Decoding complexity: how machine learning is redefining scientific discovery

Ricardo Vinuesa[1,2*], **Paola Cinnella**[3], **Jean Rabault**[4], **Hossein Azizpour**[5,2], **Stefan Bauer**[6,7], **Bingni W. Brunton**[8], **Arne Elofsson**[9,2], **Elias Jarlebring**[10,2], **Hedvig Kjellström**[5,2], **Stefano Markidis**[11,2], **David Marlevi**[12,13], **Javier García-Martínez**[14*], and **Steven L. Brunton**[15*]

[1]FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Swedish e-Science Research Centre, (SeRC), Stockholm, Sweden
[3]Institut Jean le Rond D'Alembert, Sorbonne Université, France
[4]IT Department, Norwegian Meteorological Institute, 0313 Oslo, Norway
[5]Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden
[6]TUM School of Computation, Information and Technology, Technical University Munich, Munich, Germany
[7]Helmholtz AI, Helmholtz Center Munich, Munich, Germany
[8]Department of Biology, University of Washington, Seattle, WA 98195, USA
[9]Dept. of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, 171 21 Solna
[10]Dept. Mathematics, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden
[11]Department of Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
[12]Dept. Molecular Medicine and Surgery, Karolinska Institutet, 171 77 Stockholm, Sweden
[13]Inst. for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[14]Departamento de Química Inorgánica, Universidad de Alicante, Alicante, Spain
[15]Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, USA
[*]E-mails for correspondence: rvinuesa@mech.kth.se, j.garcia@ua.es, sbrunton@uw.edu

ABSTRACT

As modern scientific instruments generate vast amounts of data and the volume of information in the scientific literature continues to grow, machine learning (ML) has become an essential tool for organising, analysing, and interpreting these complex datasets. This paper explores the transformative role of ML in accelerating breakthroughs across a range of scientific disciplines. By presenting key examples – such as brain mapping and exoplanet detection – we demonstrate how ML is reshaping scientific research. We also explore different scenarios where different levels of knowledge of the underlying phenomenon are available, identifying strategies to overcome limitations and unlock the full potential of ML. Despite its advances, the growing reliance on ML poses challenges for research applications and rigorous validation of discoveries. We argue that even with these challenges, ML is poised to disrupt traditional methodologies and advance the boundaries of knowledge by enabling researchers to tackle increasingly complex problems. Thus, the scientific community can move beyond the necessary traditional oversimplifications to embrace the full complexity of natural systems, ultimately paving the way for interdisciplinary breakthroughs and innovative solutions to humanity's most pressing challenges.
Keywords: machine learning (ML); deep learning (DL); artificial intelligence (AI); scientific discovery; complexity; physics; life sciences; computer science

Introduction

Machines have played a critical role in scientific discovery (ι.e. to obtain fundamental and formalized knowledge about Nature) by providing the tools to observe, measure, and analyze natural phenomena. Scientific instruments, such as telescopes and microscopes, have historically enabled groundbreaking discoveries by revealing details invisible to the naked eye, expanding our understanding of the universe and the microscopic world [1]. With the advent of modern scientific instruments, including DNA sequencers, astronomical observatories, and high-resolution imaging devices, research facilities are producing terabytes or even petabytes of information. As data volumes grow, computers play a critical role in organizing, analyzing, and interpreting this information. Advanced computational methods help to reduce the complexity of the data, making it possible to extract meaningful insights [2, 3]. However, even with the most advanced computers, the sheer volume of data generated by large-scale projects such as the Large Hadron Collider (LHC) and the Square Kilometre Array (SKA), and the vast amount of

information available in the scientific literature, make traditional analysis methods impractical. Complex problems such as weather forecasting, drug discovery, and genomic analysis often involve highly complex data sets and processes that cannot be efficiently managed without the assistance of machine learning (ML), which can help sift through massive data streams, identify patterns, and extract valuable insights that would be impossible for humans or traditional computational methods alone. Complexity in these problems arises from non-linearity, high dimensionality, and multiscale dynamics, posing significant challenges for traditional mathematical tools and even recent simulation-based approaches. Despite advancements in simulations and big data, we still struggle to fully understand phenomena like turbulence or biological processes at a deeper level. Advanced ML tools are also improving decision-making, enabling faster and more accurate interpretations of complex phenomena, and addressing challenges in diverse scientific fields [4]. However, it also introduces some challenges and the need for ethical guidelines to ensure the appropriate use of ML for scientific research [5]. A key issue is algorithmic bias, which can distort outcomes and lead to incorrect conclusions, particularly in areas like health care, where biased predictions can impact patient care or drug development. Additionally, the black-box nature of ML models complicates transparency, making it hard for researchers to verify decisions. Lastly, data ownership and privacy concerns arise, especially with sensitive data like genetic or health records. These issues require the development of robust ethical guidelines to ensure that ML contributes to science in a fair, transparent, and socially responsible way [6].

In this work, we explore the potential of ML and artificial intelligence (AI) in three types of scientific problems: (i) those where all governing equations are known, (ii) those with partial knowledge and (iii) those where little is known. We illustrate this with examples from the physical and life sciences, including turbulent flows, dark matter, drug discovery, and brain research. Figure 1 summarizes how different uses of ML help address complexity in these areas. As discussed below, when the system complexity grows, the distinction between full, partial, and no knowledge of the governing equations becomes increasingly blurred, but ML plays a crucial role in tackling these challenges. A good example of a complex system with vast amounts of data that traditional tools cannot process efficiently is brain research. ML enables the reconstruction of countless brain slices into highly accurate three-dimensional (3D) maps. In a recent study, Google researchers used AI to process 300 million brain images from Harvard, creating the largest-ever interactive 3D brain tissue model, now available online. This innovation is crucial for understanding neurological disorders, as ML can detect patterns that traditional methods might miss, supporting early diagnosis and treatment planning [7]. One of the most widespread applications of ML is in drug discovery, addressing the time and cost challenges of traditional methods. The drug-development process can take over a decade and billions of dollars due to the complexity of identifying viable candidates. ML is transforming this by rapidly analyzing vast biological and chemical data, uncovering patterns that might remain hidden, and streamlining the identification of promising candidates. Additionally, ML enhances predictive modeling, allowing researchers to forecast drug efficacy and safety earlier in the process. By enabling virtual screening of drug-target interactions, ML reduces the need for costly lab experiments and helps design more efficient clinical trials, accelerating development and optimizing resources [8].

Figure 2 provides a comprehensive visualisation of the key themes discussed in this paper, focusing on the role of machine learning (ML) in scientific discovery and the challenges it presents, particularly with respect to data. The top-left panel illustrates the significant increase in data generated by modern scientific instrumentation over time. Initially, humans organised and interpreted the data through manual analysis and generalisation. Since then, computational methods have facilitated the management of large amounts of data, contributing significantly to fields such as genomics, chemistry and mathematics. More recently, with the advent of large-scale scientific facilities such as CERN, ML techniques have become essential for processing vast amounts of data. This shift has reduced human involvement in direct observation and interpretation, raising concerns about our diminishing understanding of the discoveries made by these technologies. The top-right panel outlines four key challenges related to ML in science: data quality and availability, inherent biases, explainability, and the risk of overfitting. The bottom-left panel highlights a conceptual dilemma: while ML accelerates discovery, there is growing debate about what constitutes a true scientific breakthrough, and whether ML can only "rediscover" existing concepts rather than uncover new insights. Finally, the bottom-right panel emphasises the increasing need for high-quality data to enable new discoveries through ML, while highlighting the limitations imposed by our current gaps in scientific knowledge. All of these issues are discussed in detail in the following sections of this paper.

Artificial Intelligence is playing a transformative role in physics by improving data analysis, model development, and experimental interpretation. In astronomy, ML is improving the search for exoplanets by boosting the accuracy and efficiency of data analysis. AI-powered algorithms, particularly convolutional neural networks, can process massive data sets from telescopes to detect Earth-like exoplanets in noisy signals more precisely than traditional
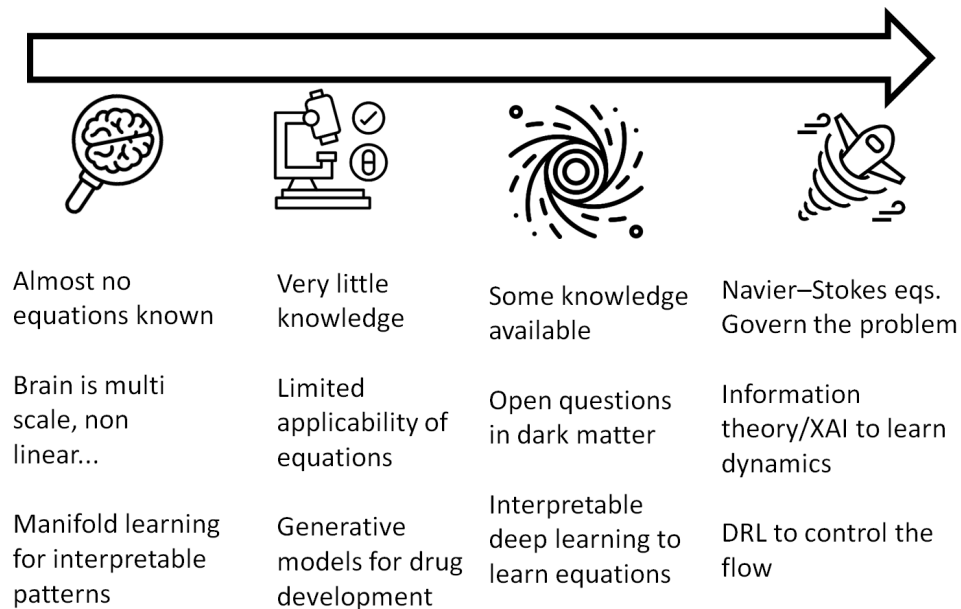
HOW MUCH WE KNOW

Almost no equations known	Very little knowledge	Some knowledge available	Navier–Stokes eqs. Govern the problem
Brain is multi scale, non linear...	Limited applicability of equations	Open questions in dark matter	Information theory/XAI to learn dynamics
Manifold learning for interpretable patterns	Generative models for drug development	Interpretable deep learning to learn equations	DRL to control the flow

Figure 1. Schematic representation of the various applications of ML for scientific discovery, depending on the amount of knowledge available in each category. A number of examples are provided, including brain research, drug discovery, dark matter and fluid mechanics.

methods. The transit method, which detects exoplanets by observing mini-eclipses as they pass in front of their stars, can be complicated by planetary interactions that disrupt periodicity. To address this, researchers from the University of Geneva, University of Bern, and Disaitek applied ML and image recognition techniques to predict these interactions. By training a neural network on numerous examples, they built a model capable of detecting subtle exoplanet signals that might be missed by traditional methods. Their work led to the discovery of exoplanets Kepler-1705b and Kepler-1705c, advancing our understanding of planetary systems [9]. ML is also crucial in areas like the Standard Model of particle physics [9]. Automated algorithms were central to the discovery of the Higgs boson [10], and future experiments will need to operate at higher energies and intensities, generating data volumes too large for traditional methods to handle. For instance, it is expected that the Large-Hadron Collider (LHC) will increase proton collision rates by an order of magnitude in the next decade, requiring data analysis tools such as ML, to identify trends, uncover hidden relationships, and design more effective experiments.

Another scientific area in which ML is increasingly assisting research is mathematics, in this case by improving the process of theorem proving, mathematical method development and discovery. ML systems have already demonstrated their ability to automate aspects of theorem proving; for example, Meta AI's neural theorem prover successfully solved 10 International Math Olympiad (IMO) problems, far exceeding the performance of previous ML systems. Another notable example is DeepMind's collaboration with mathematicians, which has resulted in ML contributing to new mathematical methods in knot theory and representation theory. Specifically, in 2021, ML was used in a new constructive way to suggest mathematical proofs. A collaboration between mathematicians and DeepMind demonstrated that ML can complement human intuition in proving or suggesting complex theorems. The team used ML to investigate long-standing conjectures, including the Kazhdan–Lusztig polynomials, and discovered new connections in knot theory. This illustrates AI's potential to accelerate mathematical research and open new frontiers [11].

Embracing complexity

Machine learning comprises a growing set of algorithms, enabled by high-performance computing and increasingly vast data, that show incredible promise for handling complexity [12, 13]. Neural networks, despite being governed by simple rules, can perform complex tasks for which no traditional algorithms exist. Although they are fully
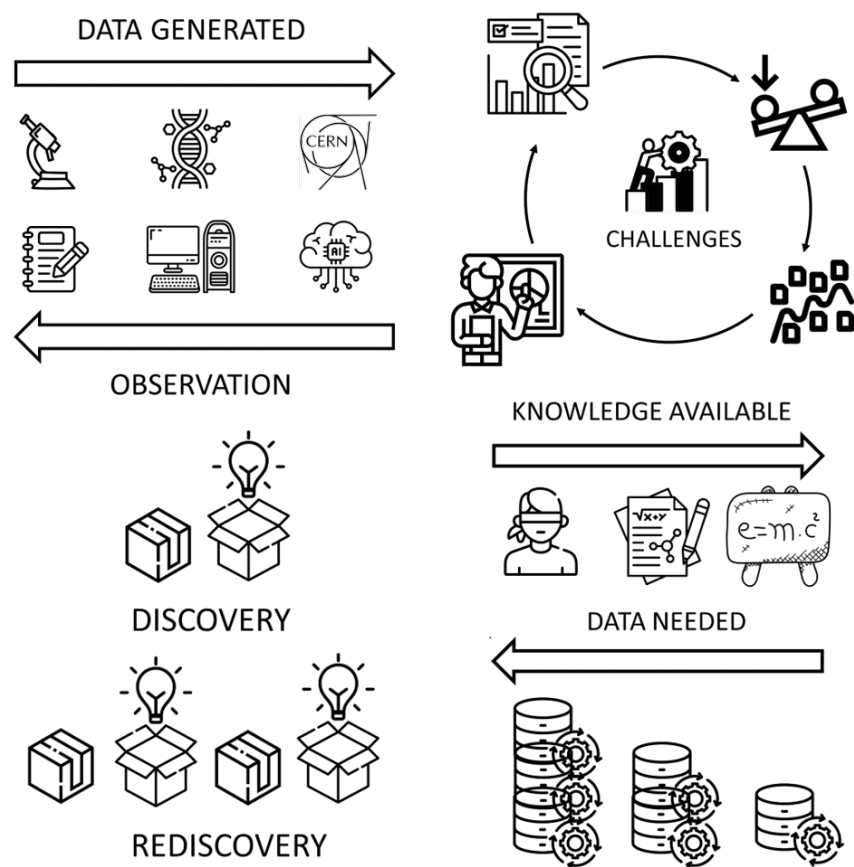
Figure 2. Visual summary of the paper illustrating some of its main ideas. (Top left) The increase in the amount of data generated by scientific instrumentation over time and the shift from data organization by humans, computers, and finally by machine-learning techniques, which translates into less observation, intervention and understanding of humans on scientific discoveries. (Top right) The four challenges we have identified in the paper on the use of ML techniques for scientific discovery, these being: data quality and availability, potential biases, explainability and overfitting. (Bottom left) What constitutes a scientific discovery and the possibility of ML making original breakthroughs versus simply rediscovering known ideas, concepts or laws. (Bottom right) The need to have more and better data to be able to make scientific discoveries with ML as we have less knowledge on the subject of study

observable and deterministic, we often cannot explain their decisions. However, they have led to groundbreaking discoveries, such as a new class of antibiotics [14]. This creates challenges, such as the need for explainable AI (XAI) [15, 16]. Symbolic approaches, such as gene-expression programming, sparse regression, and sparse Bayesian learning [17–19], have been successful, but their complexity grows exponentially with the search-space size. Recent efforts to combine symbolic and deep-learning approaches have enabled advances, such as discovering new materials [20]. This raises fundamental questions about the limits of ML in scientific discovery; for example, can a complex system understand its own complexity? And how much can AI discover beyond its training data [21]? These issues highlight the opportunities and challenges that data-driven methods bring to science. Some believe in the "unreasonable effectiveness of data" [22], particularly in deep learning [23], but the practical implications for future discoveries remain uncertain. A key question is whether ML can provide not only computational solutions but also fundamental scientific understanding. Note that ML's applications in science are not limited to discovery. For example, AI is revolutionizing optimization, laboratory automation, and solving governing equations, such as partial differential equations (PDEs). Recent studies [24] have focused on the potential of self-supervised learning in experiments and simulations (including representations of scientific data), while others [25] are exploring AI's role in formulating scientific questions. Note that the unique contribution of this work is the study of how ML can tackle complexity to reach scientific discoveries, acknowledging that different levels of knowledge of the governing

equations (see Figure 1) will require completely different ML approaches. This may constitute a revolution in how we organize disciplines when using ML for scientific discovery, where apparently different communities may share many similarities in terms of how much is known regarding the governing equations and therefore in terms of the ML methods to be developed.

The emergence of scientific foundation models (SFMs) and large language models (LLMs) is further pushing the boundaries of ML methods. Foundation models are large machine-learning or deep-learning generative models trained on vast amounts of data so they can be applied on a wide range of cases. They predict masked data regions to learn associations, excelling in multiple tasks without large, labeled datasets. Despite risks like bias and transparency issues [26], foundation models have revolutionized AI, especially through chatbots like ChatGPT [27]. LLMs now assist with tasks from writing and coding to guiding scientific experiments [28, 29] and generating ideas [30]. Foundation models also show promise in non-text-focused areas, such as protein-structure prediction [31], protein design [32] and climate simulations [33]. An intriguing feature of foundation models, and particularly LLMs, is that they demostrate so-called "emergent abilities" [34, 35]. The term refers to unexpected, not explicitly programmed, capabilities that arise as model scale increases, and do not consist in mere extrapolations of smaller models' performance. Such abilities were first observed in complex natural systems, e.g. phase changes in materials, complex behavior in flocks of birds or fishes, as well as in computational systems such as cellular automata and agent-based models, and the term was originally popularized by the Nobel-prize winner P.W. Anderson [36]. Examples of LLM emergent abilities include in-context learning, complex reasoning, and multi-step problem-solving, which are highly valuable for scientific research. Thank to this, LLMs are found to generalize to new tasks without prior examples (zero-shot learning) or with minimal data (few-shot learning), making them particularly useful in domains where data is scarce or highly specialized. More in general, they have the potential to aid hypothesis generation, automate literature reviews, predict protein structures, and accelerate scientific discovery, especially when combined with reinforcement learning (RL) and a suitable system of rewards/penalties to optimize specific scientific tasks, such as designing experiments or iteratively refining hypotheses based on feedback. Synergy with scientists may allow to leverage the models' reasoning capabilities in dynamic, real-world scenarios, enabling more efficient exploration of complex scientific problems. On the other hand, these abilities are often unpredictable and non-linear, raising challenges in reliability, interpretability, and ethical use. Careful oversight is then needed to ensure these models align with scientific goals and produce trustworthy results. Risks include the potential for generating misleading or incorrect outputs, amplifying biases present in training data, and over-reliance on automated systems without sufficient human validation. Additionally, the opacity of how these models arrive at their conclusions can hinder their adoption in critical scientific applications, raising important questions about their benefits and risks in science [37]. Fajardo-Fontiveros *et al.* [38] discuss when it is possible to learn models from data and the acceptable noise levels for accurate learning.

Discovery versus re-discovery by machine learning

While ML has proven invaluable in refining existing knowledge, its real potential lies in detecting patterns and correlations that may not be immediately apparent due to the vast amounts of data involved. This raises the question of whether ML is truly discovering new insights or merely reinterpreting existing knowledge. Historically, scientific discovery has been rooted in inquiry, observation, and experimentation, with creativity playing a crucial role in generating new insights into phenomena or theories [39, 40]. While ML excels at analyzing large datasets, identifying patterns, and generating hypotheses (which are key components of discovery), its ability to make truly autonomous, original discoveries is still debated. ML models operate without a genuine understanding of the underlying mechanisms they analyze, meaning that their "discoveries" tend to be instrumental rather than original. Groundbreaking discoveries usually require creativity, broader contextual understanding and sometimes leaps beyond available data, all of which ML currently lacks. Although ML is revolutionizing research by uncovering complex trends, the idea that it can independently produce original scientific discoveries remains controversial. Several studies have explored this question [41], and workshops, such as the one organized by the National Academies on How ML Is Shaping Scientific Discovery, have debated the issue [42]. Nevertheless, ML has already contributed to significant scientific breakthroughs. In drug discovery, ML identified new antibiotics, such as halicin, which were unknown and can kill antibiotic-resistant bacteria [43]. ML also plays a crucial role in materials science, predicting new materials for batteries and superconductors using physical and experimental data. These achievements highlight AI's potential to navigate uncharted scientific territory [44].

One of the most significant advancements in AI-driven scientific discovery is AI-Hilbert. Developed by IBM researchers, AI-Hilbert acts as an "AI scientist" that transforms existing theories and data into new, consistent mathematical models. Its goal is to accelerate scientific discovery by automating hypothesis generation and testing.

AI-Hilbert helps scientists uncover new knowledge by analyzing large scientific datasets and revealing patterns overlooked by traditional methods. It also refines theories by managing conflicting data [45]. Furthermore, Cornelio *et al.* [46] have developed a new ML tool which combines axiomatic knowledge with experimental data to derive scientific models. By integrating logical reasoning and symbolic regression, it has rediscovered laws like Kepler's third law and Einstein's time-dilation law. This tool can distinguish between competing formulas, even with limited data. While efficient at replicating human discoveries, these tools often confirm known theories rather than offering new insights. ML can be biased towards existing patterns, limiting its ability to generate novel hypotheses. In these cases, AI's role is more supportive, validating established knowledge rather than challenging it [47]. ML holds immense potential for making groundbreaking discoveries, particularly in fields like drug development and astrophysics. However, its frequent rediscovery of known scientific principles illustrates the current limitations of its creativity. For ML to drive new knowledge, it must evolve beyond confirming human findings. Currently, ML lacks an intrinsic understanding of the mechanisms it analyzes, so its discoveries often complement human expertise, relying on human interpretation and creativity to fully appreciate and exploit the insights gained.

Machine-learning-driven scientific discovery when complete information is available

There are several applications within Physical Sciences where the underlying governing equations are known perfectly but the high-level global dynamics are still not well understood. For instance, while we know the quantum equations behind biomolecular dynamics, complexity makes biology a partial-knowledge problem since only limited biological systems can be measured or simulated, and simulating the full human brain is impossible. Similarly, turbulent flows, described by the Navier–Stokes equations, are only partially understood due to their chaotic nature as the Reynolds number increases [48]. While the large scales in the flow can be simulated, the smaller scales are much more difficult to simulate, making turbulence a major challenge [49]. However, some ML techniques, such as symbolic regression and reduced-order modeling, can help uncover unknown flow features and physical properties from large direct-numerical-simulation (DNS) datasets [50, 51], as indicated in Figure 3. Advances in turbulence modeling using supervised ML have also improved closure models for Reynolds-averaged Navier–Stokes (RANS) and large-eddy-simulation (LES) turbulence models [52, 53]. In astrophysics, the supervised classifier SPOCK predicts long-term stability in multi-planet systems (which requires integration of the laws of gravitation over billions of orbital periods) using short-term simulations and effectively generalizing to larger systems [54].

Optimal control of complex physical and biological systems is another challenging area where ML, particularly deep reinforcement learning (DRL), shows promise. DRL has led to breakthroughs in quantum physics, astronomy, turbulence control, and tokamak-instability control, offering insights into complex systems and discovering novel strategies [55]. DRL has even uncovered previously unknown thermodynamic cycles [56]. ML can dramatically accelerate simulations of complex systems with known equations. Autoencoders can be used to discover latent-space representations that enable faster time integrators and simulations, leading to better optimization and systematic studies [57, 58]. For instance, high-energy physics relies on comparing observed particle detector data with simulations, which are computationally intensive. Fast generative models like generative adversarial networks (GANs) and variational autoencoders (VAEs) offer a faster alternative, although on-going research aims at ensuring that the required accuracy is achieved [59, 60]. In climate science, foundation models for weather forecasting are revolutionizing climate studies, offering potential breakthroughs in understanding climate change and paleoclimates [13]. In applied mathematics, DRL and large language models (LLMs) are generating novel algorithms and optimizations. Notable examples include matrix operations [61] and combinatorial problems, where AI aids in discovering new strategies that can be validated with classical algorithms [62]. In these cases, AI provides valuable heuristic methods, although it must be highlighted that finding adequate solutions still remains challenging.

Machine-learning-driven scientific discovery when only partial information is available

In contrast to systems with well-understood governing equations, in some scientific problems we only have access to partial knowledge of the underlying mechanisms, as seen in complex materials (*e.g.*, composites or textured materials) and certain fluids (*e.g.*, granular or multiphase flows). These systems can exhibit simple microscopic behaviors yet result in complex macroscopic phenomena, such as the spread of infectious diseases or "active turbulence" in biological matter [63, 64]. In these cases, inductive biases, like frame invariance or symmetry constraints, can be incorporated into ML models to improve the discovery process [65, 66]. Physical constraints (*e.g.*, thermodynamics) help create more generalizable models. For example, Moya *et al.* [67] proposed a neural
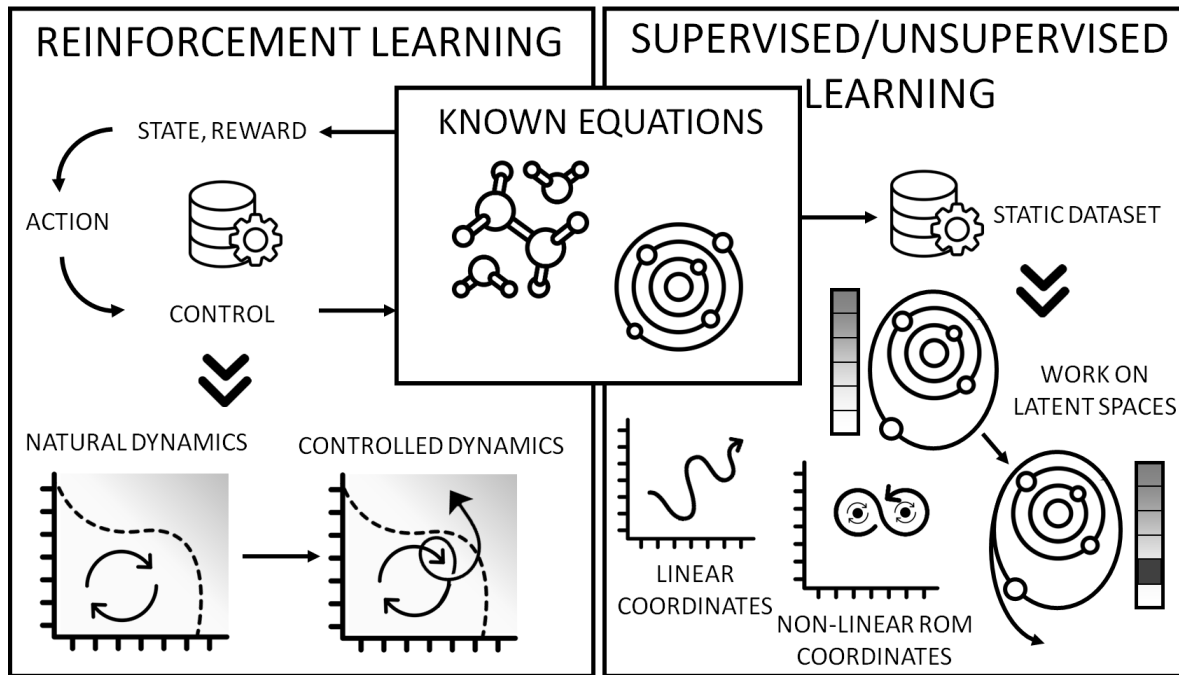
Figure 3. Schematic representation of ML directions to enable scientific discoveries when complete information about the governing equations is available. In such a case, both supervised, unsupervised, and reinforcement-learning methodologies can be used. Supervised and unsupervised methodologies are made possible by generating large datasets of synthetic data simulated from the governing equations. This allows the deployment of a variety of ML techniques that can discover complex hidden relations, nonlinear coordinate systems, hidden dynamics or solve otherwise intractable problems. Reinforcement learning can also be used by coupling it to the physics simulator, which has already proven successful at discovering previously unknown control strategies and regimes of complex systems or generating high-quality heuristic guesses that can be tested in the case of problems where solution verification is easy, but the suggestion of good candidate solutions is hard.

network constrained by known thermodynamic properties, enabling broader applications across physical systems. It is also important to highlight digital twins, where only some data are available, and they combine data-driven aspects with simulations while preserving generalization properties [68]. Another example of taking advantage of known physical properties of the system is embedding symmetries in autoencoders, intending to develop reduced-order models (ROMs) in physical systems invariant to input transformations [69, 70]. These data-driven and physics-informed techniques are crucial for scientific discovery; for instance, ML is helping to derive constitutive laws for materials with complex rheologies. De Lorenzis and collaborators [71] introduced a hybrid framework (EUCLID) to learn constitutive equations for hyperelastic solids, and the SpaRTA framework for data-driven turbulence modeling [72] has been adapted to model elastic solids [73]. Such approaches have also been used to infer constitutive equations from crystal structures and rheology of complex fluids [74] (see Figure 4).

In the context of Life Sciences, structural biology is a field where ML has made significant advances despite partial knowledge of the phenomena. AlphaFold [75], for instance, folds proteins into their three-dimensional (3D) native form from a one-dimensional(1D) amino acid sequence using deep learning with embedded biases like multiple sequence alignment (MSA) and 3D equivariance. AlphaFold has also spurred innovations in ML, including single-sequence methods like ESMfold [76] and generative models for protein design [77, 78]. Diffusion models are also being used to generate protein-backbone structures and molecular ensembles, bypassing the need for simulations [79, 80]. Furthermore, generative models have also been used to discover physical models, integrating prior knowledge with data to generalize to new scenarios [81]. Transformers, for instance, are now used in Chemistry to complete chemical reactions and predict reaction yields [82]. When it comes to quantum systems, deep reinforcement learning (DRL) has enabled new approaches to manipulate quantum states, providing insights into quantum mechanics [83]. Additionally, ML aids in reducing noise in quantum-computing systems, while quantum computing improves ML performance [84]. In climate modeling, ML has developed new LES
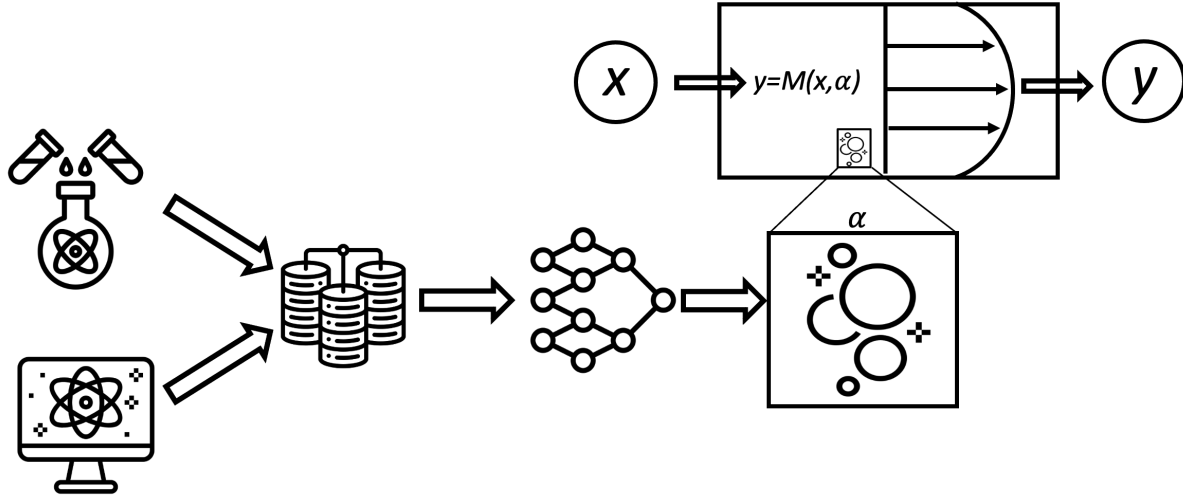
Figure 4. Example of machine learning applied to a case where partial knowledge is available about the underlying system, illustrating a model (for instance a flow with complex rheology or a flow through a porous medium, top right of the picture) which depends on a set of known inputs **x** (*e.g.* geometry, boundary conditions, *etc.*) as well as on a set of hidden (unobservable) variables $\boldsymbol{\alpha}$ describing, *e.g.*, the fluid constitutive behavior. The latter may involve small-scale phenomena that can be difficult or impossible to describe. In such conditions, experimental or numerical data for observable quantities (*e.g.* velocity fields or stresses **y**) can be used to infer the unknown field by training a machine learning model (here represented as a neural network, although other ML approaches are possible), subjected to physical constraints (*e.g.* positivity, symmetries or invariances). The whole process allows, on the one hand, to train a data-driven closure model for the hidden variables $\boldsymbol{\alpha}$ and, on the other hand, to gain *a-posteriori* physical knowledge of the fluid constitutive properties.

models to ensure stable long-term forecasts [85]. ML has also been shown to enhance traditional weather-prediction systems [86].

Machine-learning-driven scientific discovery when little information is available

There are numerous phenomena across scientific disciplines whose origins and underlying principles remain elusive. In these cases, the absence of well-established governing equations or foundational physical models makes it difficult to fully capture and understand their critical dynamics. For example, neuroscience has no first-principle equations, as there are no known conservation laws or symmetries to derive generalizable differential equations. Even with equations describing molecules or cells, the brain's complexity makes full-scale simulation unfeasible. Despite this, advances in neural data acquisition, such as large-scale neural recordings and connectomics, produce unprecedented datasets, promising a new era of ML-driven discovery in neuroscience and behavior [87, 88]. Although full brain simulations remain beyond reach, data-driven models can replicate key input-output relationships, generating predictions vital for discovery. In neuroscience, perturbation experiments (such as activating neuron populations during visual tasks) provide insights into visual perception. Data-driven models can generate testable hypotheses and refine their predictions iteratively through experiments. ML also synthesizes diverse experimental data, such as the MICrONS dataset, which links functional imaging with structural reconstructions of cortical neurons [89].

ML can learn dynamics where no a-priori knowledge exists, as seen in the Hodgkin–Huxley equations modeling neural dynamics [90]. Approaches like SINDy (sparse identification of nonlinear dynamics) [19], genetic algorithms [91] and reinforcement learning [92] can help uncover system dynamics. Neural ordinary differential equations (NODEs) [93] are another method for modeling continuous dynamics, although they often lack scientific insight. Hybrid models relying on transformers [94] or SINDy-inspired architectures [95] aim to bridge this gap. Interpretable-ML models can also discover biomarkers for diseases or predict treatment outcomes from patient data [96]. In systems lacking governing equations, interventional data like CRISPR-Ko experiments in single-cell biology [97] are enabling ML to identify some of the underlying mechanisms. Automated setups are advancing[98], leading to research in designing experiments for system identification, especially for non-linear

models like NODEs [99]. Representation-learning techniques, such as variational autoencoders [100], are essential for simplifying the characterization of complex systems by identifying latent spaces that reduce dimensionality and reveal causal structures [101]. Causality and its application to dynamical systems have gained prominence [102], especially in Earth sciences [103] and molecular biology [104], with some studies using invariance from heterogeneous experiments as a signal to identify causal ODEs [105]. Learning structured latent spaces is of crucial importance since it provides an effective coordinate system in which the dynamics have a simple representation, which is a key requirement for generalization and interpretability [106]. In Figure 5 we provide a schematic representation of the identification of an underlying causal structure from observations where the variables of interest are not directly observed.
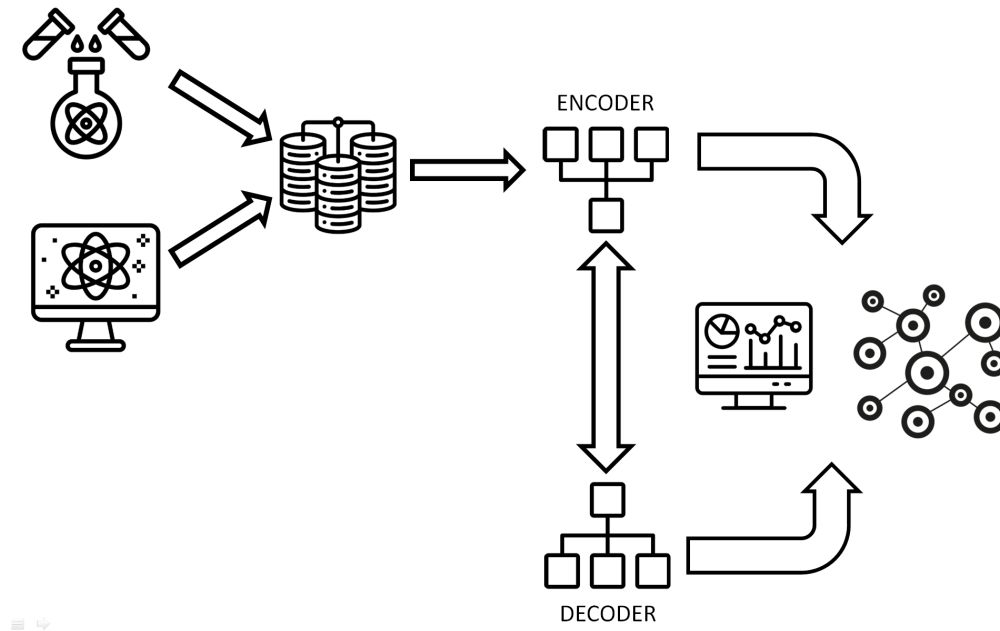
Figure 5. Schematic representation of a model (for instance, the observed symptoms of an unknown or complex disease within a population, or observed opinion dynamics within a social network) where the behavior as observed in data depends on an unknown dynamic or causal structure. The observed behavior or dynamics might occur on several different spatial and temporal scales, and the observed data might reflect more or fewer aspects of the underlying system. In such conditions, *representation-learning* methods can be employed to distil out an explanation of the observed data in the form of a system of ODEs or as a causal-graph representation.

Recent research has also explored improving diffusion models and solvers [107], which are useful for generating state-of-the-art results in areas like image generation [108], protein modeling [109] and materials science [110]. Despite their success, large pre-trained models provide limited scientific insights [111], and therefore constitute an opportunity for future research. Machine learning has also enhanced data collection and processing in systems with indirect or incomplete measurements. ML imputation and generative modeling can complete time-series data, improving downstream applications [112]. Furthermore, computer-vision techniques have also automated previously manual tasks, such as segmentation in microscopy, enabling large-scale analysis of cell populations [113]. In ethology, ML has transformed video data into animal kinematics and poses, linking behavior to underlying neural computations [114]. Notably, large language models (LLMs) and scientific foundation models (SFMs) are opening new avenues for extracting scientific insights directly from data. Genome-scale language models (GenSLMs), for instance, are helping to learn the evolutionary landscape of SARS-CoV-2 genomes [115], while LLMs are enhancing neuroscience research by integrating diverse datasets and summarizing insights across isolated subfields [116].

In sum, it is hard to overstate the ongoing impact of machine learning as a critical tool that, when used in conjunction with other approaches (*e.g.* experiments, causality analysis, development of an adequate coordinate system, *etc.*), catalyzes advances in scientific fields where no information on the phenomenon under study is available.

The drawbacks, limitations and challenges of machine learning for scientific discovery

Despite the significant advances that ML has brought to scientific discovery, there are key areas that need to be addressed to accelerate its contributions and realize its full potential while minimizing some clear risks, as discussed below:

- Data-related challenges: The success of ML in scientific discovery relies on large high-quality, structured datasets, but scientific data is often incomplete, noisy, or imbalanced, leading to biased models. Unstructured data, especially in fields like biology, chemistry, and geology, complicate the use of ML applications, since these algorithms are not inherently designed to handle such data. However, it was recently shown that foundation models could help for prediction tasks in small general datasets [117]. Furthermore, the absence of labeled data complicates the use of supervised-ML techniques [118].

- Bias and ethical issues: ML models are prone to data bias, distorting results and hindering scientific discovery. Bias in data collection or model training can reinforce existing assumptions instead of revealing novel insights [119]. This is especially concerning in fields such as drug discovery, where existing models might overlook demographic groups, making the discoveries less generalizable [120]. Ethical concerns, especially in medicine, highlight risks in applying ML to decisions affecting human life [121].

- Explainability and interpretability: ML models, particularly those based on deep learning, often function as "black boxes", making accurate predictions without revealing their decision-making processes [16]. This lack of transparency is problematic in fields like drug discovery, where understanding why a model predicts a certain interaction is crucial [122]. Advances in explainable AI (XAI) have improved interpretability in areas like materials science [123], chemistry [124], and medicine [125].

- Overfitting and generalization: ML models often overfit their training data, performing poorly on unseen data, a fact that limits their ability to generalize. In scientific contexts, this can produce misleading results and fail to capture complex, nonlinear relationships, as seen in chemistry, biology, and astronomy [126–128]. Overfitting restricts the potential of ML to develop universal theories essential for broad scientific understanding [129].

Improving data quality and diversity is critical for ML models to generalize across disciplines. Initiatives like the Open Reaction Database [130] and the Crystallography Open Database [131] aim to enhance data management under the FAIR (findability, accessibility, interoperability and reusability) principles [132, 133]. Incorporating bias detection and fairness-aware algorithms can reduce biased data impacts [134]. XAI methods are being developed to improve the transparency of ML models, particularly in healthcare applications like brain-tumor segmentation [135]. Hybrid approaches that combine ML with physics-based models yield promising results by ensuring that the predictions adhere to known scientific principles [136, 137]. It is also important to note that ethical frameworks are also needed to ensure fairness, transparency, and accountability, especially in medicine [138, 139].

While AI has shown great potential in scientific discovery, significant challenges remain. Issues related to data quality, bias, interpretability, and overfitting must be addressed to harness the full potential of AI in advancing science. By improving data access, enhancing model transparency, and developing hybrid models, the scientific community can overcome these obstacles and drive meaningful, trustworthy discoveries using AI which may play an instrumental role in areas as far-reaching as gravitational waves [140], space exploration [141] or even the discovery of extraterrestrial life [142].

Conclusions and outlook

Karl Popper famously stated in *The Open Universe: An Argument for Indeterminism from the Postscript to The Logic of Scientific Discovery* that "science can be described as the art of systematic oversimplification" [143] and this remains true due to the historical limitations in processing and analyzing vast amounts of data. As a result, scientists have been forced to simplify their objects of study by focusing on isolated phenomena or simple models. While simplification has benefits, for example in the teaching and the systematization of science, oversimplification carries significant risks, as essential information and key aspects of problems may be lost in the process. For example, in ecological research, studies often focus on individual species interactions without considering the broader dynamics of ecosystems or other effects, such as climate change [144].

ML methods have already enabled several scientific and technological advancements, from solving image classification to winning at Go. This article highlights how modern data-driven methods are enabling breakthroughs in scientific discovery, focusing on state-of-the-art techniques that push beyond previous limitations. We categorize

these approaches by how much knowledge about the underlying mechanisms is available, ranging from well-understood systems to those with unknown governing principles. As shown in Table 1, ML's potential spans a wide spectrum of applications, including discovering physical laws, evaluating complex systems, inferring unknown behaviors, and uncovering multiscale principles. These advances apply across fields like Physics, Mathematics, Chemistry, and Life Sciences, promising faster scientific progress than ever before. However, using ML for scientific discovery brings challenges. A key advantage of ML is its ability to model complex systems, but this often requires extensive training data. In areas like astrophysics, rare diseases, or new pharmaceuticals, data scarcity is a frequent obstacle. Fortunately, complementary ML methods can either generate needed data or bypass large datasets through techniques like self-supervised learning. Even when ML does not directly result in discoveries, it can facilitate breakthroughs where data are limited. Validation is another challenge, especially in cases where the governing principles are unknown. But ML-driven discoveries can be confirmed using traditional scientific methods, such as hypothesis testing, observational confirmation and benchmarking. Additionally, ML models often function as "black boxes", making it hard to derive formal knowledge from their results. This is a significant issue for science, where understanding is key. However, explainable- and interpretable-ML methods offer solutions, helping to achieve discoveries in the context of established scientific principles. Despite these challenges, ML is becoming an essential tool across disciplines, and its continued evolution promises even more opportunities for scientific discovery. With further refinement, ML techniques are likely to address the current limitations, enabling even greater advancements.

The recent Nobel Prizes in Physics [145] and Chemistry [146] underscore the transformative impact of machine learning, demonstrating its ability to revolutionise scientific discovery. By providing cutting-edge tools for data analysis, ML is accelerating breakthroughs across multiple fields, deepening our understanding and tackling complex challenges in ways previously unimaginable. This capability is crucial, as only by embracing this complexity can we connect the dots in science, grasp the intricate interdependencies of natural systems, and facilitate breakthrough, unpredictable discoveries. The real challenge – and opportunity – of AI-driven scientific discovery lies in moving beyond solving narrow, well-defined problems to tackling complex, open-ended questions. This shift calls for the development of artificial general intelligence (AGI) or even artificial superintelligence (ASI), which possesses the potential to engage in multifaceted problem-solving across diverse domains. AGI for scientific discovery could connect seemingly disparate areas of knowledge, leading to breakthroughs that single-focused AI cannot achieve. For instance, an AGI model could integrate insights from pharmacology, neuroscience, and data science to not only propose novel drug candidates but also predict their interactions and effects on complex biological systems. This capacity for interdisciplinary connections would facilitate original contributions to scientific knowledge, allowing for innovative solutions to intricate challenges. Ultimately, harnessing AGI's multifaceted problem-solving abilities presents a transformative opportunity to revolutionize how we approach scientific discovery and address some of humanity's most pressing challenges.

References

1. Tucci, P. History of scientific instrumentation and history of science. In Pisano, R. (ed.) *A History of Physics: Phenomena, Ideas and Mechanisms*, vol. 42 of *History of Mechanism and Machine Science*, DOI: 10.1007/978-3-031-26174-9_6 (Springer, Cham, 2025).

2. Szalay, A. & Gray, J. Science in an exponential world. *Nature* **440**, 413–414, DOI: 10.1038/440413a (2006).

3. Eisenstein, M. Big data: The power of petabytes. *Nature* **527**, S2–S4, DOI: 10.1038/527S2a (2015).

4. Huang, L. & Peissl, W. Artificial intelligence—a new knowledge and decision-making paradigm? In Hennen, L. *et al.* (eds.) *Technology Assessment in a Globalized World*, DOI: 10.1007/978-3-031-10617-0_9 (Springer, Cham, 2023).

5. Resnik, D. & Hosseini, M. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. *AI Ethics* DOI: 10.1007/s43681-024-00493-8 (2024).

6. UNESCO. Recommendation on the ethics of artificial intelligence. *SHS/BIO/PI/2021/1* 1–43 (2022).

7. Badrulhisham, F., Pogatzki-Zahn, E., Segelcke, D., Spisak, T. & Vollert, J. Machine learning and artificial intelligence in neuroscience: A primer for researchers. *Brain, Behav. Immun.* **115**, 470–479, DOI: 10.1016/j.bbi.2023.11.005 (2024).

LEVEL OF KNOWLEDGE	MODE OF DISCOVERY	METHODS	APPLICATIONS
COMPLETE INFORMATION *Systems with pre-existing deterministic governing equations*	Enabling	Supervised ML - Symbolic regression - Autoencoders - Reduced order models - Foundation models Unsupervised ML	Computational speed-up - Training data generation - Case evaluation - High through-put screening Parametric system sweep across complex systems (chemical compound spaces, biological system landscapes, …) Experimental automation and synthesis
	Revealing	Explainable ML DRL	Fluid mechanics - Turbulence dynamics - Flow control Material science - Stability predictions - Compound discovery - Self-correcting processing Quantum physics - Ground state predictions - Chemical reactivity Algorithmic optimization Molecular dynamics
PARTIAL INFORMATION *Systems with known input behaviour, but where model complexity prohibits complete knowledge*	Enabling	Generative modelling DRL	Shifting known equations into unknown settings - Fluid mechanics - Structural biology - Drug development - Material science and compound discovery - Quantum physics
	Revealing	Explainable ML Physics-constrained learning	Discovery of analytical equations of complex systems - Structural biology - Constitutive material modelling - Multiscale systems
	Inducing	Reduced order models Physics- constrained learning	Digital twins of known systems - Drug development - Material science and compound discovery - Fluid mechanics - Quantum physics and chemistry (exploration of large-scale chemical compound space)
LITTLE INFORMATION *Systems with complex multiscale nature going beyond current deterministic knowledge*	Causal inference and induction	Interpretable ML Explainable ML DRL Foundation models	Multidisciplinary - Multiorgan systems - Multiscale systems - Multidomain systems Biomarker identification

Table 1. Summarizing overview of the opportunities for machine learning in scientific discovery. Based on the differentiation presented in our work, the level of prior, deterministic knowledge (*left*) can be used to differentiate methods (*second right*) and applications (*right*) across which scientific advancements can be made by means of dedicated AI systems. This also allows for various modes of discovery (*second left*), ranging from cases where machine learning enables discovery by allowing for efficient computational usage, parametric sweeps, etc., to cases where machine learning is used to causally infer underlying mechanistic behaviours in complex multidisciplinary systems.

8. Hutson, M. How ai is being used to accelerate clinical trials. *Nature* **627** (2024).

9. Leleu, A. *et al.* Alleviating the transit timing variation bias in transit surveys - i. rivers: Method and detection of a pair of resonant super-earths around kepler-1705. *Astron. & Astrophys.* **655**, A66, DOI: 10.1051/0004-6361/202141471 (2021).

10. Radovic, A., Williams, M., Rousseau, D. *et al.* Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560**, 41–48, DOI: 10.1038/s41586-018-0361-2 (2018).

11. Davies, A., Veličković, P., Buesing, L. *et al.* Advancing mathematics by guiding human intuition with ai. *Nature* **600**, 70–74, DOI: 10.1038/s41586-021-04086-x (2021).

12. Degrave *et al*, J. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602**, 414–419 (2022).

13. Wong, C. How AI is improving climate forecasts. *Nature* (2024).

14. Wong, F. *et al.* Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185, DOI: 10.1038/s41586-023-06887-8 (2024).

15. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

16. Vinuesa, R. & Sirmacek, B. Interpretable deep-learning models to help achieve the Sustainable Development Goals. *Nat. Mach. Intell.* **3**, 926 (2021).

17. Ferreira, C. *Gene expression programming: mathematical modeling by an artificial intelligence*, vol. 21 (Springer, 2006).

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

19. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**, 3932–3937 (2016).

20. Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85, DOI: 10.1038/s41586-023-06735-9 (2023).

21. Leslie, D. Does the sun rise for ChatGPT? Scientific discovery in the age of generative AI. *AI Ethics* 1–6 (2023).

22. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE intelligent systems* **24**, 8–12 (2009).

23. Sejnowski, T. J. The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci.* **117**, 30033–30038 (2020).

24. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).

25. Zenil, H. *et al.* The future of fundamental science led by generative closed-loop artificial intelligence. *Prepr. arXiv:2307.07522* (2023).

26. Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R. & Bockting, C. L. Chatgpt: five priorities for research. *Nature* **614**, 224–226 (2023).

27. OpenAI. Chatgpt: Optimizing language models for dialogue. *https://openai.com/blog/chatgpt/* (2022).

28. M Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat Mach Intell* **6**, 525–535, DOI: 10.1038/s42256-024-00832-8 (2024).

29. Boiko, D., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578, DOI: 10.1038/s41586-023-06792-0 (2023).

30. Wang, Q., Downey, D., Ji, H. & Hope, T. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259* (2023).

31. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

32. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *ICML* DOI: 10.1101/2022.04.10. 487779 (2022).

33. Chen, S., Long, G., Jiang, J., Liu, D. & Zhang, C. Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014* (2023).

34. Wei, J. *et al.* Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

35. Schaeffer, R., Miranda, B. & Koyejo, S. Are emergent abilities of large language models a mirage? *Adv. Neural Inf. Process. Syst.* **36** (2024).

36. Anderson, P. W. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science* **177**, 393–396 (1972).

37. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).

38. Fajardo-Fontiveros, O. *et al.* Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **14**, 1043 (2023).

39. Bechtel, W. & Richardson, R. *Discovering Complexity* (Princeton University Press, Princeton, 1993).

40. Ben-Menahem, Y. *Causation in Science* (Princeton University Press, Princeton, 2018).

41. Krenn, M., Pollice, R., Guo, S. Y. *et al.* On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769, DOI: 10.1038/s42254-022-00518-3 (2022).

42. of Sciences, N. A. How ai is shaping scientific discovery (2023). AI for Scientific Discovery, Proceedings of a Workshop (2024).

43. Stokes, J. M., Yang, K., Swanson, K. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13, DOI: 10.1016/j.cell.2020.01.021 (2020).

44. AI and the new age of AI (2023).

45. Cory-Wright, R., Cornelio, C., Dash, S. *et al.* Evolving scientific discovery by unifying data and background knowledge with ai hilbert. *Nat. Commun.* **15**, DOI: 10.1038/s41467-024-50074-w (2024).

46. Cornelio, C., Dash, S., Austel, V. *et al.* Combining data and theory for derivable scientific discovery with ai-descartes. *Nat. Commun.* **14**, DOI: 10.1038/s41467-023-37236-y (2023).

47. Bianchini, S., Müller, M. & Pelletier, P. Artificial intelligence in science: An emerging general method of invention. *Res. Policy* **51**, 104604, DOI: 10.1016/j.respol.2022.104604 (2022).

48. Chapman, S. & Cowling, T. G. *The mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction and diffusion in gases* (Cambridge university press, 1990).

49. Fefferman, C. L. Existence and smoothness of the Navier-Stokes equation. *The millennium prize problems* **57**, 67 (2000).

50. Cremades, A. *et al.* Identifying regions of importance in wall-bounded turbulence through explainable deep learning. *Nat. Commun.* **15**, 3864 (2024).

51. Lozano-Durán, A. & Arranz, G. Information-theoretic formulation of dynamical systems: Causality, modeling, and control. *Phys. Rev. Res.* **4**, 023195 (2022).

52. Duraisamy, K., Iaccarino, G. & Xiao, H. Turbulence modeling in the age of data. *Annu. Rev. Fluid Mech.* **51**, 357–377 (2019).

53. Brunton, S. L., Noack, B. R. & Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508 (2020).

54. Tamayo, D. *et al.* Predicting the long-term stability of compact multiplanet systems. *Proc. Natl. Acad. Sci.* **117**, 18194–18205 (2020).

55. Silver *et al*, D. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

56. Beeler, C. *et al.* Optimizing thermodynamic trajectories using evolutionary and gradient-based reinforcement learning. *Phys. Rev. E* **104**, 064128 (2021).

57. Solera-Rico, A. *et al.* β-Variational autoencoders and transformers for reduced-order modelling of fluid flows. *Nat. Commun.* **15**, 1361 (2014).

58. Park, S. *et al.* Optimization of physical quantities in the autoencoder latent space. *Sci. Reports* **12**, 9003 (2022).

59. Goodfellow, J. *et al.* Generative adversarial networks. *Prepr. arXiv:1406.2661* (2014).

60. Albertsson, K. *et al.* Machine learning in high energy physics community white paper. In *Journal of Physics: Conference Series*, vol. 1085, 022008 (IOP Publishing, 2018).

61. Fawzi *et al*, A. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022).

62. Romera-Paredes, B. *et al.* Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).

63. Souza, L. F., Rocha Filho, T. M. & Moret, M. A. Relating SARS-CoV-2 variants using cellular automata imaging. *Sci. Reports* **12**, 10297 (2022).

64. Alert, R., Casademunt, J. & Joanny, J.-F. Active turbulence. *Annu. Rev. Condens. Matter Phys.* **13**, 143–170 (2022).

65. Cranmer, M. *et al.* Discovering symbolic models from deep learning with inductive biases. *34th Conf. on Neural Inf. Process. Syst. (NeurIPS 2020), Vancouver, Can.* (2020).

66. Liu, Z., Chen, Y., Du, Y. & Tegmark, M. Physics-augmented learning: A new paradigm beyond physics-informed learning. *arXiv preprint arXiv:2109.13901* (2021).

67. Moya, B., Badías, A., González, D., Chinesta, F. & Cueto, E. A thermodynamics-informed active learning approach to perception and reasoning about fluids. *Comput. Mech.* **72**, 577–591 (2023).

68. Kapteyn, M. G., Pretorius, J. V. R. & Willcox, K. E. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nat. Comput. Sci.* **1**, 337–347 (2021).

69. Kneer, S., Sayadi, T., Sipp, D., Schmid, P. & Rigas, G. Symmetry-aware autoencoders: s-PCA and s-NLPCA. *Prepr. arXiv:2111.02893v3* (2022).

70. Otto, S. E., Zolman, N., Kutz, J. N. & Brunton, S. L. A unified framework to enforce, discover, and promote symmetry in machine learning. *Prepr. arXiv:2311.00212* (2023).

71. Flaschel, M., Kumar, S. & De Lorenzis, L. Automated discovery of generalized standard material models with euclid. *Comput. Methods Appl. Mech. Eng.* **405**, 115867 (2023).

72. Schmelzer, M., Dwight, R. P. & Cinnella, P. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow, Turbul. Combust.* **104**, 579–603 (2020).

73. Wang, M., Chen, C. & Liu, W. Establish algebraic data-driven constitutive models for elastic solids with a tensorial sparse symbolic regression method and a hybrid feature selection technique. *J. Mech. Phys. Solids* **159**, 104742 (2022).

74. Mahmoudabadbozchelou, M., Kamani, K. M., Rogers, S. A. & Jamali, S. Digital rheometer twins: Learning the hidden rheology of complex fluids through rheology-informed graph neural networks. *Proc. Natl. Acad. Sci.* **119**, e2202234119 (2022).

75. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

76. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

77. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106, DOI: 10.1038/s41587-022-01618-2 (2023).

78. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56, DOI: 10.1126/science.add2187 (2022).

79. Watson, J. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100, DOI: 10.1038/s41586-023-06415-8 (2023).

80. Ingraham, J. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078, DOI: 10.1038/s41586-023-06728-8 (2023).

81. Cornelio, C. *et al.* Combining data and theory for derivable scientific discovery with ai-descartes. *Nat. Commun.* **14**, 1777 (2023).

82. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**, 015022 (2022).

83. Melnikov, A. A. *et al.* Active learning machine learns to create new quantum experiments. *Proc. Natl. Acad. Sci.* **115**, 1221–1226 (2018).

84. Melnikov, A., Kordzanganeh, M., Alodjants, A. & Lee, R.-K. Quantum machine learning: from physics to software engineering. *Adv. Physics: X* **8**, 2165452 (2023).

85. Frezat, H., Sommer, J., Fablet, R., Balarac, G. & Lguensat, R. A posteriori learning for quasi-geostrophic turbulence parametrization. *J. Adv. Model. Earth Syst.* **14**, e2022MS003124 (2022).

86. Molina, M. J. *et al.* A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artif. Intell. for Earth Syst.* **2**, 220086 (2023).

87. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273, DOI: 10.1038/s41586-019-1787-x (2019). Number: 7786 Publisher: Nature Publishing Group.

88. Yao, S. *et al.* A whole-brain monosynaptic input connectome to neuron classes in mouse visual cortex. *Nat. neuroscience* **26**, 350–364 (2023).

89. Consortium, M. *et al.* Functional connectomics spanning multiple areas of mouse visual cortex. *BioRxiv* 2021–07 (2021).

90. Nelson, M. & Rinzel, J. The Hodgkin–Huxley model. *The Book Genes.* (1995).

91. Chen, Y., Luo, Y., Liu, Q., Xu, H. & Zhang, D. Symbolic genetic algorithm for discovering open-form partial differential equations (sga-pde). *Phys. Rev. Res.* **4**, 023174 (2022).

92. Du, M., Chen, Y. & Zhang, D. Discover: Deep identification of symbolic open-form PDEs via enhanced reinforcement-learning. *Prepr. arXiv:2210.02181* (2022).

93. Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. *Adv. neural information processing systems* **31** (2018).

94. Becker, S., Klein, M., Neitz, A., Parascandolo, G. & Kilbertus, N. Predicting ordinary differential equations with transformers. In *International Conference on Machine Learning*, 1978–2002 (PMLR, 2023).

95. Sahoo, S., Lampert, C. & Martius, G. Learning equations for extrapolation and control. In *International Conference on Machine Learning*, 4442–4450 (Pmlr, 2018).

96. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain* **143**, 1920–1933 (2020).

97. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Syst.* **12**, 522–537 (2021).

98. MacLeod, B. P. *et al.* A self-driving laboratory advances the pareto front for material properties. *Nat. Commun.* **13**, 995 (2022).

99. Du, J., Futoma, J. & Doshi-Velez, F. Model-based reinforcement learning for semi-Markov decision processes with neural ODEs. *Adv. Neural Inf. Process. Syst.* **33**, 19805–19816 (2020).

100. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis machine intelligence* **35**, 1798–1828 (2013).

101. Schölkopf, B. *et al.* Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).

102. Camps-Valls, G. *et al.* Discovering causal relations and equations from data. *Phys. Reports* **1044**, 1–68 (2023).

103. Runge, J. *et al.* Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553 (2019).

104. Lobentanzer, S., Rodriguez-Mier, P., Bauer, S. & Saez-Rodriguez, J. Molecular causality in the advent of foundation models. *Prepr. arXiv:2401.09558* (2024).

105. Pfister, N., Bauer, S. & Peters, J. Learning stable and predictive structures in kinetic systems. *Proc. Natl. Acad. Sci.* **116**, 25405–25411 (2019).

106. Champion, K., Lusch, B., Kutz, J. N. & Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci.* **116**, 22445–22451 (2019).

107. Lu, C. *et al.* Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Adv. Neural Inf. Process. Syst.* **35**, 5775–5787 (2022).

108. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695 (2022).

109. Watson, J. L. *et al.* De novo design of protein structure and function with rfdiffusion. *Nature* **620**, 1089–1100 (2023).

110. Zeni, C. *et al.* Mattergen: a generative model for inorganic materials design. *Prepr. arXiv:2312.03687* (2023).

111. Nichani, E., Damian, A. & Lee, J. D. How transformers learn causal structure with gradient descent. *Prepr. arXiv:2402.14735* (2024).

112. Vetter, J., Macke, J. H. & Gao, R. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *bioRxiv* 2023–08 (2023).

113. Kirillov, A. *et al.* Segment anything (2023). 2304.02643.

114. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289, DOI: 10.1038/s41593-018-0209-y (2018). Number: 9 Publisher: Nature Publishing Group.

115. Zvyagin, M. *et al.* Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The Int. J. High Perform. Comput. Appl.* **37**, 683–705 (2023).

116. Bzdok, D. *et al.* Data science opportunities of large language models for neuroscience and biomedicine. *Neuron* **112**, 698–717 (2024).

117. Hollmann, N. *et al.* Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326, DOI: 10.1038/s41586-024-08328-6 (2025).

118. Singh, S. & Hooda, S. A study of challenges and limitations to applying machine learning to highly unstructured data. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–6, DOI: 10.1109/ICCUBEA58933.2023.10392115 (Pune, India, 2023).

119. Jones, C., Castro, D. C., De Sousa Ribeiro, F. *et al.* A causal perspective on dataset bias in machine learning for medical imaging. *Nat. Mach. Intell.* **6**, 138–146, DOI: 10.1038/s42256-024-00797-8 (2024).

120. Vaidya, A., Chen, R. J., Williamson, D. F. K. *et al.* Demographic bias in misdiagnosis by computational pathology models. *Nat. Medicine* **30**, 1174–1190, DOI: 10.1038/s41591-024-02885-z (2024).

121. van Giffen, B., Herhausen, D. & Fahse, T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *J. Bus. Res.* **144**, 93–106, DOI: 10.1016/j.jbusres.2022.01.076 (2022).

122. Savage, N. Breaking into the black box of artificial intelligence. *Nature* DOI: 10.1038/d41586-022-00858-1 (2022).

123. Zhong, X., Gallagher, B., Liu, S. *et al.* Explainable machine learning in materials science. *npj Comput. Mater.* **8**, DOI: 10.1038/s41524-022-00884-7 (2022).

124. Gallegos, M., Vassilev-Galindo, V., Poltavsky, I. *et al.* Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat. Commun.* **15**, DOI: 10.1038/s41467-024-48567-9 (2024).

125. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P. *et al.* Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-86327-7 (2021).

126. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nat. Methods* **13**, 703–704, DOI: 10.1038/nmeth.3968 (2016).

127. Welcome to the ai future? *Nat. Astron.* **7**, 1, DOI: 10.1038/s41550-023-01891-4 (2023).

128. Sapoval, N., Aghazadeh, A., Nute, M. G. *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**, DOI: 10.1038/s41467-022-29268-7 (2022).

129. Ridhawi, I. A., Otoum, S., Aloqaily, M. & Boukerche, A. Generalizing ai: Challenges and opportunities for plug and play ai solutions. *IEEE Netw.* **35**, 372–379, DOI: 10.1109/MNET.011.200037 (2021).

130. The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826, DOI: 10.1021/jacs.1c09820 (2021).

131. Jacobsson, T. J., Hultqvist, A., García-Fernández, A. *et al.* An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat. Energy* **7**, 107–115, DOI: 10.1038/s41560-021-00941-3 (2022).

132. Gražulis, S. *et al.* Crystallography open database–an open-access collection of crystal structures. *J. Appl. Crystallogr.* **42**, 726–729 (2009).

133. Barba, L. *Reproducibility and replicability in science* (National Academies Press, 2019).

134. Singh, P. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Sci. Manag.* **6**, 144–157, DOI: 10.1016/j.dsm.2023.06.001 (2023).

135. Yang, J., Soltan, A. A. S., Eyre, D. W. *et al.* Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat. Mach. Intell.* **5**, 884–894, DOI: 10.1038/s42256-023-00697-3 (2023).

136. Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N. *et al.* Considerations for addressing bias in artificial intelligence for health equity. *npj Digit. Medicine* **6**, DOI: 10.1038/s41746-023-00913-9 (2023).

137. Zheng, P., Zubatyuk, R., Wu, W. *et al.* Artificial intelligence-enhanced quantum chemical method with broad applicability. *Nat. Commun.* **12**, 7022, DOI: 10.1038/s41467-021-27082-9 (2021).

138. Davies, J. Program good ethics into artificial intelligence. *Nature* DOI: 10.1038/538291a (2016).

139. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.* **3**, 659–666, DOI: 10.1038/s42256-021-00373-4 (2021).

140. Huerta, E. A., Khan, A., Huang, X. *et al.* Accelerated, scalable and reproducible ai-driven gravitational wave detection. *Nat. Astron.* **5**, 1062–1068, DOI: 10.1038/s41550-021-01405-0 (2021).

141. Space missions out of this world with ai. *Nat. Mach. Intell.* **5**, 183, DOI: 10.1038/s42256-023-00643-3 (2023).

142. Ma, P. X., Ng, C., Rizk, L. *et al.* A deep-learning search for technosignatures from 820 nearby stars. *Nat. Astron.* **7**, 492–502, DOI: 10.1038/s41550-022-01872-z (2023).

143. Popper, K. *The Open Universe: An Argument for Indeterminism From the Postscript to The Logic of Scientific Discovery* (Routledge, 1992).

144. Åkesson, A., Curtsdotter, A., Eklöf, A. *et al.* The importance of species interactions in eco-evolutionary community dynamics under climate change. *Nat. Commun.* **12**, 4759, DOI: 10.1038/s41467-021-24977-x (2021).

145. Physics Nobel scooped by machine-learning pioneers. *Nature* (2024).

146. Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures. *Nature* (2024).

Acknowledgements

Author contributions

RV and HA initiated the idea for this article following a workshop celebrated in November 2022 at KTH. All the authors contributed equally to the rest of this work.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.