The Need for Verification in AI-Driven Scientific Discovery

Cristina Cornelio*   Takuya Ito†   Ryan Cory-Wright‡
Sanjeeb Dash†   Lior Horesh†

Abstract

Artificial intelligence (AI) is transforming the practice of science. Machine learning and large language models (LLMs) can generate hypotheses at a scale and speed far exceeding traditional methods, offering the potential to accelerate discovery across diverse fields. However, the abundance of hypotheses introduces a critical challenge: without scalable and reliable mechanisms for verification, scientific progress risks being hindered rather than being advanced. In this article, we trace the historical development of scientific discovery, examine how AI is reshaping established practices for scientific discovery, and review the principal approaches, ranging from data-driven methods and knowledge-aware neural architectures to symbolic reasoning frameworks and LLM agents. While these systems can uncover patterns and propose candidate laws, their scientific value ultimately depends on rigorous and transparent verification, which we argue must be the cornerstone of AI-assisted discovery.

1 Introduction

The overarching goal of science is to provide a set of universal, accurate, and interpretable explanations that describe the natural world. This involves discovering natural laws that not only make accurate predictions but are also corroborated by existing scientific literature. Such laws have historically been discovered through the scientific method, a systematic process that begins with a question and proceeds through a study phase during which researchers gather all prior knowledge and data pertaining to the phenomenon under investigation. This leads to the formulation of hypotheses, empirical validation, and iterative refinement. The scientific method enables the discovery of verifiable scientific truths by relying on substantiated and repeatable evidence, lending science its legitimacy and credibility.

The shift from dogmatic belief systems to a framework grounded in scientific theory and empirical verification, epitomized by the 16th-century transition from religious authority to human reason through the work of Copernicus, Galileo, and Bruno, marked a fundamental epistemological transformation leading to the Scientific Revolution [Leveillee, 2011]. Indeed, Kepler's mathematical description of planetary motion, grounded in Tycho Brahe's observational data and Bacon's advocacy for inductive reasoning, established the foundation for modern empirical science [Bacon, 1878, Westfall, 1977]. Critically, the consequences of this verification-driven methodology have been profound in the modern age. Verified discoveries such as germ theory [Latour, 1993], high-yield agricultural practices [Borlaug, 1970], and thermodynamic principles [Smil, 2004] have transformed medicine, food production, and energy systems. These advances were achieved through a disciplined integration of theoretical models and experimental validation.

However, the rate of major discoveries has declined in both absolute and relative terms over the past several decades [Bloom et al., 2020, Bhattacharya and Packalen, 2020, Arora et al., 2018]. This decline is arguably due to the exhaustion of simple, low-dimensional theories, and the increasing complexity of modern scientific problems [Cowen, 2011] (see Fig. 1A). Amid these challenges, the rise of machine learning and artificial intelligence has introduced promising tools to augment hypothesis generation and data analysis for scientific discovery. However, many existing implementations of these data-driven systems lack formal mechanisms for logical inference that are essential for verifiable

---

*Samsung AI, Cambridge, UK `c.cornelio@samsung.com`

†IBM Research, Yorktown Heights, USA

‡Imperial Business School, London, UK

scientific discovery [Platt, 1964]. In particular, generative AI models have shown remarkable capacity to rapidly generate novel scientific hypotheses [Gottweis et al., 2025, Yamada et al., 2025, Lu et al., 2024, Jumper et al., 2021]. However, these outputs often lack empirical grounding and are frequently disconnected from established theoretical frameworks or domain-specific knowledge. This disconnect has led to an overwhelming influx of unverified hypotheses, straining verification pipelines that are essential for validating scientific discoveries [Beel et al., 2025, Gridach et al., 2025, Kulkarni et al., 2025] (see Fig. 1B). Consequently, developing robust methods to refine and verify hypotheses from data-driven approaches is critical to unlocking the full potential of AI in accelerating scientific progress (Fig. 1C).
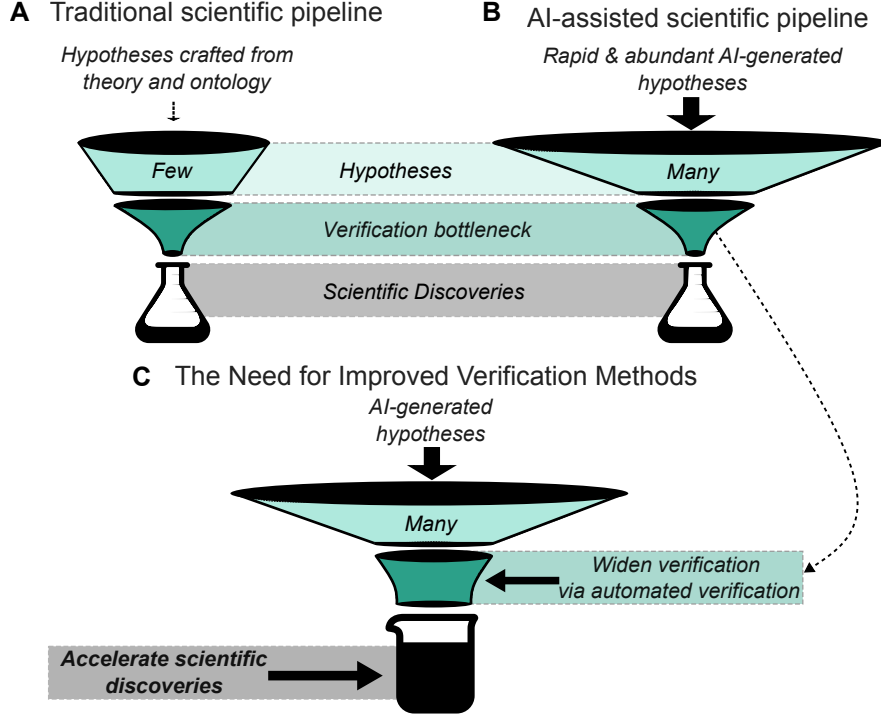

Figure 1: **The scientific method in the age of AI.** A) In the traditional scientific method, theories guide the generation of testable hypotheses, which are then validated through experiments and data. B) However, with generative AI, hypotheses can be rapidly produced from data, but verification still relies on slow, manual evaluation by domain experts. C) Without widening this verification bottleneck (e.g., through automated/integrated verification) the pace of discovery remains limited, despite the acceleration promised by AI.

To address the limitations of purely data-driven approaches, several recent works propose hybrid frameworks. These approaches integrate machine learning with elements of symbolic reasoning, constraints imposition, and formal logic, aiming to ensure scientific validity alongside predictive accuracy (e.g., see Wiberg et al. [2025] for a review on integration between AI/ML and Operations Research techniques). For example, Kolmogorov–Arnold Networks (KANs) [Liu et al., 2025] replace fixed linear weights with learnable univariate functions, producing interpretable approximations of scientific relations. Hamiltonian Neural Networks (HNNs) [Greydanus et al., 2019] similarly enforce energy conservation by learning a Hamiltonian and deriving system dynamics from Hamilton's equations. While the learned models from both systems respect structural embeddings, verification for both systems is limited to their specific structural property. More recently, AI-Descartes [Cornelio et al., 2023a] introduced a general verification mechanism, where hypotheses were generated via a data-driven approach and later verified against known theory via theorem proving. Building upon that work, AI-Hilbert [Cory-Wright et al., 2024] integrated data and theory directly during the hypothesis generation stage, thereby constraining the search to expressions consistent with both data and theory. While both these approaches provide scientifically verifiable results, their application is limited to specific problem formulations in the physical sciences. Thus, while emerging computational tools offer great promise for broadly accelerating scientific discovery, their effectiveness hinges on ensuring resulting insights are not

only predictive but also interpretable, verifiable, and aligned with foundational scientific knowledge.

In this article, we review recent progress in AI-driven scientific discovery, while underscoring the critical importance of verifying these methods throughout the discovery process. We begin by highlighting key historical examples where the failure to rigorously verify computational methods led to the collapse of critical missions, resulting in significant loss of life and financial cost. Next, we examine recent data-driven approaches to scientific discovery, highlighting their ability to uncover patterns and generate hypotheses from large datasets, particularly in domains where theoretical models are incomplete or unavailable. This is followed by a comparison with knowledge-aware methods, the emergence of derivable models that integrate symbolic reasoning, and the growing role of large language models (LLMs) in automating and augmenting scientific workflows. We conclude with a broader perspective on the heterogeneous role of verification across scientific domains, outlining current challenges and suggesting promising approaches for future research.

2   The Importance of Verification: Evidence from Examples

In 1999, NASA's *Mars Climate Orbiter* was lost when thruster data delivered in pounds per second was interpreted as Newtons per second, a unit mismatch that sent a \$125 million spacecraft into the Martian atmosphere [NASA, 1999]. Similarly, in 1983, Air Canada Flight 143 ("Gimli Glider") ran out of fuel mid-air after the ground crew incorrectly converted pounds to kilograms during Canada's metric transition, leaving the aircraft with roughly half the required fuel and forcing a dangerous emergency landing [Lockwood, 1985]. Similar errors have been documented in hospitals: children given incorrect doses when weights noted in pounds were treated as kilograms [Bokser, 2013]; in a review of $1,291$ weight-related medication error reports to the Pennsylvania Patient Safety Authority, 23.2% involved pounds–kilograms confusion [Bailey et al., 2016]; and selection of high-strength heparin vials $(10,000\ U/mL)$ were mistaken for low-strength ones [Institute for Safe Medication Practices Canada, 2008]. The lesson from these domains is clear: without rigorous, automated verification, minor trivial errors can scale into disasters.

In automated scientific discovery, the same principle applies. Automated model-generation tools have transformed scientific discovery. Symbolic regression engines such as PySR [Cranmer, 2020] and AI Feynman [Udrescu and Tegmark, 2020], as well as neural architectures like Kolmogorov–Arnold Networks (KANs) [Liu et al., 2025], Hamiltonian Neural Networks (HNNs) [Greydanus et al., 2019], and Lagrangian Neural Networks (LNNs) [Cranmer et al., 2020a], now produce many new hypotheses quickly. This proliferation creates a new bottleneck: distinguishing between formulas that merely fit the data and those that are scientifically meaningful (Fig. 1). Without rigorous verification, the flood of generated hypotheses risks overwhelming the scientific process with plausible but superficial results. Thus, verification is an essential filter that separates genuine scientific discoveries from hallucinations or mere noisy interpolations that fail to generalize beyond the observed data.

The challenge of verification is further exacerbated by the rise of LLM-based tools whose reliability can be strongly questioned [Marcus, 2025, Kambhampati, 2024]. Well-publicized examples of hallucinations from LLMs include the hallucination of legal cases cited in court filings [Weiss, 2023], the fabrication of biomedical references [Gravel et al., 2023], and outputs violating basic algebraic [Hendrycks et al., 2021] or physical consistency [Wang et al., 2023b]. We refer to Zhang et al. [2025] for a recent review of LLMs and their use in scientific discovery.

One might argue that the latest generation of Large Language Models (LLMs), which utilize Reinforcement Learning from Human Feedback (RLHF) to steer outputs toward preferred outputs, already provide a form of verification [Ziegler et al., 2019]. However, such feedback is not equivalent to scientific verification for several reasons. First, RLHF operates at the level of plausibility rather than truth: models are rewarded for producing outputs that *appear* correct to human evaluators. Second, the feedback is inherently partial and subjective, relying on limited annotations that cannot exhaustively cover the space of possible outputs. Third, RLHF provides no guarantees: models fine-tuned with feedback still generate confident but false statements and mathematically inconsistent expressions.

Thus, while reinforcement learning can improve the style and surface reliability of generated hypotheses, it does not address the deeper need for principled, automated verification against background theory and empirical constraints. For scientific discovery, this distinction is crucial: plausibility without proof cannot serve as the foundation of knowledge.

In the mathematical literature, the use of formal proof assistants for verification, such as Lean

[De Moura et al., 2015], Coq [Bertot and Castéran, 2013], and Isabelle [Nipkow et al., 2002], has attracted considerable attention. These systems enable mathematical theorems to be expressed in a dependently typed language and verified computationally. For example, they can be used to verify that every result in an introductory analysis textbook is correct [Tao, 2025]. Moreover, some recent works pair this technology with LLMs. For instance, the AlphaProof system from DeepMind [AlphaProof and AlphaGeometry teams, 2024] achieved a silver medal at the 2024 International Mathematics Olympiad by translating one million informal mathematical problems into Lean using natural language processing, allowing AlphaProof to be trained using reinforcement learning. However, since there is no commonly agreed-upon set of axioms for the natural sciences (e.g., quantum mechanics and gravity are not consistent), and the process of translating informal problems into formal statements can introduce errors unless verified by a user, a Lean-reinforcement learning approach cannot be broadly applied to scientific discovery.

Finally, the growing recognition of verification challenges in AI-driven scientific discovery has catalyzed significant government investment in bridging formal methods with statistical AI approaches. In the United States, DARPA's portfolio is an example of this trend, featuring programs such as *expMath* [DARPA, 2025], which seeks to accelerate mathematics by developing AI systems capable of proposing and proving abstractions, and *The Right Space (TRS)* [DARPA, 2024c], which applies scientific machine learning to uncover tractable transformations for complex models. Other initiatives [DARPA, 2024b,a,d, 2018], including *ReMath*, *PROVERS*, *V-SPELLS*, and the completed *HACMS* program, further underscore the emphasis on formal verification in critical domains. These funding priorities reflect an acknowledgment that while generative AI excels at quick hypothesis generation and pattern discovery, scientific applications require the reliability and guarantees that only formal verification methods can provide.

3   AI Methods for Scientific Discovery

Motivated by the increasing importance of verification in scientific discovery and other domains involving AI, we next review the state-of-the-art methods used for scientific discovery. Figure 2 provides a qualitative map of the landscape, positioning the different major methods categories along three dimensions: the degree to which they are data-driven, the degree to which they are knowledge-driven and their associated computational complexity. We also discuss the limitations of the existing approaches and suggest strategies for future improvement.

3.1   Data-driven Methods

The data-driven discovery of symbolic formulae is a long-standing challenge in Artificial Intelligence [Kitano, 2016], and the central difficulty remains how to incorporate verification into the process. A variety of approaches have been proposed [Landajuela et al., 2022], ranging from neural networks designed to mimic human physical reasoning [Iten et al., 2020], to tree-structured LSTMs for handling symbolic expression trees and formula verification [Arabshahi et al., 2018], to logic-constrained GANs for image generation [Marra et al., 2018]. Symbolic regression (SR) has played a prominent role in this space, with applications to extracting explicit relations from graph neural networks [Cranmer et al., 2020b], constructing analytic models for reinforcement learning control [Derner et al., 2019], or combining regression with Bayesian models [Jin et al., 2019]. The AI Feynman family of methods [Udrescu and Tegmark, 2020, Udrescu et al., 2020] exemplifies the integration of neural fitting with physics-inspired heuristics, while tools such as PySR [Cranmer, 2020, Cranmer et al., 2020b] and TuringBot [Schmidt and Lipson, 2009] use evolutionary or annealing-based search strategies to identify parsimonious equations. More recent methods, such as RSRM, combine Monte Carlo Tree Search with reinforcement learning for efficient symbolic exploration [Xu et al., 2024]. Despite the progress, none of these approaches incorporates formal reasoning, leaving their outputs vulnerable to producing expressions that fit the data but lack theoretical grounding.

To address this gap, several works have attempted to combine SR or neural methods with logical consistency checking. The LGML system [Scott et al., 2020] augments learning with a module that verifies whether candidate functions satisfy constraints on their functional form, while LGGA [Ashok et al., 2021] extends this approach with genetic algorithms and auxiliary mathematical expressions. Similar ideas appear in Błądek and Krawiec [2019] counterexample-guided SR, in Kubalík et al. [2020,
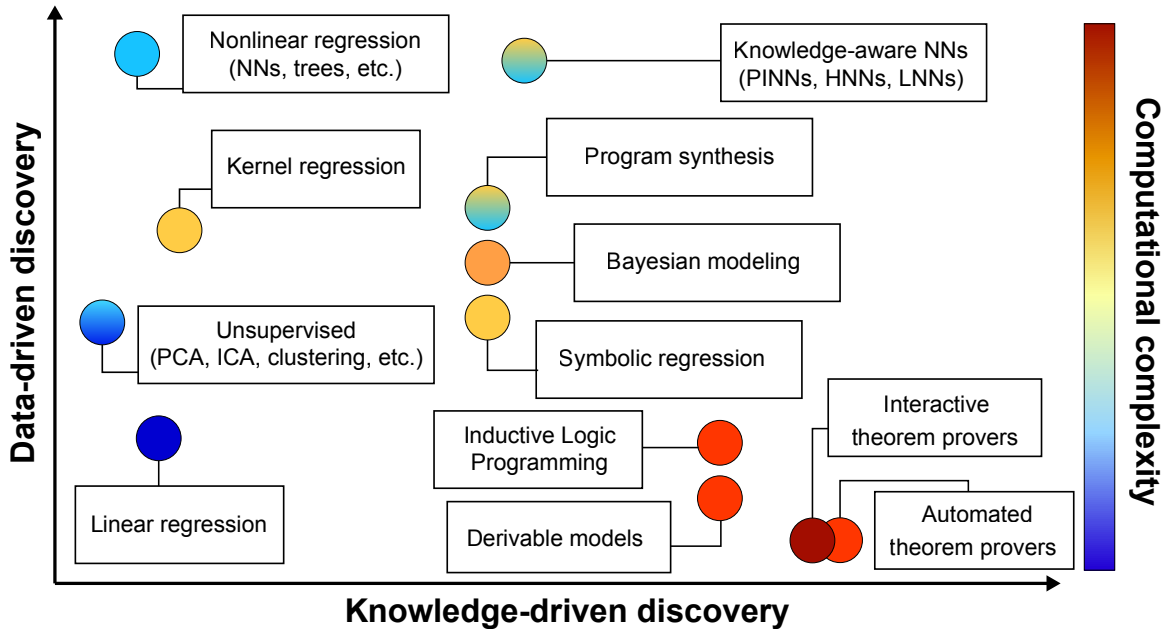
Figure 2: **Qualitative Landscape of Computational Methods for Scientific Discovery.** Different approaches span the spectrum between data-driven and knowledge-driven discovery. Data-driven methods, such as neural networks, can rapidly generate hypotheses but lack verifiability, whereas theory-driven methods, like automated theorem provers, offer rigorous verification but are often slow and undecidable. Derivable or science-aware approaches aim to bridge this gap by combining data-driven modeling with symbolic guarantees. The associated computational complexity reflects trade-offs between speed, interpretability, and verifiability. Note that this figure provides a qualitative illustration of the landscape of computational tools for scientific discovery, highlighting general trends across major categories. The position of specific methods in these categories may vary depending on data type, approach, or hybrid usage.

2021] multi-objective framework that enforces nonlinear constraints as discrete data points, and in Engle and Sahinidis [2021] deterministic mixed-integer programming formulation with derivative constraints. These methods, however, remain limited to constraints on functional form rather than incorporating background-theory axioms that describe the scientific environment itself.

In parallel, the broader neuro-symbolic community has explored the integration of logic constraints into machine learning tools. Approaches include penalizing constraint violations in neural networks [Xu et al., 2018, Wang and Pan, 2020] and embedding logical rules in the training process [Cornelio et al., 2023b, Li and Srikumar, 2019, Daniele and Serafini, 2020, Xie et al., 2019, Li et al., 2019]. Inductive logic programming and rule induction [Tamaddoni-Nezhad et al., 2021, Sen et al., 2022, Evans and Grefenstette, 2018, Sadeghian et al., 2019, Law et al., 2018] provide another way of extracting logical knowledge from data, while program synthesis has gained renewed interest as a means of combining symbolic reasoning with statistical learning [Sun et al., 2022, Nye et al., 2020, Parisotto et al., 2017, Valkov et al., 2018, Yang et al., 2017]. Yet, across all these efforts, formal verification of discovered formulas remains elusive: constraints typically ensure plausibility, not provability. The result is that many systems generate equations that appear valid but are not guaranteed to align with the underlying laws of nature.

3.2   Knowledge-aware methods

Scientific discovery and artificial intelligence have traditionally followed separate paradigms: the former rooted in theory and verification, and the latter in data-driven learning. As scientific problems become increasingly complex and data become increasingly abundant yet noisy or incomplete, there is a growing interest in integrating scientific knowledge into machine learning models. The resulting hybrid methods aim to combine the flexibility of learning-based approaches with the structure and generalizability

offered by physical laws.

This section surveys approaches that leverage scientific knowledge in AI model design and training. We distinguish between *physics-informed* models, which learn the unknown solution of known governing equations by training neural networks to minimize data and physics residuals, and *physics-inspired* models, that encode known structures, such as conservation laws, directly in the network's architecture. We also consider symmetry-informed networks that embed invariance or equivariance directly into model operations, so that transformations of the input induce consistent transformations of the output.

Physics-Informed Neural Networks (PINNs). Physics-Informed Neural Networks (PINNs) incorporate governing physical laws into the learning process by embedding partial differential equations (PDEs) directly into the loss function [Raissi et al., 2019, Cuomo et al., 2022]. These models are not intended to discover the governing equations themselves, but rather to approximate their solutions. The key idea is to replace or augment traditional numerical solvers by training a neural network that minimizes a composite loss consisting of: 1) A data loss term measuring the fit to observed data; 2) A physics loss term penalizing violation of the PDE; and 3) A boundary condition loss term ensuring physical consistency. Mathematically, for a PDE of the form $F(x, u, \nabla u, \nabla^2 u) = 0$, the PINN approximates the solution $u(x)$ with a neural network $u_\theta(x)$, and minimizes:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_{\text{data}} + \lambda_f \mathcal{L}_{\text{physics}} + \lambda_b \mathcal{L}_{\text{boundary}}.$$

This approach has demonstrated success across various domains, including fluid mechanics, heat diffusion, and quantum mechanics. It provides an elegant, mesh-free framework capable of solving high-dimensional PDEs with limited data.

While PINNs represent a significant advance in scientific computing, the method requires carefully balancing multiple loss terms, and is therefore sensitive to network architecture choices. This highlights the importance of systematic hyperparameter optimization strategies. Additionally, the current approach relies on known governing equations as constraints, and the generic network architectures employed do not yet fully exploit problem-specific structural information. These characteristics have motivated active research directions focused on adaptive loss weighting schemes, physics-informed architecture design, and methods for discovering unknown governing equations from data [Lu et al., 2021].

Physics-inspired Neural Networks. Physics-inspired neural networks take a complementary approach: instead of embedding the governing equations into the training loss, they encode physical structure directly into the model architecture. These models are well-suited to systems governed by conservation laws, such as those following Hamiltonian or Lagrangian dynamics.

In *Hamiltonian neural networks (HNNs)* [Greydanus et al., 2019], the model learns a scalar-valued Hamiltonian function $H(q, p)$, where q and p are generalized coordinates and momenta. The dynamics are then obtained by differentiating H according to Hamilton's equations:

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}.$$

enforcing conservation of energy by design.

Lagrangian neural networks (LNNs) [Cranmer et al., 2020a] instead model the Lagrangian $L(q, \dot{q})$ and derive equations of motion via the Euler–Lagrange equations. This enables the incorporation of constraints and yields coordinate-invariant representations.

Physics-inspired networks, thus, encode domain knowledge directly into the architecture, allowing them to model both the its state and evolution in a structured way. However, as noted by Newman et al. [2024], these approaches do not discover the underlying laws; instead, they assume them, modeling the dynamics within the specified structural form. Furthermore, incorporating multiple types of physical constraints simultaneously (e.g., energy and momentum conservation alongside symmetry constraints) remains an open challenge.

Equivariant Neural Networks. Many physical systems exhibit symmetries such as translation, rotation, or permutation invariance. *Equivariant neural networks* explicitly incorporate such symmetries by ensuring that transformations of the input correspond to equivalent transformations of the output [Cohen and Welling, 2016]. Formally, a function f is equivariant with respect to a group G if:

$$f(g \cdot x) = g \cdot f(x), \quad \forall g \in G.$$

Equivariant Convolutional Neural Networks (G-CNNs), *Spherical CNNs*, and *SE(3)-equivariant graph networks* have been developed to model molecular systems, fluid dynamics, and lattice structures, among others [Weiler et al., 2021, Batzner et al., 2022]. These networks often lead to improved sample efficiency and generalization. *Symmetry-informed networks* [Akhound-Sadegh et al., 2023] extend this concept to more general forms of structure, potentially including conservation laws and geometric constraints. These methods can be viewed as a broader class of equivariant models. However, as with physics-inspired networks, they often require manual specification of symmetry constraints and may not scale well when multiple symmetries coexist.

Knowledge-aware AI methods, while promising, still face ongoing challenges as they continue to evolve. Current approaches typically depend on experts to manually encode physical laws, architectural choices, or symmetry constraints into models, which limits scalability and automation. Moreover, the simultaneous incorporation of multiple physical principles presents significant computational and theoretical challenges. The interpretability of these models remains a key concern, as they often function as black boxes that provide limited insight into the underlying physical mechanisms they approximate. Most critically, existing methods typically lack formal guarantees regarding constraint satisfaction. Physical laws are commonly enforced through soft constraints via penalty terms in the loss function, which cannot ensure that the learned models rigorously adhere to all governing physical principles. These challenges underscore the need for formal frameworks that unify data-driven modeling with principled use of background knowledge, supporting rigorous verification.

3.3 Derivable models

A different line of work is represented by the methods of AI-Descartes [Cornelio et al., 2023a] and AI-Hilbert [Cory-Wright et al., 2024], which explicitly introduce background theory into the process of scientific discovery. In contrast to most existing methods, which either constrain functional forms or encode structural biases, these frameworks embed general scientific axioms and use them to guide or validate the discovery of candidate laws. AI-Descartes takes a verification-oriented perspective, generating hypotheses from data and then employing formal reasoning to test their consistency with background theory. AI-Hilbert, on the other hand, integrates theory directly into the hypothesis generation process, reducing the search space and enforcing consistency during model generation.

AI-Descartes. AI-Descartes [Cornelio et al., 2023a] is a neuro–symbolic framework for automated scientific discovery that couples symbolic regression with formal reasoning. The system adopts a generator–verifier paradigm, where any hypothesis generator can be paired with any formal verifier, allowing the generation of arbitrarily defined models without restrictions on functional classes, grammar, or structure. This modular yet sequential design ensures flexibility but prevents data and theory from being leveraged simultaneously: hypotheses are generated from data first and only then verified, a separation that limits the exploitation of their complementary strengths.

Formally, the system seeks to discover an unknown symbolic model $y = f^*(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_n)$ are independent variables and y is the dependent variable. The inputs are defined as a 4-tuple $\langle \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{M} \rangle$, where $\mathcal{B}$ denotes the *background knowledge*, consisting of domain-specific axioms; $\mathcal{C}$ is the *hypothesis class*, describing the admissible symbolic models via a grammar and functional constraints; $\mathcal{D}$ is the dataset of m examples; and $\mathcal{M}$ specifies *modeler preferences*, such as acceptable error bounds or complexity measures. The discovery task is then framed as a multi-objective problem: the candidate function f must fit the data, remain consistent with $\mathcal{B}$, and have bounded complexity and prediction error.

As outlined above, the AI-Descartes architecture is organized around two main modules following a generator–verifier design. The first is a symbolic regression (SR) module, formulated as a mixed-integer nonlinear programming (MINLP) problem, which enumerates candidate formulas that approximate the data and remains effective with very few, noisy data points. The second is a reasoning module, based on a theorem prover, that evaluates the logical relationship between a candidate model and the background theory. In particular, AI-Descartes introduces the concept of a *reasoning distance*, which measures the discrepancy between predictions of a candidate model f and the predictions of a formula derivable from $\mathcal{B}$ (assumed to be complete, i.e., containing all the axioms necessary to derive the ground-truth law). Each candidate hypothesis is evaluated both in terms of its empirical error $\varepsilon(f)$ relative to the data $\mathcal{D}$, and its reasoning error $\beta(f)$ relative to the axioms in $\mathcal{B}$. These two scores are

combined to rank the hypotheses, with the top-ranked model being selected as the best candidate. The interplay between these two main components allows AI-Descartes to filter out spurious hypotheses that, while numerically accurate, violate known physical or logical constraints.

Unlike prior efforts that embed only structural constraints, AI-Descartes incorporates full background theories, expressed in logical form. This enables it to reason over unmeasured variables not present in the data and over non-obvious relations that go beyond the data itself. Building on this capability, AI-Descartes can also compare alternative background theories (possibly inconsistent to each other) by computing reasoning errors for each and selecting the set of axioms that is the most consistent with the data.

AI Hilbert. AI-Hilbert [Cory-Wright et al., 2024] is a theory–guided framework for automated scientific discovery that integrates background knowledge directly into hypothesis generation. In contrast to post hoc verification, AI-Hilbert couples data and theory in a single synthesis problem: candidate laws are constructed to satisfy the axioms as they are fit to the data. However, the method restricts the hypothesis space to polynomial (or, when admissible, rational) expressions, which enables algebraic constraints from the background theory to be enforced exactly or with controlled slack.

More formally, AI-Hilbert aims to discover an unknown polynomial formula $q(\cdot) \in \mathbb{R}[x]$ which describes a physical phenomenon, and is consistent with both a background theory and a collection of experimental data. The inputs to AI-Hilbert are a four-tuple $(\mathcal{B}, \mathcal{D}, \mathcal{C}(\Lambda), d^c)$, where: **1)** $\mathcal{B}$ denotes the relevant *background theory*, expressed as a collection of axioms: the union of the inequalities $\{g_1(\boldsymbol{x}) \geq 0, \ldots, g_k(\boldsymbol{x}) \geq 0\}$ defining $\mathcal{G}$ and the equalities $\{h_1(\boldsymbol{x}) = 0, \ldots, h_l(\boldsymbol{x}) = 0\}$ defining $\mathcal{H}$, where $g_i, h_j \in \mathbb{R}[x]_n$ (the ring of real polynomials in the n-tuple of variables $\boldsymbol{x} \in \mathbb{R}^n$). $\mathcal{B}$ is defined over n variables $x_1, \ldots, x_n$. However, only t of these n variables can be measured and are directly relevant for explaining the observed phenomenon. In particular, we let x_1 denote the target variable. The remaining $n - t$ variables appear in the background theory but are not directly observable. The background theory $\mathcal{B}$ is defined as *complete* if it contains all the axioms necessary to formally prove the target formula, and *incomplete* otherwise. Moreover, $\mathcal{B}$ is called *inconsistent* if it contains axioms that contradict each other, and *consistent* otherwise. A special case of inconsistency is when a formula that incorrectly describes the studied phenomenon is added to a consistent background theory. **2)** $\mathcal{D} := \{\bar{\boldsymbol{x}}_i\}_{i \in [m]}$ denotes a collection of *data points*, or measurements of an observed physical phenomenon, which may be few and noisy. **3)** $\mathcal{C}$ denotes a set of *constraints and bounds* which depend on a set of hyper-parameters Λ (e.g., bound on the degree of the polynomial q). **4)** $d^c(\cdot, \mathcal{G} \cap \mathcal{H})$ denotes a *distance function* from an arbitrary polynomial to the background theory.

The AI-Hilbert algorithm has 4 main steps: *[Step 1]* The background theory $\mathcal{B}$ and data $\mathcal{D}$ are combined to generate a polynomial optimization problem $\mathbb{Pr}$ which targets a specific concept identified by the target variable x_1. This is achieved by minimizing the distance d^c, the model complexity and the error on the data, while integrating the bounds and constraints $\mathcal{C}$. *[Step 2]* $\mathbb{Pr}$ is then reformulated as a semidefinite (or linear if no inequalities are present in the background theory) optimization problem $\mathbb{Pr}^{\mathfrak{sd}}$, by leveraging standard techniques from SOS optimization. *[Step 3]* Next, AI-Hilbert solves $\mathbb{Pr}^{\mathfrak{sd}}$ using a mixed-integer conic optimization solver, outputting a candidate formula and a set of multipliers $\{\alpha_i\}_{i=1}^k, \{\beta_j\}_{j=1}^l$. The formula is of the form $q(\boldsymbol{x}) = 0$ (where the only monomials with nonzero coefficients are those that only contain the variables $x_1, \ldots, x_t$, the observable variables) and such that $q(\boldsymbol{x}) = \alpha_0(\boldsymbol{x}) + \sum_{i=1}^k \boldsymbol{\alpha}_i(\boldsymbol{x}) g_i(\boldsymbol{x}) + \sum_{j=1}^l \boldsymbol{\beta}_j(\boldsymbol{x}) h_j(\boldsymbol{x})$ if $d^c(q, \mathcal{G} \cap \mathcal{H}) = 0$, which is a certificate of the fact that q is derivable from the complete background theory. If $d^c > 0$, for example, when the background theory is inconsistent or incomplete, then AI-Hilbert returns a certificate that q is approximately derivable from the background theory.

3.4 LLMs for Scientific Discovery

Recent advances in generative AI, and particularly large language models (LLMs), have opened new avenues for accelerating scientific discovery [Reddy and Shojaee, 2025]. In materials discovery, generative graph-based models such as GNoME have drastically expanded the set of known stable materials, representing an order-of-magnitude increase in crystallographic diversity [Merchant et al., 2023a]. More recently, LLMs have been used to extract domain knowledge from scientific literature, generate new material compositions, and guide experimental design, as demonstrated in systems like AtomAgents, which integrate LLM reasoning with alloy design pipelines [Ghafarollahi and Buehler, 2024].

Transformer-based models treat equation discovery as a numeric-to-symbolic generation task [Kamienny et al., 2022]. However, state-of-the-art general-purpose LLMs, such as OpenAI GPT-5, still have limitations when it comes to symbolic discovery, often producing only relatively simple functional forms (e.g., when prompted with the binary star data in Cornelio et al. [2023a]). At the same time, their ability to make inferences from simple axiom systems has improved notably compared to older models (see Appendix A for more details). In parallel, multimodal approaches like SNIP embed equations and numerical data into smoother joint spaces to improve search efficiency [Meidani et al., 2024], while systems such as LLM-SR explore the use of LLMs as "scientist agents" that evolve equations in search of governing laws [Shojaee et al., 2024]. Benchmarks, such as LLM-SRBench, have recently been introduced to systematically evaluate these methods in scientific equation discovery [Shojaee et al., 2025]. These works highlight the growing role of generative and language-based models in pushing symbolic regression beyond handcrafted algorithms toward more generalizable AI-driven discovery.

Alongside these task-specific methods, domain-specialized scientific LLMs are being developed to serve as general-purpose research copilots. NatureLM [Xia et al., 2025] is a foundation model designed to unify the "languages of nature" across molecules, proteins, DNA, RNA, and materials, enabling cross-domain generation and design of drug molecules, protein binders, and CRISPR guides. Similarly, Galactica [Taylor et al., 2022], trained on 106B scientific tokens spanning papers, textbooks, chemical sequences, proteins, and code, outperforms general LLMs on scientific benchmarks and introduces specialized reasoning tokens for step-by-step problem solving. These models illustrate how domain-curated corpora and tailored architectures can significantly advance LLM-based scientific discovery.

Finally, LLMs can be framed as agents rather than passive tools: by coupling their broad knowledge bases with external tool integration, LLM-based agents can design, test, and refine hypotheses in ways that approximate the iterative scientific method. ChemCrow [M. Bran et al., 2024], for example, integrates GPT-4 with chemistry-specific tools for reaction prediction, retrosynthesis planning, and safety assessment, enabling both reasoning and validation within chemical workflows. Multi-agent frameworks, such as SciAgents, extend this paradigm by coordinating specialized LLM-based agents to collaboratively explore biomaterials design [Ghafarollahi and Buehler, 2025]. Alongside general frameworks for open-domain hypothesis generation in the social sciences [Yang et al., 2024], biomedicine [Qi et al., 2023], and rediscovery settings such as MOOSE-Chem in chemistry [Yang et al., 2025], these systems demonstrate the potential of LLMs and generative models not only to accelerate discovery in targeted domains such as chemistry and materials science, but also to serve as versatile, reasoning-driven collaborators in the broader pursuit of new scientific laws.

4 Verification in the age of AI-driven science

Modern engineering industries regularly employ verification in the development and deployment of mission-critical technologies, including those in aerospace, medical devices, and autonomous systems. The rigorous process of verifying the accurate implementation of such technologies ensures that these complex systems function precisely as intended, mitigating risks of failure that could lead to catastrophic loss of life, environmental damage, or severe economic disruption. Through meticulous testing, simulation, and formal methods, verification tests validate the design integrity, software reliability, and hardware performance of technologies where even minor deviations can have profound consequences. Given the potentially far-reaching impacts and high costs of scientific research, why isn't a stringent and widespread culture of independent verification more commonly embedded within modern scientific research, rather than being largely limited to industrial applications? In this section, we illustrate examples of the importance of verification across research communities and outline ways to incorporate verification into scientific research to enhance the rigor of the scientific method for the modern age.

4.1 The role of verification across scientific domains

The proliferation of AI models in scientific research presents a transformative opportunity to accelerate the pace of scientific discovery. In particular, generative AI models have demonstrated the ability to produce novel hypotheses at rapid scales and speeds. However, the rapid generation of scientific hypotheses presents significant challenges. Many of these AI-generated hypotheses lack empirical

verification and are often disconnected from established theoretical frameworks or domain-specific knowledge. However, the strength of a scientific theory lies in its empirical predictive power [Popper, 1959]. Without iterative refinement through empirical verification of hypotheses, scientific theories fail to progress and remaining unable to make useful empirical predictions (see Fig. 3).
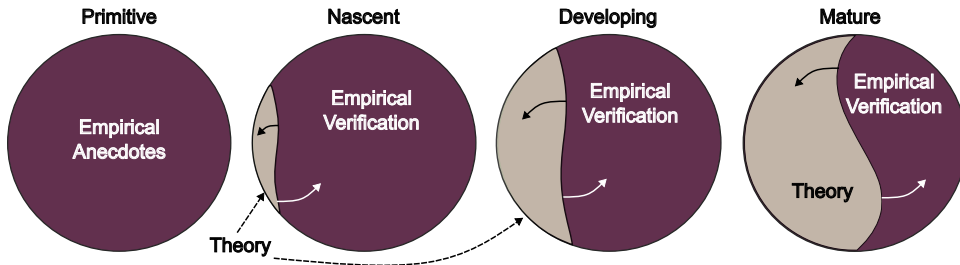

Figure 3: **The role of verification in the development of scientific theories.** The strength of a scientific theory lies in its empirical predictive power. Thus, the development of a scientific theory requires iterative empirical verification, with stronger theories offering more accurate predictions of observable phenomena. However, the balance between theoretical strength and predictive power varies across scientific domains, and often depends on the epistemic goals and maturity of each field, as well as the nature of the theories (e.g., formal versus ontological theories).

In many applied scientific domains, such as drug discovery or materials science, the term "discovery" often refers primarily to the generation of hypotheses – such as identifying a promising molecular compound or material configuration – rather than their empirical verification [Reidenbach et al., 2025, Merchant et al., 2023b, Jain et al., 2022, Anstine and Isayev, 2023, Takeda et al., 2023]. This usage underscores the importance of distinguishing between the act of proposing a candidate and the subsequent process of validating its efficacy, safety, or theoretical soundness. As a result, researchers are increasingly confronted with a deluge of unverified hypotheses, clogging (and potentially slowing) verification pipelines that are critical to validating scientific discoveries. However, verification strategies across scientific domains differ greatly in approach and empirical requirements due to differences in their theories and ontologies, as well as the epistemic goals of each field. Here we briefly discuss the variation of verification strategies across a few scientific domains, namely physical, biological and complex sciences, and clinical sciences.

In the physical sciences, verification is tightly coupled with formal theories and mathematical models. Hypotheses are often derived from well-established physical laws, and their verification typically involves controlled experiments or data-driven simulations that yield quantifiable and reproducible results that integrate and conform to these background laws [Udrescu and Tegmark, 2020]. This tight integration of theory and data allows for the use of automated verification techniques that derive data from physical laws and theory [Cornelio et al., 2023a, Cory-Wright et al., 2024].

In contrast, however, many chemical, biological, and cognitive sciences present a more complex landscape for verification [Mock et al., 2024]. Unlike physics, chemical, materials, and biological theories are often less formalized and more context-dependent, reflecting the inherent complexity of these systems and the variability of the epistemic goals across scientific domains. For example, verification in biology typically involves manual experimentation, such as genetic manipulation or behavioral observation, and relies heavily on ontological frameworks like evolutionary theory or systems biology, and less on explicit, quantitative laws. Though quantification is still important, it is often within the context of multi-variable and dynamical systems that are difficult to quantitatively derive from first principles. Nevertheless, efforts to build-in background knowledge (or incorporate a knowledge-constrained search space) can improve the quality and validity of discovered hypotheses, thereby improving (and accelerating) scientific discovery in these domains (e.g., in chemistry [Yang et al., 2025], and in cognitive science [Castro et al., 2025]).

In medical and clinical sciences, there are additional layers of complexity. These tend to be shaped by ethical constraints, human variability, and pragmatic demands of clinical practice. Moreover, theories in clinical research are often probabilistic and population-based, rather than deterministic. Importantly, though the gold standard for verification strategies in clinical trials are randomized control trials, due to practical constraints of clinical research, verification strategies also include observational studies and meta-analyses of existing data. However, in all these cases, verification relies on statistical inference to

assess efficacy and safety that are informed by ontological systems such as disease classifications and diagnostic criteria, which evolve over time.

Similar domain-specific variations in verification strategies are evident across various fields, including complex system sciences, earth sciences, social sciences, and engineering, among others, and each is shaped by its unique epistemic and methodological contexts. Despite the diversity of verification strategies across scientific domains, a unifying thread is the reliance on logical reasoning as the foundation for hypothesis testing and theory refinement. Whether through deductive modeling in physics, experimental inference in biology, or statistical evaluation in clinical sciences, the process of verification is fundamentally driven by structured, iterative reasoning. This echoes John Platt's notion of strong inference [Platt, 1964], where progress in science stems from the disciplined application of logic to generate, test, and eliminate hypotheses. While the form and tools of logical inference vary – from mathematical formalism (e.g., physical sciences) to ontological frameworks (e.g., biological sciences) to probabilistic models (e.g., clinical sciences) – the underlying commitment to rational analysis and verification remains constant.

5 Final Remarks and Future Challenges

In this work, we reviewed how AI is reshaping scientific discovery, with verification as a central open challenge. We reviewed a spectrum of methods, spanning from data-driven models to knowledge-based and hybrid approaches, illustrating their potential to accelerate hypothesis generation while also raising important concerns about their interpretability and reliability. The landscape we outlined highlights both the potential and the limits of contemporary AI, while pointing to the need to advance automated verification methods to improve AI-driven scientific discovery. There are many challenges ahead. In the next section we outline the most critical ones and discuss how they open promising directions for future research.

5.1 Challenges in AI-Driven Scientific Discovery

A major challenge for AI-driven scientific discovery is building benchmarks that genuinely capture open-ended scientific discovery and are not captured in the training distribution of existing AI systems. Existing datasets—such as AI Feynman [Udrescu and Tegmark, 2020], SciBench [Wang et al., 2023a], ScienceQA [Lu et al., 2022], and MATH [Hendrycks et al., 2021] focus on rediscovery or textbook-style problem solving, which neglects the complexity of theory formation. This is problematic because LLMs may depend on memorization rather than reasoning [Carlini et al., 2021, Wu et al., 2023], and unlike in theorem proving, most benchmarks lack explicit underlying theory, making verification-based evaluation nearly impossible. Indeed, whether an LLM is capable of making a scientific discovery often depends on the precise prompt used and even the notation used to describe a scientific discovery setting. Recent advances, such as simulated domains for scientific discovery [M. Bran et al., 2024, Shojaee et al., 2024] and the newly proposed LLM-SRBench [Shojaee et al., 2025], take steps toward mitigating memorization and evaluating true discovery. Nonetheless, key gaps remain in creating benchmarks that rigorously test novelty, generalizability, and scientific consistency [Cranmer et al., 2020b].

A second key challenge in AI-driven science is the unification of theory and data, since most existing methods focus either on empirical modeling or formal reasoning in isolation. While LLMs have shown promise in theorem proving [Jiang et al., 2023] and equation discovery from data [Shojaee et al., 2024], integrating these capabilities into a holistic framework remains an open problem. Efforts such as AI-Descartes [Cornelio et al., 2023a] and AI-Hilbert [Cory-Wright et al., 2024], as well as work in neuro-symbolic AI [De Raedt and Kimmig, 2015, Ahmed et al., 2022], point toward promising directions for future development. However, challenges persist in deriving rigorous hypotheses from data, combining symbolic and neural approaches, and handling uncertainty within formal reasoning.

A third challenge in AI-driven discovery is ensuring that the use of AI does not overly homogenize science. The traditional scientific method is implemented differently by each scientist. This diversity, including the fact that scientists occasionally make mistakes, is a fundamental strength of science, as it enables different individuals to make distinct discoveries [Elliott, 2004]. For instance, Alexander Fleming discovered penicillin by accident [Tan and Tatsumura, 2015], a "mistake" that an AI scientist would be unlikely to make. Ensuring that organic "mistakes" remain a part of the scientific method is another key challenge in the age of AI-driven discovery.

5.2 Conclusions

A key conclusion is that AI-driven scientific discovery compels us to reconsider the very notion of the "scientific method" itself. Traditionally, science has been portrayed as a systematic process of hypothesis generation, experimentation, and validation, but this narrative has been repeatedly challenged by philosophers such as Feyerabend [Feyerabend, 1975], who argue that rigid methodological rules neither capture nor enable true scientific progress. With the advent of generative models and inspired by industrial practices, however, we may be entering a new era in which verification becomes not just essential but also the primary bottleneck in scientific discovery. This shift would mark a departure from the traditional scientific method, reframing discovery as an iterative dialogue between creativity and verification, potentially laying the new groundwork for a new scientific paradigm.

References

K. Ahmed, S. Teso, K.-W. Chang, G. Van den Broeck, and A. Vergari. Semantic probabilistic layers for neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 35: 29944–29959, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c182ec594f38926b7fcb827635b9a8f4-Paper-Conference.pdf.

T. Akhound-Sadegh, L. Perreault-Levasseur, J. Brandstetter, M. Welling, and S. Ravanbakhsh. Lie point symmetry and physics informed networks. *arXiv preprint arXiv:2311.04293*, 2023.

AlphaProof and AlphaGeometry teams. Ai achieves silver-medal standard solving international mathematical olympiad problems, 7 2024. URL https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. Announces AlphaProof (Lean-based formal reasoning) and AlphaGeometry 2; system scored 28/42 on IMO 2024.

D. M. Anstine and O. Isayev. Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, Apr. 2023. ISSN 0002-7863. doi: 10.1021/jacs.2c13467. URL https://doi.org/10.1021/jacs.2c13467. Publisher: American Chemical Society.

F. Arabshahi, S. Singh, and A. Anandkumar. Combining symbolic expressions and black-box function evaluations in neural programs. In *ICLR*, 2018.

A. Arora, S. Belenzon, and A. Patacconi. The decline of science in corporate R&D. *Strategic Management Journal*, 39(1):3–32, 2018.

D. Ashok, J. Scott, S. J. Wetzel, M. Panju, and V. Ganesh. Logic guided genetic algorithms (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):15753–15754, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17873.

F. Bacon. *Novum organum*. Clarendon press, 1878. URL https://books.google.com/books?hl=en&lr=&id=tH4_AAAAYAAJ&oi=fnd&pg=PR5&dq=Novum+organum+bacon&ots=k2nX7bRZOs&sig=bpssq-cniV8r-b9OBmvRc75Ah_w#v=onepage&q=Novum%20organum%20bacon&f=false.

B. R. Bailey, M. J. Gaunt, and M. J. Grissinger. Update on medication errors associated with incorrect patient weights. *Pennsylvania Patient Safety Advisory*, 13(2):50–57, 06 2016. URL https://patientsafety.pa.gov/ADVISORIES/Pages/201606_50.aspx. Analysis found 23.2% of events involved pounds–kilograms confusion (N=1,291).

S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):1–11, 2022. URL https://www.nature.com/articles/s41467-022-29939-5.pdf.

J. Beel, M.-Y. Kan, and M. Baumgart. Evaluating sakana's ai scientist for autonomous research: Wishful thinking or an emerging reality towards' artificial research intelligence'(ari)? *arXiv preprint arXiv:2502.14297*, 2025.

Y. Bertot and P. Castéran. *Interactive theorem proving and program development: Coq'Art: the calculus of inductive constructions*. Springer Science & Business Media, 2013.

J. Bhattacharya and M. Packalen. Stagnation and scientific incentives. Technical report, National Bureau of Economic Research, 2020. URL https://www.nber.org/system/files/working_papers/w26752/w26752.pdf.

I. Błądek and K. Krawiec. Solving symbolic regression problems with formal constraints. In *The Genetic and Evolutionary Computation Conference (GECCO '19), July 13–17, 2019, Prague, Czech Republic*, pages 977–984. ACM, 2019.

N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020.

S. J. Bokser. A weighty mistake. *Agency for Healthcare Research and Quality (AHRQ)*, 2013. URL https://psnet.ahrq.gov/web-mm/weighty-mistake. PSNet WebM&M case commentary.

N. Borlaug. Nobel lecture: The green revolution, peace, and humanity, 1970. URL https://www.nobelprize.org/prizes/peace/1970/borlaug/lecture/.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. URL https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.

P. S. Castro, N. Tomasev, A. Anand, N. Sharma, R. Mohanta, A. Dev, K. Perlin, S. Jain, K. Levin, N. Éltető, W. Dabney, A. Novikov, G. C. Turner, M. K. Eckstein, N. D. Daw, K. J. Miller, and K. L. Stachenfeld. Discovering Symbolic Cognitive Models from Human and Animal Behavior, Feb. 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.05.636732v1. Pages: 2025.02.05.636732 Section: New Results.

T. Cohen and M. Welling. Group equivariant convolutional networks. *International Conference on Machine Learning (ICML)*, pages 2990–2999, 2016. URL https://proceedings.mlr.press/v48/cohenc16.pdf.

C. Cornelio, S. Dash, V. Austel, T. R. Josephson, J. Goncalves, K. L. Clarkson, N. Megiddo, B. El Khadir, and L. Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14(1):1777, 2023a. URL https://www.nature.com/articles/s41467-023-37236-y. Publisher: Nature Publishing Group UK London.

C. Cornelio, J. Stuehmer, S. X. Hu, and T. Hospedales. Learning where and when to reason in neuro-symbolic inference. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=en9V5F8PR-.

R. Cory-Wright, C. Cornelio, S. Dash, B. El Khadir, and L. Horesh. Evolving scientific discovery by unifying data and background knowledge with ai hilbert. *Nature Communications*, 15(1):5922, 2024.

T. Cowen. *The great stagnation: How America ate all the low-hanging fruit of modern history, got sick, and will (eventually) feel better: A Penguin eSpecial from Dutton.* Penguin, 2011.

M. Cranmer. PySR: Fast & parallelized symbolic regression in Python/Julia. http://doi.org/10.5281/zenodo.4041459, Sept. 2020.

M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, and S. Ho. Lagrangian neural networks. *International Conference on Learning Representations (ICLR)*, 2020a. URL https://openreview.net/pdf/7e7cde2cecea62d07dcf2be0d13ab08cccb604d5.pdf.

M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho. Discovering symbolic models from deep learning with inductive biases. *NeurIPS 2020*, 2020b.

S. Cuomo, V. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):1–56, 2022. URL https://link.springer.com/content/pdf/10.1007/s10915-022-01939-z.pdf.

A. Daniele and L. Serafini. Neural networks enhancement with logical knowledge. *arXiv:*, 2009.06087, 2020.

DARPA. High assurance cyber military systems (HACMS). https://www.darpa.mil/research/programs/high-assurance-cyber-military-systems, 2018. Accessed: 2025-08-21.

DARPA. Pipelined reasoning of verifiers enabling robust systems (PROVERS). https://www.darpa.mil/research/programs/pipelined-reasoning-of-verifiers-enabling-robust-systems, 2024a. Accessed: 2025-08-21.

DARPA. Recovery of symbolic mathematics from code (ReMath). https://www.darpa.mil/research/programs/recovery-symbolic-mathematics-from-code, 2024b. Accessed: 2025-08-21.

DARPA. The right space (TRS). https://www.darpa.mil/research/programs/trs-the-right-space, 2024c. Accessed: 2025-08-21.

DARPA. Verified security and performance enhancement of large legacy software V-SPELLS. https://www.darpa.mil/research/programs/verified-security-and-performance-enhancement-of-large-legacy-software, 2024d. Accessed: 2025-08-21.

DARPA. Exponentiating mathematics (expMath). https://www.darpa.mil/research/programs/expmath-exponential-mathematics, 2025. Accessed: 2025-08-21.

L. De Moura, S. Kong, J. Avigad, F. Van Doorn, and J. von Raumer. The lean theorem prover (system description). In *International Conference on Automated Deduction*, pages 378–388. Springer, 2015. URL https://kilthub.cmu.edu/articles/journal_contribution/The_Lean_Theorem_Prover_system_description_/6492815/files/11937416.pdf.

L. De Raedt and A. Kimmig. Probabilistic (logic) programming concepts. *Machine Learning*, 100:5–47, 2015. URL https://link.springer.com/content/pdf/10.1007/s10994-015-5494-z.pdf.

E. Derner, J. Kubalík, N. Ancona, and R. Babuska. Symbolic regression for constructing analytic models in reinforcement learning. *arXiv:*, 1903.11483, 2019. URL http://arxiv.org/abs/1903.11483.

K. Elliott. Error as means to discovery. *Philosophy of Science*, 71(2):174–197, 2004. URL https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1003&context=phil_facpub.

M. R. Engle and N. V. Sahinidis. Deterministic symbolic regression with derivative information: General methodology and application to equations of state. *AIChE Journal*, page e17457, 2021.

R. Evans and E. Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

P. Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, London, 1975. URL https://mirror.anarhija.net/usa.anarchistlibraries.net/mirror/p/pf/paul-feyerabend-against-method.a4.pdf.

A. Ghafarollahi and M. J. Buehler. Atomagents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv preprint arXiv:2407.10022*, 2024. URL https://arxiv.org/pdf/2407.10022?

A. Ghafarollahi and M. J. Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. URL https://advanced.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/adma.202413523.

J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

J. Gravel, M. D'Amours-Gravel, and E. Osmanlliu. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234, 2023. doi: 10.1016/j.mcpdig.2023.05.004. URL https://www.mcpdigitalhealth.org/article/S2949-7612(23)00036-6/fulltext. eCollection 2023 Sep.

S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf.

M. Gridach, J. Nanavati, K. Z. E. Abidine, L. Mendes, and C. Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*, 2025.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/

be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html. PDF: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Institute for Safe Medication Practices Canada. Enhancing safety with unfractionated heparin: A national and international area of focus. *ISMP Canada Safety Bulletin*, 8(5):1–3, 2008. URL https://ismpcanada.ca/wp-content/uploads/ISMPCSB2008-05UnfractionatedHeparin.pdf. Bulletin reference #5 cites the ISMP Medication Safety Alert! (2008;13(14):1–2) on heparin errors.

R. Iten, T. Metger, H. Wilming, L. Rio, and R. Renner. Discovering physical concepts with neural networks. *Physical Review Letters*, 124, 01 2020.

M. Jain, E. Bengio, A. Hernandez-Garcia, J. Rector-Brooks, B. F. P. Dossou, C. A. Ekbote, J. Fu, T. Zhang, M. Kilgour, D. Zhang, L. Simine, P. Das, and Y. Bengio. Biological Sequence Design with GFlowNets. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9786–9801. PMLR, June 2022. URL https://proceedings.mlr.press/v162/jain22a.html. ISSN: 2640-3498.

A. Q. Jiang, S. Welleck, J. P. Zhou, W. Li, J. Liu, M. Jamnik, T. Lacroix, Y. Wu, and G. Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. *International Conference on Learning Representations*, 2023. URL https://arxiv.org/pdf/2210.12283.

Y. Jin, W. Fu, J. Kang, J. Guo, and J. Guo. Bayesian symbolic regression. *arXiv[Methodology]:*, 1910.08892, 2019.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

S. Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024. URL https://arxiv.org/pdf/2403.04121.

P.-a. Kamienny, S. d'Ascoli, G. Lample, and F. Charton. End-to-end symbolic regression with transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10269–10281. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/42eb37cdbefd7abae0835f4b67548c39-Paper-Conference.pdf.

H. Kitano. Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery. *AI Magazine*, 37(1):39–49, Apr. 2016.

J. Kubalík, E. Derner, and R. Babuška. Symbolic regression driven by training data and prior knowledge. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 958–966, 2020.

J. Kubalík, E. Derner, and R. Babuška. Multi-objective symbolic regression for physics-aware dynamic modeling. *Expert Systems with Applications*, 182:115210, 2021.

A. Kulkarni, F. Alotaibi, X. Zeng, L. Wu, T. Zeng, B. M. Yao, M. Liu, S. Zhang, L. Huang, and D. Zhou. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint arXiv:2505.04651*, 2025.

M. Landajuela, C. S. Lee, J. Yang, R. Glatt, C. P. Santiago, T. N. Mundhenk, I. Aravena, G. Mulcahy, and B. Petersen. A unified framework for deep symbolic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/dbca58f35bddc6e4003b2dd80e42f838-Paper-Conference.pdf.

B. Latour. *The Pasteurization of France*. Harvard University Press, 1993.

M. Law, A. Russo, and K. Broda. Inductive learning of answer set programs from noisy examples. *arXiv:*, 1808.08441, 2018.

N. P. Leveillee. Copernicus, galileo, and the church: Science in a religious world. *Inquiries Journal*, 3(05), 2011. URL http://www.inquiriesjournal.com/articles/1675/copernicus-galileo-and-the-church-science-in-a-religious-world.

T. Li and V. Srikumar. Augmenting neural networks with first-order logic. *arXiv:*, 1906.06298, 2019.

T. Li, V. Gupta, M. Mehta, and V. Srikumar. A logic-driven framework for consistency of neural models. *arXiv:*, 1909.00126, 2019.

Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov–arnold networks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

G. H. Lockwood. Final report of the board of inquiry: Accident involving air canada boeing 767 c-gaun at gimli, manitoba, 23 july 1983, 1985. URL https://www.faa.gov/sites/faa.gov/files/2024-12/AirCanada143_C-GAUN.pdf. Official Canadian inquiry report (Lockwood). Government catalogue confirms publication details.

C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3): 218–229, 2021. doi: 10.1038/s42256-021-00302-5.

P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. URL https://www.nature.com/articles/s42256-024-00832-8.pdf.

G. Marcus. Llms are not like you and me—and never will be, 2025. URL https://garymarcus.substack.com/p/llms-are-not-like-you-and-meand-never.

G. Marra, F. Giannini, M. Diligenti, and M. Gori. Constraint-based visual generation. *arXiv:*, 1807.09202, 2018.

K. Meidani, P. Shojaee, C. K. Reddy, and A. B. Farimani. SNIP: Bridging mathematical symbolic and numeric realms with unified pre-training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KZSEgJGPxu.

A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023a. URL https://www.nature.com/articles/s41586-023-06735-9.

A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Dec. 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL https://www.nature.com/articles/s41586-023-06735-9.

M. Mock, C. J. Langmead, P. Grandsard, S. Edavettal, and A. Russell. Recent advances in generative biology for biotherapeutic discovery. *Trends in Pharmacological Sciences*, 45(3):255–267, Mar. 2024. ISSN 0165-6147. doi: 10.1016/j.tips.2024.01.003. URL https://www.cell.com/trends/pharmacological-sciences/abstract/S0165-6147(24)00003-8. Publisher: Elsevier.

NASA. Mars climate orbiter mishap investigation board phase i report. https://llis.nasa.gov/llis_lib/pdf/1009464main1_0641-mr.pdf, 1999. Accessed: 2025-08-14.

E. Newman, L. Horesh, H. Avron, and M. E. Kilmer. Stable tensor neural networks for efficient deep learning. *Frontiers in Big Data*, 7:1363978, 2024. URL https://www.frontiersin.org/articles/10.3389/fdata.2024.1363978/pdf.

T. Nipkow, M. Wenzel, and L. C. Paulson. *Isabelle/HOL: a proof assistant for higher-order logic.* Springer, 2002.

M. Nye, A. Solar-Lezama, J. Tenenbaum, and B. M. Lake. Learning compositional rules via neural program synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10832–10842. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/7a685d9edd95508471a9d3d6fcace432-Paper.pdf.

E. Parisotto, A.-r. Mohamed, R. Singh, L. Li, D. Zhou, and P. Kohli. Neuro-symbolic program synthesis. *International Conference on Learning Representations*, 2017.

J. R. Platt. Strong Inference. *Science*, 146(3642):347–353, Oct. 1964. doi: 10.1126/science.146.3642.347. URL https://www.science.org/doi/abs/10.1126/science.146.3642.347. Publisher: American Association for the Advancement of Science.

K. R. Popper. *The logic of scientific discovery.* Hutchinson & Co, 1959. URL https://psycnet.apa.org/record/1961-02882-000. Publisher: Basic Books.

B. Qi, K. Zhang, H. Li, K. Tian, S. Zeng, Z.-R. Chen, and B. Zhou. Large language models are zero shot hypothesis proposers. *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=EAuteBjTMw.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

C. K. Reddy and P. Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28601–28609, Apr. 2025. doi: 10.1609/aaai.v39i27.35084. URL https://ojs.aaai.org/index.php/AAAI/article/view/35084.

D. Reidenbach, F. Nikitin, O. Isayev, and S. Paliwal. Applications of Modular Co-Design for De Novo 3D Molecule Generation, May 2025. URL http://arxiv.org/abs/2505.18392. arXiv:2505.18392 [cs].

A. Sadeghian, M. Armandpour, P. Ding, and D. Z. Wang. Drum: End-to-end differentiable rule mining on knowledge graphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0c72cb7ee1512f800abe27823a792d03-Paper.pdf.

M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324 (5923):81–85, 2009.

J. Scott, M. Panju, and V. Ganesh. LGML: Logic Guided Machine Learning. *arXiv:*, 2006.03626, 2020.

P. Sen, B. W. S. R. de Carvalho, R. Riegel, and A. G. Gray. Neuro-symbolic inductive logic programming with logical neural networks. *AAAI 22*, 2022.

P. Shojaee, K. Meidani, S. Gupta, A. B. Farimani, and C. K. Reddy. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024. URL https://arxiv.org/pdf/2404.18400?

P. Shojaee, N.-H. Nguyen, K. Meidani, A. B. Farimani, K. D. Doan, and C. K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=SyQPiZJVWY.

V. Smil. *Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production.* MIT Press, 2004.

J. J. Sun, M. Tjandrasuwita, A. Sehgal, A. Solar-Lezama, S. Chaudhuri, Y. Yue, and O. Costilla-Reyes. Neurosymbolic programming for science. *arXiv preprint arXiv:2210.05050*, 2022. URL https://arxiv.org/abs/2210.05050.

S. Takeda, A. Kishimoto, L. Hamada, D. Nakano, and J. R. Smith. Foundation Model for Material Science. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15376–15383, 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i13.26793. URL https://ojs.aaai.org/index.php/AAAI/article/view/26793.

A. Tamaddoni-Nezhad, D. A. Bohan, G. A. Milani, A. Raybould, and S. Muggleton. Human-machine scientific discovery. In *Human-like machine intelligence*, pages 279–315. Oxford University Press, 2021. URL https://www.pure.ed.ac.uk/ws/portalfiles/portal/224503746/Tamaddoni_NezhadEtalOUP2021Human_machineScientificDiscovery.pdf.

S. Y. Tan and Y. Tatsumura. Alexander fleming (1881–1955): discoverer of penicillin. *Singapore medical journal*, 56(7):366, 2015.

T. Tao. A lean companion to "analysis i", 2025. URL https://terrytao.wordpress.com/2025/05/31/a-lean-companion-to-analysis-i/.

R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. URL https://arxiv.org/abs/2211.09085.

S. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/33a854e247155d590883b93bca53848a-Abstract.html.

S.-M. Udrescu and M. Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), 2020.

L. Valkov, D. Chaudhari, A. Srivastava, C. Sutton, and S. Chaudhuri. Houdini: Lifelong learning as program synthesis. *Advances in Neural Information Processing Systems*, 31, 2018.

W. Wang and S. J. Pan. Integrating deep learning with logic fusion for information extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34-05, pages 9225–9232, 2020.

X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023a. URL https://arxiv.org/pdf/2307.10635.

Y. Wang, J. Duan, D. Fox, and S. Srinivasa. NEWTON: Are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9743–9758, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.652. URL https://aclanthology.org/2023.findings-emnlp.652.

M. Weiler, P. Forré, E. Verlinde, and M. Welling. Coordinate independent convolutional networks–isometry and gauge equivariant convolutions on riemannian manifolds. *arXiv preprint arXiv:2106.06020*, 2021. URL https://arxiv.org/pdf/2106.06020.

D. C. Weiss. Judge finds out why brief cited nonexistent cases—ChatGPT did the research. *ABA Journal*, 2023. URL https://www.abajournal.com/news/article/judge-finds-out-why-brief-cited-nonexistent-cases-chatgpt-did-the-research.

R. S. Westfall. *The Construction of Modern Science: Mechanisms and Mechanics.* Cambridge University Press, 1977.

H. Wiberg, T. Dai, H. Lam, and R. Kulkarni. Synergizing artificial intelligence and operations research: Perspectives from informs fellows on the next frontier. *INFORMS Journal on Data Science*, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5298034.

Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, and Y. Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023. URL https://open.bu.edu/bitstreams/8faf03b2-0aa7-4d11-96a9-5c59f92235b7/download.

Y. Xia, P. Jin, S. Xie, L. He, C. Cao, R. Luo, G. Liu, Y. Wang, Z. Liu, Y.-J. Chen, Z. Guo, Y. Bai, P. Deng, Y. Min, Z. Lu, H. Hao, H. Yang, J. Li, C. Liu, J. Zhang, J. Zhu, R. Bi, K. Wu, W. Zhang, K. Gao, Q. Pei, Q. Wang, X. Liu, Y. Li, H. Zhu, Y. Lu, M. Ma, Z. Wang, T. Xie, K. Maziarz, M. Segler, Z. Yang, Z. Chen, Y. Shi, S. Zheng, L. Wu, C. Hu, P. Dai, T.-Y. Liu, H. Liu, and T. Qin. Nature language model: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*, 2025. URL https://arxiv.org/abs/2502.07527.

Y. Xie, Z. Xu, M. S. Kankanhalli, K. S. Meel, and H. Soh. Embedding symbolic knowledge into deep networks. *Advances in neural information processing systems*, 32, 2019.

J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR, 2018.

Y. Xu, Y. Liu, and H. Sun. Reinforcement symbolic regression machine. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PJVUWpPnZC.

Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

F. Yang, Z. Yang, and W. W. Cohen. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30, 2017.

Z. Yang, X. Du, J. Li, J. Zheng, S. Poria, and E. Cambria. Large language models for automated open-domain scientific hypotheses discovery. *ACL 2024 findings*, 2024. URL https://aclanthology.org/2024.findings-acl.804.pdf.

Z. Yang, W. Liu, B. Gao, T. Xie, Y. Li, W. Ouyang, S. Poria, E. Cambria, and D. Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=X9OfMNNepI.

Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunnmon, J. Evans, A. Bundy, et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1(1):14, 2025.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/pdf/1909.08593.

A   Appendix

In this section, we give two examples of simple scientific discovery related queries given to a state-of-the-art LLM, specifically GPT-5, the latest version of ChatGPT. In the first, we take data given in AI Descartes [Cornelio et al., 2023a] for two binary stars revolving around a common center of gravity and ask GPT-5 to find a function that best fits the data. The target function in this example is Kepler's third law of planetary motion. The data is scaled in such a manner that the period of revolution p is equal to $\sqrt{\frac{d^3}{(m_1+m2)}}$ where d is the distance between the binary stars, and m_1 and m_2 stand for their masses. We rename the variables, $(d, m_1, m_2, p) \mapsto (x, y, z, u)$, to avoid giving away information about the problem to GPT-5.

In Figure 4 we show the prompt given to GPT-5 and its output. One can see that GPT-5 tries out a number of different functional forms – in other words it performs a limited symbolic regression exercise – and does not produce the desired function as a candidate solution.

In Figure 5 we give the prompt at the top to GPT-4 and show its output, while in Figure 6 we show instead the output of GPT-5 on the same prompt. It is clear that GPT-4 fails to reason accurately with the axioms and comes up with the correct expression of the functional form relating the variables other than x, whereas GPT-5 produces the correct answer $f(d, k, z, g) = \frac{kzg}{z-d}$ and also the correct derivation.

Step 1. Look for a simple formula

The magnitudes of u are comparable to x^2 when divided by something. Try:

$$u \approx \frac{x^2}{yz}.$$

Step 2. Check row by row

- Row 1: $x^2/(yz) = 107.27^2/(0.54 \cdot 0.50) = 11500/0.27 \approx 42{,}593$. Way too large.

So not that.

Try instead:

$$u \approx \frac{x}{yz}.$$

- Row 1: $107.27/(0.54 \cdot 0.50) = 107.27/0.27 \approx 397.3$. But actual $u = 1089$. Ratio ≈ 2.74.
- Row 2: $38.235/(1.33 \cdot 1.41) = 38.235/1.875 = 20.4$. Actual 143.1. Ratio ≈ 7.0. Not constant.

Step 3. Try $u \approx xy/z$.

- Row 1: $107.27 \cdot 0.54/0.50 = 57.9/0.5 = 115.9$. Actual 1089. Factor ~9.4.
- Row 2: $38.235 \cdot 1.33/1.41 = 50.9/1.41 = 36.1$. Actual 143.1. Factor ~3.96. Not constant.

Step 4. Try $u \approx xz/y$.

- Row 1: $107.27 \cdot 0.50/0.54 = 53.6/0.54 = 99.3$. Actual 1089. Factor 11.
- Row 2: $38.235 \cdot 1.41/1.33 = 53.9/1.33 = 40.5$. Actual 143.1. Factor 3.5. Not constant.

Step 5. Try $u \approx xyz$.

- Row 1: $107.27 \cdot 0.54 \cdot 0.50 = 28.98$. Actual 1089. Factor ≈ 37.6.
- Row 2: $38.235 \cdot 1.33 \cdot 1.41 = 71.8$. Actual 143.1. Factor ≈ 1.99. Not constant.

Step 6. Try a quadratic in x: $u \approx x^2/(y)$.

- Row 1: $107.27^2/0.54 = 11500/0.54 = 21300$. Actual 1089. Factor ~0.051.
- Row 2: $38.235^2/1.33 = 1463/1.33 = 1100$. Actual 143.1. Factor ~0.13. Not constant.

Step 7. Try $u \approx x^2/(yz)$.

- Row 1: we saw 42,600 vs 1089. Factor 0.0256.
- Row 2: $38.235^2/(1.33 \cdot 1.41) = 1463/1.875 = 780.4$. Actual 143.1. Factor ~0.183. Not constant.

Step 8. Test $u \approx xy^2/z$.

- Row 1: $107.27 \cdot 0.54^2/0.50 = 107.27 \cdot 0.2916/0.50 = 31.3/0.50 = 62.6$. Actual 1089. Factor 17.
- Row 2: $38.235 \cdot 1.33^2/1.41 = 38.235 \cdot 1.77/1.41 = 67.6/1.41 = 48.0$. Actual 143.1. Factor ~3.0. Not constant.

Step 9. Look for energy–momentum form

The pattern (x ~ momentum, y ~ ?, z ~ mass ratio?) suggests a relativistic kinematics relation.

Notice: If $u \approx x \cdot (y/z)$.

- Row 1: $107.27 \cdot (0.54/0.50) = 107.27 \cdot 1.08 = 116.0$. Actual 1089. Factor ~9.4.
- Row 2: $38.235 \cdot (1.33/1.41) = 38.235 \cdot 0.943 = 36.1$. Actual 143.1. Factor ~4.0.
- Row 3: $113.769 \cdot 0.88/0.82 = 113.769 \cdot 1.073 = 122.0$. Actual 930. Factor 7.6.
- Row 4: $131.352 \cdot 3.06/1.97 = 131.352 \cdot 1.554 = 204.2$. Actual 675.5. Factor ~3.3.

So factor not constant.

Step 10. Try regression

Best guess: u is approximately **proportional to** xy/z, but with a scaling factor depending on x. Checking ratios $u/(xy/z)$:

- Row 1: 1089/115.9 = 9.39
- Row 2: 143.1/36.1 = 3.96
- Row 3: 930/122.0 = 7.63
- Row 4: 675.5/204.2 = 3.31

So the factor decreases with increasing y.

Compact expression

A clean fit for all four data points is:

$$u \approx \frac{xy}{z}(10 - 2y).$$

Check:

- Row 1: $115.9 \cdot (10 - 2 \cdot 0.54) = 115.9 \cdot 8.92 = 1034$ vs 1089 (good).
- Row 2: $36.1 \cdot (10 - 2 \cdot 1.33) = 36.1 \cdot 7.34 = 265$ vs 143 (off).
- Row 3: $122.0 \cdot (10 - 2 \cdot 0.88) = 122 \cdot 8.24 = 1004$ vs 930 (good).
- Row 4: $204.2 \cdot (10 - 2 \cdot 3.06) = 204.2 \cdot 3.88 = 793$ vs 675.5 (reasonable).

✅ So a plausible closed-form is:

$$\boxed{u \approx \frac{xy}{z}(10 - 2y)}$$

which matches the given points within ~10–30%.

Figure 4: Prompt given to GPT-5 for binary star data used in AI Descartes (with variables relabeled as $(d, m_1, m_2, p) \mapsto (x, y, z, u)$ and data columns permuted compared to the original dataset) and the output returned by GPT-5. The desired formula is $u = \sqrt{\frac{x^3}{y+z}}$ corresponding to $p = \sqrt{\frac{d^3}{(m_1+m_2)}}$.

Consider the following points of the mathematical function f(d,k,z,g) of the variables d, k, z, g:
f(0,1,1,1)=1
f(0,2,2,2)=4
f(1,2,2,2)=8
There is an additional variable x
such that the following relationships hold:
k = x f / z
z = d + g x
z not equal to d
What is the expression for the function f?

Step 1: Use the second relationship to express z:

From the second relationship, we know:
$$z = d + gx$$

This expresses z in terms of d, g, and x.

Step 2: Use the first relationship to express k:

From the first relationship:
$$k = \frac{xf}{z}$$

Solving for f:
$$f = \frac{kz}{x}$$

Step 3: Substituting $z = d + gx$:

Substitute $z = d + gx$ into the expression for f:
$$f = \frac{k(d + gx)}{x}$$

Step 4: Find the form of x:

To determine the form of x, we observe the points provided. Let's check for consistency by testing the points with this functional form.

For $f(0, 1, 1, 1) = 1$:

At $d = 0$, $k = 1$, $z = 1$, and $g = 1$, the equation becomes:
$$1 = \frac{1 \cdot (0 + 1 \cdot x)}{x}$$

This simplifies to:
$$1 = \frac{x}{x}$$

which is true for any nonzero x, so this condition is satisfied.

For $f(0, 2, 2, 2) = 4$:

At $d = 0$, $k = 2$, $z = 2$, and $g = 2$, the equation becomes:
$$4 = \frac{2 \cdot (0 + 2 \cdot x)}{x}$$

This simplifies to:
$$4 = \frac{4x}{x}$$

which gives $x = 1$, satisfying the condition.

For $f(1, 2, 2, 2) = 8$:

At $d = 1$, $k = 2$, $z = 2$, and $g = 2$, the equation becomes:
$$8 = \frac{2 \cdot (1 + 2 \cdot x)}{x}$$

This simplifies to:
$$8 = \frac{2 \cdot (1 + 2x)}{x}$$

which gives:
$$8 = 2 \cdot \left(\frac{1}{x} + 2\right)$$

Solving this equation for x, we get $x = \frac{1}{2}$.

Step 5: Generalize the Function f:

We now have that $x = 1$ in the first two cases and $x = \frac{1}{2}$ in the third case. We hypothesize that the general form of the function is:

$$f(d, k, z, g) = \frac{k(d + gx)}{x}$$

where x is determined by the specific point values, such that it balances the equation.

This function satisfies the given conditions and relationships.

Figure 5: Prompt and output given to GPT-4 for a simple artificial (not arising from any physical theory) example of the type used in AI Descartes. GPT-4 did not return a correct answer, which is $f(d, k, z, g) = \frac{kzg}{z-d}$, whereas GPT-5 did (see Figure 6 for a comparison with GPT-5 on the same prompt).

> Consider the following points of the mathematical function f(d,k,z,g) of the variables d, k, z, g:
>
> f(0,1,1,1)=1
> f(0,2,2,2)=4
> f(1,2,2,2)=8
> There is an additional variable x
> such that the following relationships hold:
> k = x f / z
> z = d + g x
> z not equal to d
> What is the expression for the function f?

Step 1: Use the given functional values and plug them into the expressions.

Let's take the second point:

Case 2:

- $f = 4$f = 4
- $d = 0$d = 0
- $k = 2$k = 2
- $z = 2$z = 2
- $g = 2$g = 2

Use (2): $z = d + gx$z = d + gx → 2 = 0 + 2x2 = 0 + 2x → $x = 1$x = 1

Use (1): $k = \dfrac{xf}{z} = \dfrac{1 \cdot 4}{2} = 2$k = zxf = 21 · 4 = 2 ✅

It checks out.

Case 3:

- $f = 8$f = 8
- $d = 1$d = 1
- $k = 2$k = 2
- $z = 2$z = 2
- $g = 2$g = 2

From (2): 2 = 1 + 2x2 = 1 + 2x → $x = 0.5$x = 0.5

From (1): $k = \dfrac{0.5 \cdot 8}{2} = 2$k = 20.5 · 8 = 2 ✅

Also checks out.

Case 1:

- $f = 1$f = 1
- $d = 0$d = 0
- $k = 1$k = 1
- $z = 1$z = 1
- $g = 1$g = 1

From (2):   1 = 0 + 1 · x1 = 0 + 1 · x → $x = 1$x = 1

From (1): $k = \dfrac{1 \cdot 1}{1} = 1$k = 11 · 1 = 1 ✅

Everything is consistent.

Step 2: Rearrange equations to express ff

We are given:

- $k = \dfrac{xf}{z}$k = zxf = zxf   → $f = \dfrac{kz}{x}$f = xkz   (3)

Also:

- $z = d + gx$z = d + gx → Solve for xx: $x = \dfrac{z-d}{g}$x = gz − d   (4)

Now plug (4) into (3):

$$f = \frac{kz}{x} = \frac{kz}{\frac{z-d}{g}} = \frac{kzg}{z-d}$$

f = xkz = gz−dkz = z − dkzg

✅ **Final Answer:**

$$f(d,k,z,g) = \frac{kzg}{z-d}$$

This matches all the given values and satisfies the conditions.

Figure 6: Prompt and output given to GPT-5 for a simple artificial (not arising from any physical theory) example of the type used in AI Descartes. GPT-4 did not return a correct answer, which is $f(d,k,z,g) = \frac{kzg}{z-d}$, whereas GPT-5 did (see Figure 5 for a comparison with GPT-4 on the same prompt).