments led Lavoisier to the conclusion that there must be different *subtypes* of air with different densities. This led to the development of new equipment to measure those densities, and ultimately the finding that the air of the atmosphere was in fact a composite of these subtypes, rather than an elemental root. He then showed that the reduction of calx-of-mercury with charcoal produced a different air (carbon monoxide and carbon dioxide) than the reduction of calx-of-mercury without charcoal, eventually calling the latter air "oxygen." Lavoisier explained the differences between these two reactions by positing an underlying, potentially infinite range of chemical primitives that could take the familiar three states of matter depending on how much of the "matter of fire" was coupled with them. This was the beginning of the main Chemical Revolution—actually more of an

---

[1]appearance, taste, feel and ductility and smell.

[2]the sea, the animal kingdom, mining, etc.

inversion, in that the things previously considered elemental (earth, water, air, fire) were now considered complex, whereas previously complex things like carbon were now considered elemental.

Case Study 2: The electromagnetic field

By the middle of the 19th century, Michael Faraday had published a set of discoveries and observations related to electromagnetic induction: A current could be generated in a conducting wire in the presence of a strong permanent magnet by moving the magnet or the wire. Faraday recorded the intensity of magnetic force surrounding magnets of various shapes, strengths, and number, as well as electrical circuits, arguing that the most useful representation for these data was in terms of *lines of magnetic force* [23]. He had speculated on what might be the cause of these patterns, but had been largely unsuccessful [24]. The Scottish physicist James Clerk Maxwell derived a brilliant and creative theoretical solution to this problem that provides the foundation of modern physics—the mathematical representation of the electromagnetic field.

No computational model has been proposed to emulate Maxwell's discovery. However, several influential models target the general setting of deriving physical laws from datasets of this sort [7, 8, 6]. Here we will focus on AI Feynman [7], an algorithm that uses *symbolic regression* to recover natural laws from physical data (see Box 2).

AI Feynman

The input for AI Feynman is a data table, comprising data samples (rows) of a dependent variable and several independent variables (columns) that the modeler has specified in advance for *each problem*. Variables take continuous values, correspond to measurements of the physical system, and are augmented with type information representing their fundamental physical units (meter, second, kilogram, kelvin, and volt.).

AI Feynman outputs predictions that match the dimensionality and type structure of the input, as well as a symbolic formula representing a theory of the observed system. The objective function uses a squared-error loss to assess predictions in the input space and a hard loss on whether its current solution is equivalent to the ground-truth expression.

AI Feynman cycles through a set of computational strategies premised on commonalities in the functional forms of solutions to known physical problems (Figure 2). The inputs and outputs to physical problems tend to have units, which justifies algebraic manipulations based on their types (dimensional analysis). Solutions, or parts thereof, often contain polynomial expressions, justifying polynomial fitting; they tend to be compositional, justifying search over symbolic expressions; they tend to be smooth, justifying approximation by a neural networks; they tend to exhibit symmetry and separability, allowing a reduction of variables after transformation by the neural network components. If nothing else works, a fixed set of transformations are applied to the variables, including the transcendental functions.

For example, the data in "mystery table 5" comprises samples from one dependent variable, F, and nine independent variables corresponding to the masses and 3D positions of two objects, and Newton's constant G. The algorithm runs through its pre-determined steps: Algebraic manipulations yield a reduced set of dimensionless variables; the application of a neural network component identifies translational symmetry; a good factorization is found; then polynomials are fit to two subsets of transformed variables. The end result of this process is an equation that accounts for the data below some error threshold, ϵ (see Figure 3).

AI Feynman: Analysis

Although the choice of input variables for AI Feynman might seem logical, they in fact correspond to quite an advanced stage of problem solving—when scientists have already constructed an idealized model for the system at hand.[3] For example, Newton had to *posit* the idea of a gravitational constant, expressed implicitly in terms of proportionality; and he had to posit that these were the *only* influential factors when explaining gravity—that action-at-a-distance was the correct framework to use, rather than the transmission of forces through an underlying medium. Similarly, Maxwell *invented* dimensional analysis to help solve difficult physics problems. But he did not always choose this representation—for electromagnetism, for example, he chose to think about dynamical properties of the aether.

---

[3]This is the process that Richard Feynman went through in his lectures when giving the historical background of the problem statement.

There is also the question of which datapoints are chosen. For mystery table 5, the data are not taken from systems far from the scientist or near large masses, where the behavior of light (its speed or deflection, respectively) needs to be taken into account. Recognizing and adjusting for these factors were essential parts of proving the theory and then taking it forward.

Then there is the representation of theories. The choice of symbolic expressions to represent "natural laws" *after* the domain has been specified is reasonably unproblematic—although Newton himself did not use explicit symbols like "G" or formulae for the relationship between the motions of cosmic bodies [25]. Symbolic expressions are constrained to be "well-formed," which requires that the modeling team ensures that each operation only runs on valid inputs and only produces valid outputs, defined by the initial primitives and the fixed type structure of the operations that run on them. Again, this scheme lacks the flexibility to capture the kinds of conceptual change that would have been necessary to derive the mature form of the problem. For example, the space of symbolic expressions might include primitives and operations that were discovered in the process of formulating the problem—analogous to providing the symbol i before deriving a general solution to the problem of polynomial root finding. AI Feynman does not include primitives for the differential calculus, but the closely related algorithm SINDy does [8]. In this context, taking derivatives with non-integer powers might be required to capture the diffusion patterns in some data [26], but would not be allowed by its predefined type structure.

The kinds of laws AI Feynman can derive are also limited by its processing steps. This is motivated by an analysis of common characteristics of physical laws—they contain variables with units, low-degree polynomial structure, compositionality, smoothness, symmetry, and separability [7]. But again, these constraints arose out of analysis of the existing laws of physics, and provide constraints that restrict the subsequent class of models in an inflexible manner.

Historical perspective: Maxwell and the electromagnetic field concept

Nancy Nersessian has given a thorough cognitive-historical analysis of Maxwell and the development of the electromagnetic field concept [27]. In order to make progress given the ill-defined and heterogenous state of electrical science, Maxwell restricted his scope to Faraday's data on electromagnetic induction and lines of force. In 1855, he gave a rigorous and analyzable form to Faraday's observations and theoretical postulations using a descriptive mathematical model based on continuum mechanics of stresses in an underlying medium [28]. From 1861-1864, he tackled the deeper problem of providing a dynamical model that would explain these data [29, 30]. He began with magnetic phenomena, and showed that the constraints provided by his descriptive analysis could be fit by a vortex model. From this model he could calculate the magnetic force at any point in the medium by carrying over the system of equations describing the mechanical force exerted and replacing mechanical variables with magnetic ones [31].

When he generalized this model to a medium composed of these vortices, however, he found the model unsatisfactory because of the friction caused by adjacent vortices. This brought to mind the idle wheels interposed between rotating machine gears, from which he introduced the idea of idle-wheel-particles to communicate between vortices. Idle-wheel particles provided a good way to model electrical current, so his next step was to include electromagnetic phenomena. But this required the relaxation of the model to allow the particles to translate in conductive medium, and to rotate without generating any friction. Using the new model, he could bring in a set of equations to represent electrical current as the flux density of these particles, driven by the circumferential velocity of the vortices [31]. Maxwell continued this process of domain relaxation and model building to include electrostatic phenomena and the polarization of light.

A striking feature of Maxwell's problem solving is how explicit he was about the scope of his theories and the utility of intermediate models. Selectively restricting the domain allowed him to identify which parameters or features of the intermediate model were essential, and an analysis of those features afforded selective expansion of the domain—a process Nersessian has called "generic abstraction" [27]. Like Lavoisier, Maxwell was guided in this process of abstraction by *ontological knowledge* about the structure of different physical and mathematical systems, which also helped him sequentially assemble and modify the mathematical expressions underlying the model. Perhaps these idealized models played a role in Lavoisier's early investigations, albeit in a simpler form involving crude movements of air and changes of weight. This process is not captured by systems like AI Feynman, which are given the problem variables from the mature idealized model, and lack the flexibility to alter their own conceptual systems.

Box 2: AI Feynman

Stepwise objective

$$\begin{cases} 1 \text{ if } ||\frac{(\hat{f}(x)-x)^2}{n}||^{\frac{1}{2}} < \epsilon_{\text{step}} \\ 0 \text{ otherwise.} \end{cases}$$

Final objective

$$\begin{cases} \text{Pass if SIMPLIFY}(\hat{f} - f) == 0 \\ \text{Fail otherwise.} \end{cases}$$

AI Feynman assesses whether the Euclidean distance between predictions and data is small enough after each step, and whether its symbolic expression is the same as the ground truth law.
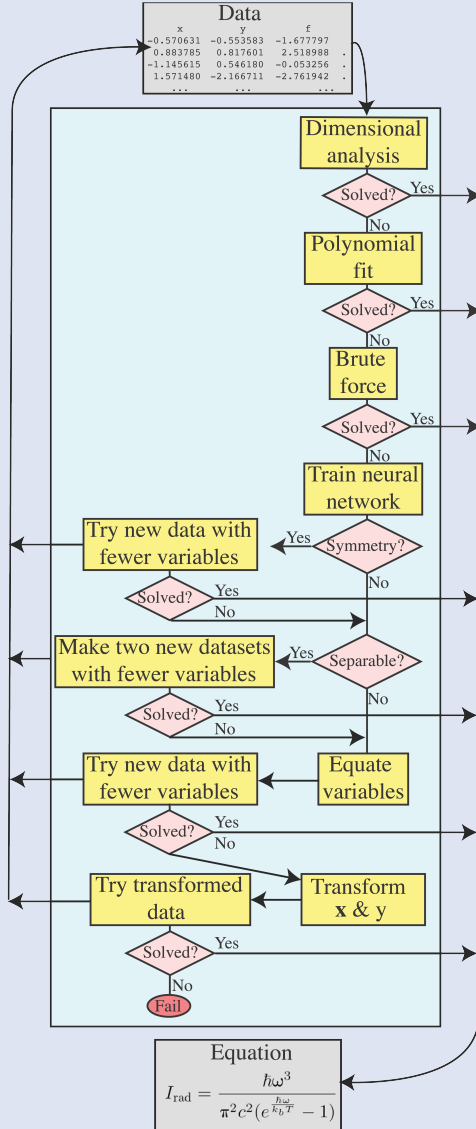

Figure 2: The steps that AI Feynman goes through when applied to solve a scientific problem, given in the form of a mystery table. Reproduced from [7].
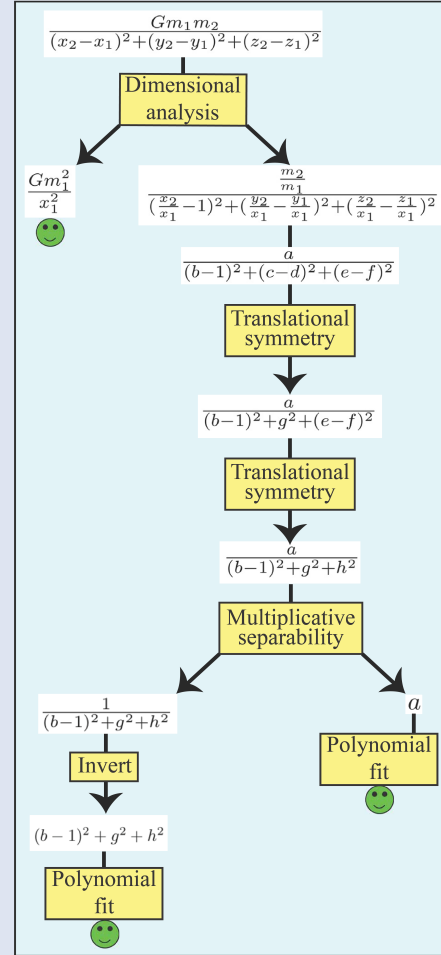
Figure 3: AI Feynman recovers the correct expression at the top of the diagram by applying its problem-solving steps. Reproduced from [7].

Case Study 3: Protein folding

Several major conceptual breakthroughs led to the "protein folding problem." The discovery that proteins are *linear* chains of amino acids goes back to the seminal sequencing of insulin by Frederick Sanger in 1951 [32], based on the isolation and recursive extraction of hydrolyzed protein fragments using various media and electrical currents. Evidence that the overall 3D structure of proteins was important for their function, rather than the identity of individual amino acids,[4] came from X-ray crystallography of oxygen-carrying proteins [33, 34], the structural effects of natural and artificial variation of amino acids [35], and catalytic-rate analyses with different cellular conditions, substrates, and inhibitors [36, 37].

The third development was the specification of the protein folding problem, primarily by Christian Anfinsen Jr. [38], who found that ribonuclease A would lose its enzymatic activity in artificial conditions and recover it when physiological conditions were re-established. This led to the "thermodynamic hypothesis" that the correctly folded protein occupied the minimum free-energy state in its natural cellular environment, and provided evidence against the competing hypothesis that proteins folded sequentially as they were synthesized.[5] The remaining step was to characterize the physical process by which the protein folded.

AlphaFold2

One of the most successful recent discovery algorithms is AlphaFold2 [11], which predicts the 3D structure of a protein given its 1D amino-acid sequence (see Box 3). When it was released, AlphaFold2 brought the average molecular deviation for a protein down from 0.3 to 0.1 nanometers, which was precise enough for biologists to make use of.

AlphaFold2's input is a multiple sequence alignment (MSA), which augments the protein of interest's 1D amino-acid sequence with additional rows containing similar amino-acid sequences from existing databases. If any of the MSA sequences have already had structures derived, 2D distograms of the pairwise distance between residues and a sequence of torsion angles between adjacent amino acid residues are added to the inputs.

AlphaFold2 outputs a set of atomic co-ordinates, a confidence score in each residue's position, torsion angles between adjacent amino-acid backbones, the 2D distogram between residues, and a prediction of any masked parts of the MSA. The objective function during training contains a loss term for each of these representations, with the most important components penalizing the 3D deviations of heavy atoms in the amino-acid chain. The loss function during "fine-tuning" contains all of these terms, plus two extra terms that penalize the final structure for violating physical constraints.

AlphaFold2 uses complex heuristics to solve this optimization problem, based on a great deal of biological and engineering knowledge. At a high level, the Evoformer module learns increasingly rich and abstract representations of the 1D primary structure and 2D distogram that the Structure module uses to build a 3D model of the protein. The network is trained end-to-end, meaning all operations are differentiable and the loss signal from the final 3D positions is back-propagated to inform the update of neural networks weights in all operations after the input.
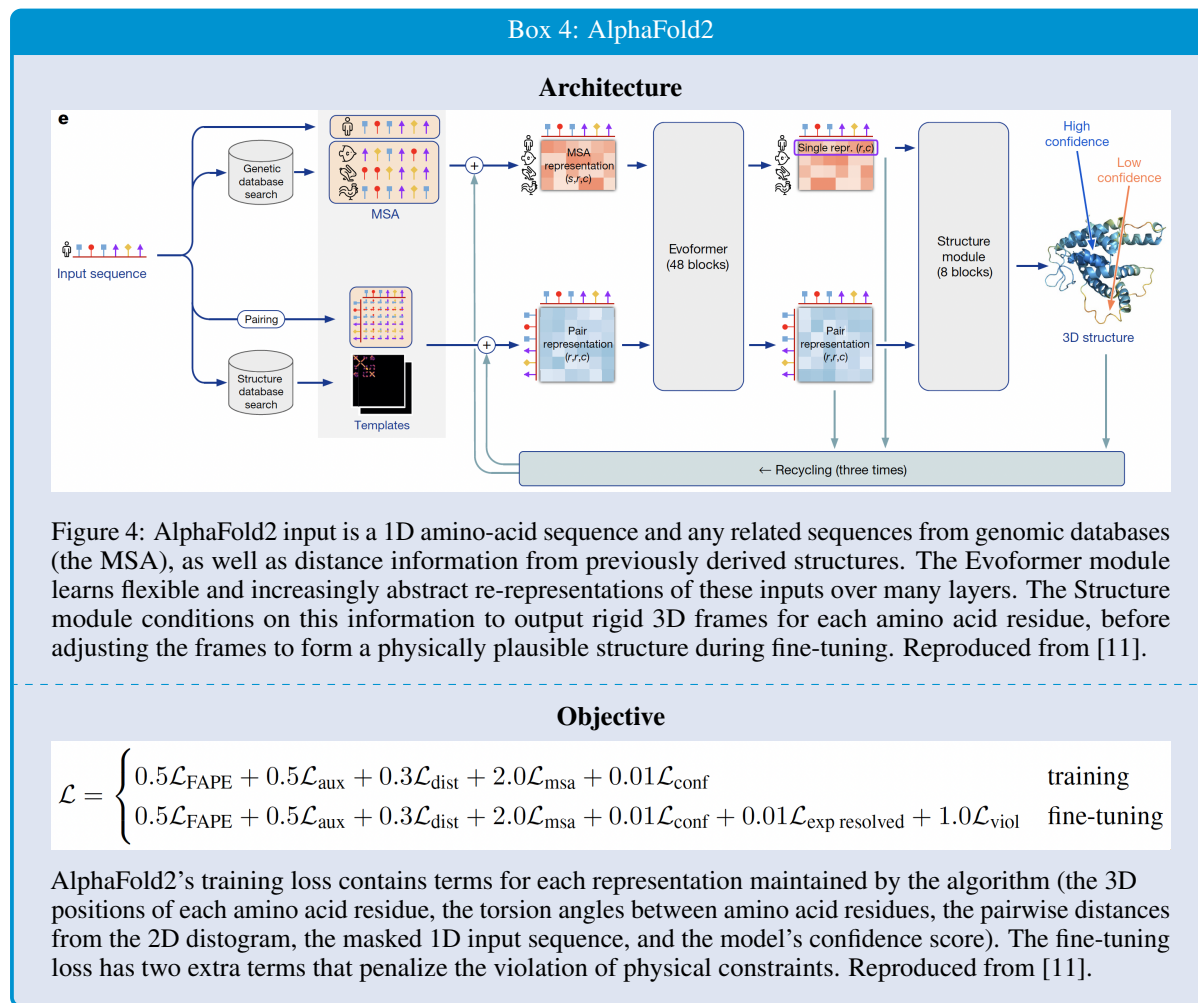
The main biological insight behind the Evoformer module is that information about the 3D protein structure can be derived by comparing its primary sequence with the sequences of similar proteins in different organisms. Some of these sequences might have had their structures discovered experimentally, which can be used directly in the 2D distance representation. But even when no related structure exists, significant covariation of residues in two different positions across multiple organisms is an indication that they are close in 3D space. These 3D dependencies might be quite far away in the 1D representation (the primary sequence), so the inductive bias of attention, which can model longer dependencies [39], is more suitable than other deep learning methods. When a particular structure is available, it biases the attention mechanism to learn similar representations for amino acids that are physically close together, and when it is not, the information flows the other way, with the covariance between amino acid positions used to infer the 2D distances.

The Structure module uses the Evoformer's final representations to iteratively move rigid frames representing each amino-acid residue as close as possible to their ground truth cognates. After residues have been aligned in the main

---

[4]The notable exception to this proposition is the identity of certain amino acids in the active site of enzymes.

[5]There were data that this theory did not apply to—for example, proteins that required enzymatic modification to renature; or assistance during folding by "chaperone" proteins.

training cycle, refinement steps add the amino acids' side chains and alter their positions during fine-tuning such that inter-atomic bonds take physically plausible values and no side-chains overlap.

Box 4: AlphaFold2

Architecture

Figure 4: AlphaFold2 input is a 1D amino-acid sequence and any related sequences from genomic databases (the MSA), as well as distance information from previously derived structures. The Evoformer module learns flexible and increasingly abstract re-representations of these inputs over many layers. The Structure module conditions on this information to output rigid 3D frames for each amino acid residue, before adjusting the frames to form a physically plausible structure during fine-tuning. Reproduced from [11].

Objective

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

AlphaFold2's training loss contains terms for each representation maintained by the algorithm (the 3D positions of each amino acid residue, the torsion angles between amino acid residues, the pairwise distances from the 2D distogram, the masked 1D input sequence, and the model's confidence score). The fine-tuning loss has two extra terms that penalize the violation of physical constraints. Reproduced from [11].

AlphaFold2: Analysis

AlphaFold2's success comes in large part from the engineering choice of problem statement. In particular, it does *not* solve the original "problem" of protein folding, the time-evolving movement of the polypeptide chain from 1D denatured to 3D functional state. The authors relax the physical requirements to define a new, related problem: The prediction of a final folded state given the 1D sequence.

This choice was motivated by an abundance of sequence data, which can be used for the new, but not the old, problem.[6] It was also based on a suite of models with the ability to use that information to solve the relaxed problem formulation—deep neural networks with attention mechanisms. The positive consequence of this choice is that some requirements of a solution to the original problem are met—we can predict the structure of hydrophilic proteins with lots of analogous evolutionary sequences well. The negative consequence is that we don't have a model of the folding dynamics which can make good predictions of the structures of orphan molecules like antibodies, lipophilic molecules with no experimentally derived homologous structures, or the effect of a new mutation or ion on the final folded structure.

The choice of inputs and outputs reflect the new problem. Evolutionary correlations have been used for some time to make arguments about folded structures and function [40], but do not obviously inform folding dynamics. And,

---

[6]The MSA can be generated using off-the-shelf tools from genomics, which now have very large sequence datasets.

the pair-wise distogram input has been chosen as a primitive because this is the form required by the algorithm that will be used upstream (attention). AlphaFold2's intermediate and output representations and transformations constitute a highly distributed theory. The most important engineering choice, in light of defining the new problem, is to partition training into a free optimization period over the residue gas representation, plays to the strengths of deep learning in iterative representation learning and stochastic gradient descent, followed by a fine-tuning stage where physical constraints are met.

Although the scale and complexity of natural biological systems warrants different types of theory and strategies of investigation, including decomposition and localization [41], there are also important commonalities—including the use of ontologies [42] and imagistic intermediate models [37, 40]. Conversely, Lavoisier and Maxwell were often tempted to re-specify constraints based on abundant data and a powerful method nearby.[7] We would like AI scientists that can likewise recognize when progress has been slow on a particular problem, but adjacent sources of data and powerful models promise fulfilment on an intersecting set of desiderata. We would also like them to recognize when and how to gather more useful data when there is a mismatch with the use case—as recent improvements on using AlphaFold2 to predict human structures have done.

Understanding the hard problem

The previous sections depict a recurring pattern: Much progress in applications of AI to science has been made, but only with the aid of humans specifying the problem formulation. Thus, these systems are essentially solving the easy problem, not the hard problem. What makes the hard problem so hard?

An important and elusive feature of problem specification is that it is not a data modeling problem. The selection of what to model and and what constraints to condition on are antecedent to any data modeling problem. It is also not reducible to a representation learning problem, in the sense of figuring out how raw sensory input maps to abstract representations. Of course, that problem also needs to be solved, but first the scientist needs to know what problems the representations are being used to solve.

Sociological, aesthetic, and utility considerations enter at the problem specification stage. Building an AI scientist is as much about shaping its tastes, style, and preferences as it is about endowing it with powerful problem-solving abilities. Again, a look at how we train human scientists is instructive: A good graduate advisor educates students about what problems matter, what phenomena are interesting, which explanations count, and so on. These considerations can't be brushed aside as subjective factors irrelevant to the purely technical problems facing AI systems; they are in fact constitutive of those technical problems. Without them, the technical problems would not exist.

A research program for attacking the hard problem should begin with the cognitive science of science [43], focusing on the understudied subjective, creative aspects discussed above and how they interact with the objective aspects of problem solving. However, this presents two immediate challenges. First, how can we gain the conceptual background necessary to understand scientists' innovations in a short enough time to iterate meaningful research? For most graduate students, arriving at the point where they can begin to generate meaningful and achievable problems within their field takes 2-5 years of dedicated higher education. Second, how can we gather enough results to make statistically robust arguments for any individual problem? There was, of course, only one Antoine Lavoisier.

Cognitive-historical analyses are one approach to deriving such insights [27, 42, 44]. In this methodology, modern cognitive theories are used to build hypotheses about how the scientists were thinking. Historical data can add content or temper these theories, and historical analyses and techniques can be used to make the retrospective analyses relatively unbiased, robust to historical contingencies, and generalizable to new contingencies. At the birth of a modern scientific field, the concepts and measurements are relatively undifferentiated, and can be acquired quickly. They are also, necessarily, edge-cases of creativity, where one or a group of scientists broke away from the normal tradition.

Spending time observing scientists' behaviors in modern operating laboratories is another way to increase the amount of data available for cognitive scientists to build theories about problem specification. For instance, the construction of intermediate *in-vitro* models as sources of analogy has been hypothesized to explain the success of scientific research practices in biochemistry [45]. Studies of scientific collaborations between people from

---

[7]Lavoisier's use of the Hales' apparatus and burning lens is a good example of this.

different fields or methodological backgrounds have emphasized the importance of visual aids and hand gestures in providing explanations [46].

The birth of the internet and online crowdsourcing platforms has allowed cognitive psychologists to scale up their studies to online experiments where the behaviors of very large numbers of computer-literate participants can be tested [47]. We can use our insights from cognitive-historical analyses and laboratory observations to design prospective tests of the key computational principles underlying the construction of problems and other aspects of discovery. Indeed, related studies already provide strong evidence that humans construct simplified mental representations to plan [48], but have not been extended to the less well-defined problem settings in science. Another rich set of problems come from tests of physical reasoning, in which previous laboratory-based work has identified iterative model-based revision of problem statements as a critical part of deriving a successful scientific solution [49, 50].

Towards scalable AI scientists that solve the hard problem

Once we understand what human scientists are doing with enough precision that we can formalize their activities, we can try to leverage these insights to build scalable AI scientists. At least initially, it is unlikely that these will be standalone systems, but rather more like research assistants or first-year grad students: Curious agents with some technical competence but in need of expert guidance. This guidance can come in the form of natural language instruction, reading curricula, and demonstrations. The growth of models beyond this requires the examination and emulation of the communal aspects of science and related cultural institutions. Lab meetings, conferences, and presentations and discussions are ultimately the place where judgements on the quality of a scientific problem are made.

The use of natural language processing for scientific discovery is at the heart of the recently proposed "AI Scientist" [51], which autonomously updates machine-learning (ML) code in order to generate scientific papers. In its inner loop the AI Scientist is given access to the training, testing, and visualization code for a simple ML model and dataset, along with several suggestions of innovative changes to the code and the overall objective of reducing the model's loss on held-out data. Its outer loop requires that it generate a range of ideas in natural language format, check their novelty using the internet, apply several ideas, write a ML paper for each of the ideas that ran successfully, review the paper, then update the paper. The proposed system comprises a carefully designed interface of language models, prompting schemes, a coding assistant, and templates for papers and conference guidelines. From the examples presented in [51], the innovative ideas that the AI Scientist generates are mostly decisions to split variables or processing pathways, add new model components or training metrics based on previously successful strategies in the literature, and combine any of the above that improve the final loss. A particularly impressive part of the work is the ability to implement these high-level conceptual changes in the code example, including producing useful visualizations.

Whether this system and its successors can produce radically innovative discoveries remains to be seen. Do such systems replicate human strategies such as ontologically guided constraint respecification, producing and modifying intermediate models, and re-specifying the problem based on knowledge of adjacent rich sources of data and available models? Natural language is certainly capable of capturing some aspects of the ontological structure of knowledge, and multimodal models should be able to create and maintain imagistic intermediate models of the scientific phenomenon.

On the other hand, many scientific developments, including those we have characterized above, come from a reflective consideration of either how to alter model constraints to capture anomalous data [52, 42], or where an alteration of model constraints affects the domain, borne out over a course of successive investigations [22, 27]. Whether current models that decouple in-context and weight-based learning can capture this type of reflective continual learning and selective conceptual respecification will require further investigation [53, 54]. For now, humans remain the only intelligent system capable of solving the hard problem. We still have much to learn about building AI scientists by studying ourselves.

Acknowledgments

References

[1] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[2] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962.

[3] Herbert A Simon. Does scientific discovery have a logic? *Philosophy of science*, 40(4):471–480, 1973.

[4] Herbert A Simon, Patrick W Langley, and Gary L Bradshaw. Scientific discovery as problem solving. *Synthese*, 47(1):1, 1981.

[5] Gary F Bradshaw, Patrick W Langley, and Herbert A Simon. Studying scientific discovery by computer simulation. *Science*, 222(4627):971–975, 1983.

[6] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[7] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

[8] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[9] David Klahr and Herbert A Simon. Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125(5):524, 1999.

[10] Donald Rose and Pat Langley. Chemical discovery as belief revision. *Machine Learning*, 1:423–452, 1986.

[11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[12] Thomas Tymoczko. The four-color problem and its philosophical significance. *The Journal of Philosophy*, 76(2):57–83, 1979.

[13] David J Chalmers, Robert M French, and Douglas R Hofstadter. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3):185–211, 1992.

[14] Karl Popper. *The logic of scientific discovery*. Hutchinson & Co, 1959.

[15] Roger N Shepard. Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In *The perception of structure: Essays in honor of Wendell R. Garner*, pages 53–71. American Psychological Association, 1991.

[16] Joseph Priestley. An account of further discoveries in air. *Philosophical Transactions*, 65:384–394, 1775.

[17] Antoine Lavoisier. *Elements of Chemistry in New Systematic Order, Containing All Modern Discoveries*. Edinburgh: William Creech, 1790.

[18] G.E. Stahl. *Fundamenta chymiae dogmaticae & experimentalis*. Nürnberg: Adelbulner für Endter, Germany, 1723.

[19] Henry Guerlac. *Lavoisier—the crucial year: the background and origin of his first experiments on combustion in 1772*. Cornell University Press, 1961.

[20] G.E. Stahl. *Zufällige Gedanken und nützliche Bedencken über den Streit, von dem sogenannten Sulphure*. Waysenhaus, Germany, 1718.

[21] Frank C Keil. *Semantic and conceptual development: An ontological perspective*. Harvard University Press, 1979.

[22] Frederic Lawrence Holmes. *Antoine Lavoisier: The Next Crucial Year: Or, the Sources of His Quantitative Method in Chemistry*. Princeton University Press, 1997.

[23] Michael Faraday. On lines of magnetic force: their definite character and their distribution within a magnet and through space. *Philosophical Transactions of the Royal Society of London*, 142:25–56., 1852.

[24] Michael Faraday. On the physical character of the lines of magnetic force. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 3(20):401–428, 1852.

[25] C. Vernon Boys. On the newtonian constant of gravitation. *Notices of the Proceedings*, 14:353–377, 1896.

[26] Alfonso Bueno-Orovio, David Kay, Vicente Grau, Blanca Rodriguez, and Kevin Burrage. Fractional diffusion models of cardiac electrical propagation: role of structural heterogeneity in dispersion of repolarization. *Journal of The Royal Society Interface*, 11(97):20140352, 2014.

[27] Nancy J Nersessian. *Creating scientific concepts*. MIT press, 2010.

[28] James Clerk Maxwell. On faraday's lines of force. In W. D. Niven, editor, *Scientific Papers*, pages 155–229. Cambridge University Press, 1855.

[29] James Clerk Maxwell. On physical lines of force. In W. D. Niven, editor, *Scientific Papers*, page 451–513. Cambridge University Press, 1861.

[30] James Clerk Maxwell. On physical lines of force. In W. D. Niven, editor, *Scientific Papers*, page 526–597. Cambridge University Press, 1864.

[31] Nancy J Nersessian. Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change. *Essays in the History and Philosophy of Science and Mathematics*, pages 129–166, 2002.

[32] Frederick Sanger and Hans Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical journal*, 49(4):463, 1951.

[33] JC Kendrew, G Bodo, HM Dintzis, RG Parrish, H Wyckoff, and DC Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[34] Max F Perutz, Michael G Rossmann, Ann F Cullis, Hilary Muirhead, Georg Will, and Anthony CT North. Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-å. resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422, 1960.

[35] MF Perutz, JC Kendrew, and HC Watson. Structure and function of haemoglobin: Ii. some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, 13(3):669–678, 1965.

[36] John A Thoma and DE Koshland Jr. Competitive inhibition by substrate during enzyme action. evidence for the induced-fit theory1, 2. *Journal of the American Chemical Society*, 82(13):3329–3333, 1960.

[37] Daniel E Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*, 44(2):98–104, 1958.

[38] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Walter M Fitch and Emanuel Margoliash. The usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol*, 4:67–109, 1970.

[41] William Bechtel and Robert C Richardson. *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press, 2010.

[42] Lindley Darden. *Theory change in science: Strategies from Mendelian genetics*. Oxford University Press, 1991.

[43] Paul Thagard. *The cognitive science of science: Explanation, discovery, and conceptual change*. Mit Press, 2012.

[44] Ruairidh M. Battleday and Samuel L. Gershman. *States of Mind: Lavoisier's Conceptual Revolution in Chemistry*. Princeton University Press, in preparation.

[45] Nancy Nersessian. In vitro analogies: Simulation modeling in bioengineering sciences. In Tarja Knuuttila, Natalia Carrillo, and Rami Koskinen, editors, *The Routledge Handbook of Philosophy of Scientific Modeling*. Routledge, 2024.

[46] J Gregory Trafton, Susan B Trickett, and Farilee E Mintz. Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science*, 10:89–106, 2005.

[47] Thomas L Griffiths. Manifesto for a new (computational) cognitive revolution. *Cognition*, 135:21–23, 2015.

[48] Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[49] Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):280–285, 2004.

[50] John Clement. Use of physical intuition and imagistic simulation in expert problem solving. In *Implicit and explicit knowledge*. Ablex Publishing, 1994.

[51] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.

[52] L. Laudan. *Progress and its problems*. University of California Press, Berkeley, CA, 1977.

[53] Melanie Mitchell. On crashing the barrier of meaning in artificial intelligence. *AI magazine*, 41(2):86–92, 2020.

[54] Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36, 2024.