



On scientific understanding with artificial intelligence

Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao  and Alán Aspuru-Guzik 

Abstract | An oracle that correctly predicts the outcome of every particle physics experiment, the products of every possible chemical reaction or the function of every protein would revolutionize science and technology. However, scientists would not be entirely satisfied because they would want to comprehend how the oracle made these predictions. This is scientific understanding, one of the main aims of science. With the increase in the available computational power and advances in artificial intelligence, a natural question arises: how can advanced computational systems, and specifically artificial intelligence, contribute to new scientific understanding or gain it autonomously? Trying to answer this question, we adopted a definition of ‘scientific understanding’ from the philosophy of science that enabled us to overview the scattered literature on the topic and, combined with dozens of anecdotes from scientists, map out three dimensions of computer-assisted scientific understanding. For each dimension, we review the existing state of the art and discuss future developments. We hope that this Perspective will inspire and focus research directions in this multidisciplinary emerging field.

Artificial intelligence (AI) has been called a revolutionary tool for science^{1,2} and it has been predicted to play a creative role in research in the future³. In the context of theoretical chemistry, for example, it is believed that AI can help solve problems “in a way such that the human cannot distinguish between this [AI] and communicating with a human expert”⁴. However, this excitement has not been shared by all scientists. Some have questioned whether advanced computational approaches can go beyond ‘numerics’^{5–9} and contribute on a fundamental level to gaining of new scientific understanding^{10–12}.

In this Perspective, we discuss how advanced computational systems, and AI in particular, can contribute to scientific understanding: we overview what is currently possible and what might lie ahead. In addition to the review of the literature, we surveyed dozens of scientists working at the interface of biology, chemistry or physics on the one hand, and AI and advanced

computational methods on the other. These personal narratives (see Supplementary Information) focus on the concrete discovery process of ideas and are a vital augmentation of the scientific literature. We discuss the literature overview and personal accounts in the context of the philosophical theory of scientific understanding recently developed by Dennis Dieks and Henk de Regt^{12,13}. We then identify three fundamental dimensions for AI contributing to new scientific understanding (FIG. 1). (We encapsulate all advanced artificial computational systems under the term AI, independent of their working principles. In this way, we are focusing on the operational objective rather than the methodology.) First, AI can act as an instrument revealing properties of a physical system that are otherwise difficult or even impossible to probe. Humans then lift these insights to scientific understanding. Second, AI can act as a source of inspiration for new concepts and ideas that are subsequently

understood and generalized by human scientists. Third, AI acts as an agent of understanding. AI reaches new scientific insight and — importantly — can transfer it to human researchers. Although there have not yet been any examples of AI acting as a true ‘agent of understanding’ in science, we outline important characteristics of such a system and discuss possible ways to achieve it.

In the first two dimensions, the AI enables humans to gain new scientific understanding, whereas in the last, the machine gains understanding itself. Distinguishing between these classes allows us to map out a vibrant and mostly unexplored field of research, and will hopefully guide direction for future AI developments in the natural sciences.

The focus of this Perspective is how advanced computational systems and AI specifically can contribute to new scientific understanding. There are many related, interesting topics that we cannot cover here. For example, we will not discuss the relationship between scientific understanding and cognitive science, but refer the reader to a good overview¹⁴. Furthermore, we will only discuss ‘understanding’ in the context of the natural sciences, in which we can use concrete criteria from the philosophy of science and, therefore, will not touch on ‘understanding’ in a broader context (such as understanding by babies and animals, language understanding in AI and related topics). Many other works contribute to related questions and should be mentioned here. One important field of research in AI is explainable AI, which aims to interpret and explain how advanced AI algorithms come up with their solutions; see, for instance, REFS.^{15–18} Whereas it is not necessary, and we believe also not sufficient, to interpret the internal workings of the AI to get new scientific understanding, many of these tools and techniques can be very useful. We will briefly explain them below with concrete examples in the natural sciences. AI pioneer Donald Michie classified machine learning (ML) into three classes: weak, strong and ultrastrong, in which ultrastrong requires the machine to teach the human¹⁹. The ultrastrong ML is related to the idea of agent of understanding,

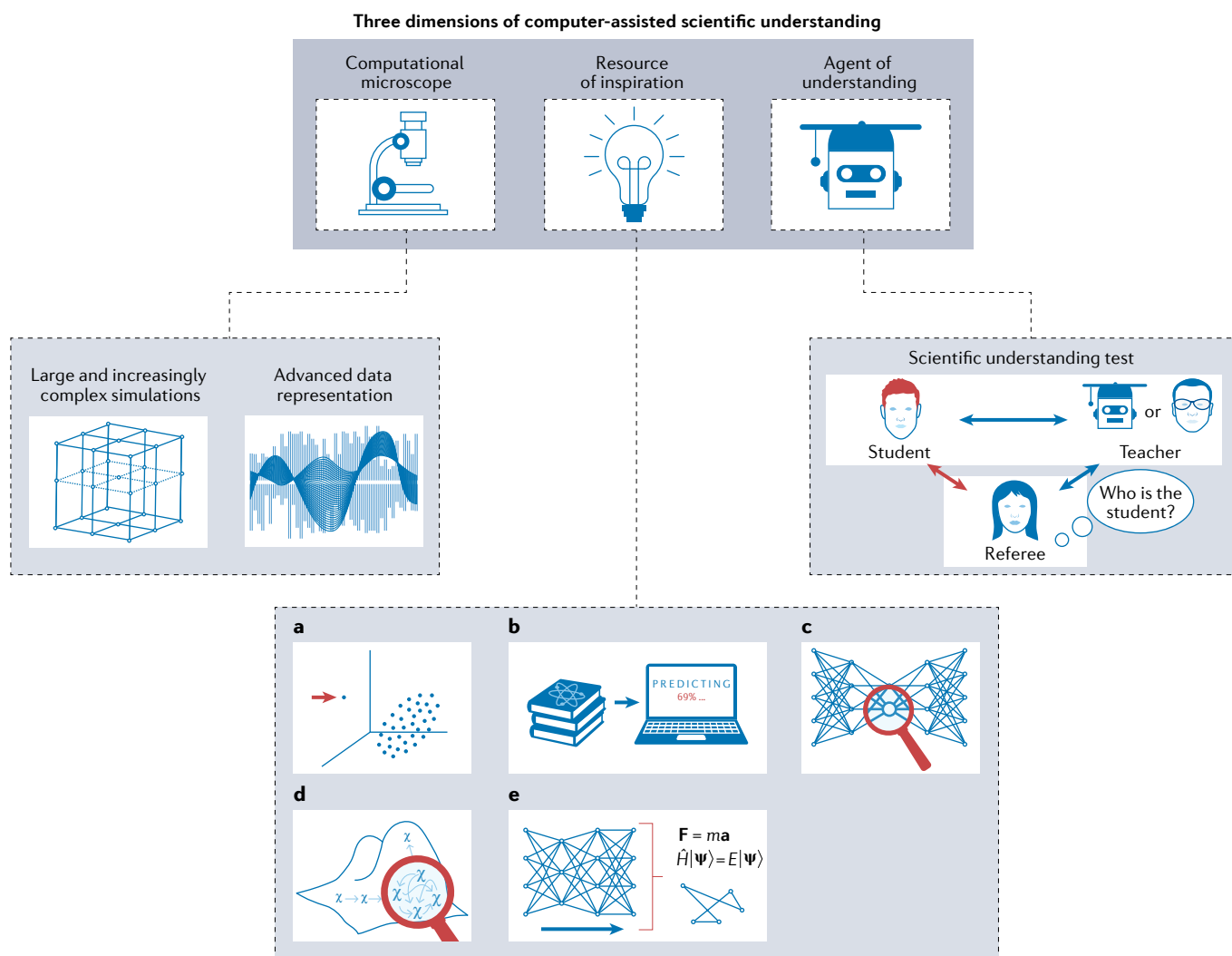


Fig. 1 | The three dimensions of computer-assisted scientific understanding. The current state-of-the-art computational microscopes could be developed further with more complex systems, which could be simulated thanks to advances in algorithms and hardware, and with more advanced data representations (left-hand panel). As resources of inspiration, computational

systems can help the human scientist by identifying surprises in data (a), identifying surprises in the scientific literature (b), finding surprising concepts by inspecting models (c), probing the behaviour of artificial agents (d) or by extracting new concepts from interpretable solutions (e). The scientific understanding test discussed in the main text is illustrated in the right-hand panel.

as we will define and go into more detail below. A very useful and comprehensible collection of computational and AI methods for science can be found in REF.²⁰. Different levels of automation in the design of molecules are described in REF.²¹, with a final step involving a computer that chooses the initial ideas. Other works have investigated what full automation might look like based on specific scientific methodologies, leading to the idea of the ‘Nobel Turing Challenge’²², the development of an AI system capable of making Nobel-prize-level scientific discoveries. We note that our take purposefully does not depend on any specific scientific method (to avoid problems at a foundational level²³). Rather, we focus on ‘scientific understanding’, how scientists can get it and

how advanced AI can help humans gain new scientific understanding.

Scientific understanding

Imagine an oracle providing non-trivial predictions that are always true. Although such a hypothetical system would have a significant scientific impact, scientists would not be satisfied. They would want “to be able to grasp how the predictions are generated, and to develop a feeling for the consequences in concrete situations”¹³. Colloquially, we refer to this goal as ‘understanding’, but what does this really mean? To find criteria for scientific understanding, we seek guidance from the philosophy of science. Although hardly any scientist would argue against ‘understanding’ as a fundamental aim of science (along with

explanation, description and prediction²⁴), this view has not always been accepted by philosophers. Carl Hempel, who made foundational contributions clarifying the meaning of ‘scientific explanation’, argued that ‘understanding’ is subjective and merely a psychological by-product of scientific activity and is, therefore, not relevant for the philosophy of science²⁵. Numerous philosophers criticized this conclusion, trying to formalize what ‘scientific understanding’ actually means. These proposals suggest that ‘understanding’ is connected to the ability to build causal models (for example, Lord Kelvin said “It seems to me that the test of ‘Do we or not understand a particular subject in physics?’ is, ‘Can we make a mechanical model of it?’”¹³), connected to providing visualizations

(or *Anschaulichkeit*, as its strong proponent Erwin Schrödinger called it^{26,27}) or that understanding corresponds to providing a unification of ideas^{28,29}.

More recently, Henk de Regt and Dennis Dieks have developed a new theory of scientific understanding, which is both contextual and pragmatic^{12,13,24}. They found that techniques such as visualization or unification are ‘tools for understanding’, thereby connecting previous ideas in one general framework. Their theory is agnostic to the specific ‘tool’ being used, making it particularly useful for application in a variety of scientific disciplines. de Regt and Dieks extended Werner Heisenberg’s insights³⁰ and, rather than merely introducing theoretical or hypothetical ideas, the main motivation behind their theory is that a “satisfactory conception of scientific understanding should reflect the actual (contemporary and historical) practice of Science”. Simply put, they argue that: “A phenomenon P can be understood if there exists an intelligible theory T of P such that scientists can recognise qualitatively characteristic consequences of T without performing exact calculations”^{12,13}. de Regt and Dieks defined two interlinked criteria:

- Criterion of understanding phenomena: a phenomenon P can be understood if a theory T of P exists that is intelligible.
- Criterion for the intelligibility of theories: a scientific theory T is intelligible for scientists (in context C) if they can recognise qualitatively characteristic consequences of T without performing exact calculations.

We decided to use this specific theory because it can be used to ‘experimentally’ evaluate whether scientists have ‘understood’ new concepts or ideas, rather than by inspecting their methodology, by simply looking at the scientific outcome and the consequences. This approach also coincides with Angelika Potochnik’s argument that “understanding requires successful mastery, in some sense, of the target of understanding”¹¹.

Scientific discovery versus scientific understanding

Scientific understanding and scientific discovery are both important aims in science. The two are distinct in the sense that scientific discovery is possible without new scientific understanding (we use the precise terminology in REFS.^{12,13}).

Let’s examine three examples. First, to design new efficient molecules for organic laser diodes, a search space of 1.6 million

was explored using ML and quantum chemistry insights³¹. The top candidate was experimentally synthesized and investigated. Thereby, the authors of this study discovered new molecules with very high quantum efficiency. Whereas these discoveries could have important technological consequences, the results do not provide new scientific understanding. From the results per se, one cannot derive qualitative consequences without performing further detailed computations. Second, the recent ML-enabled breakthrough in protein folding^{32,33} will undoubtedly change the landscape of biochemistry. However, so far, AlphaFold has been a black box: an oracle and, as such, it does not directly provide new scientific understanding in the sense of de Regt and Dieks. Third, many discoveries in physics occur before (sometimes long before) a theory or explanation, which provides scientific understanding, is uncovered. Examples include the discovery of superconductivity (and its high-temperature version), the discovery of the cosmological microwave background, neutrino oscillations and the discovery of a zoo of particles before the invention of the quark model.

These examples show that scientific discoveries can lead to scientific and technological disruption without directly contributing to scientific understanding^{11,24}.

Over the past few years, scientists working at the interface between AI and the natural sciences have been trying to rediscover physical laws or concepts with machines. Examples include the heliocentric world view³⁴, the arrow of time³⁵ or mechanical equations of motions^{36,37}. These applications are good benchmarks to show that the algorithms work in principle. The question remains, however, whether an AI that can rediscover physical laws and concepts would also be capable of contributing to new scientific understanding. We believe that this is not guaranteed. The human creators of these AI systems know what they are looking for in these case studies. Therefore, it is unclear how both conscious and unconscious biases (in the broadest sense, for example, by choosing particular representations) in the code or the data can be prevented. Consequently, even if an algorithm can rediscover interesting physical phenomena, it is unclear whether and how it can be used to advance science by uncovering new scientific ideas. To go beyond rediscovery tasks, one needs to focus explicitly on the question of how to get ‘new’ scientific understanding.

Loosely speaking, gaining new understanding from advanced computational systems means uncovering new ideas, principles, concepts or even theories that scientists can apply and use in different situations without (complete) computations. In the next sections, we will outline how this could be done, what previous approaches have achieved and how we can go further.

Three dimensions of computer-assisted understanding

As already alluded to in the introduction, we surveyed the scientific literature and used personal anecdotes from dozens of scientists. Then, within the framework provided by the philosophy of science, we introduced a new classification of AI’s contribution to scientific understanding that helps map out different directions of future investigation. We call this classification ‘dimensions’, as they are independent and non-exclusive.

An AI system can contribute to new scientific understanding in three ways. First, as a ‘computational microscope’, it can provide information not (yet) attainable through experimental means. Second, as a ‘resource of inspiration’ or an artificial muse, expanding the scope of human imagination and creativity. In those two dimensions, the human scientist is essential to identify and refine the new insight and inspiration and develop it to full understanding. For the former, the machine creates new data (and represents it potentially in advanced ways) and, thereby, the human scientist extracts from it her new understanding. In the latter, the machine explicitly looks for surprising or interesting new ideas or unexpected connections, which it presents to the human scientist, who uses them to reach new scientific understanding. These two dimensions could exist even without advanced computational systems or AI. However, AI can significantly boost and extend their opportunities.

The third dimension is that of AI as an ‘agent of understanding’, replacing the human in generalizing observations and transferring these new scientific concepts to different phenomena, and — importantly — conveying these insights to human scientists.

As the focus of this Perspective is AI-assisted new scientific understanding, we stress that such an agent of understanding must be able to transfer its insights to humans. We will briefly discuss the situation of an agent that cannot convey its insights.

The three dimensions described above should not be understood dogmatically

but, rather, provide a framework to guide future directions. In the following sections, based on concrete examples, we discuss each dimension in more detail and propose avenues for pushing the boundaries of what is currently possible.

Computational microscope

Microscopes are perhaps the best known type of instrument that allows the investigation of objects and phenomena invisible to the naked eye. Similarly, computational microscopes enable the investigation of objects or processes that cannot be visualized or probed in any other way, for example, biological, chemical or physical processes that happen at length and time scales not accessible in experiments.

In the context of ‘understanding’, the new computer-generated data by a computational microscope need to be generalized to other contexts without complete computation¹³. We illustrate this with two concrete examples.

The first example is molecular dynamics simulations of SARS-CoV-2. The authors of REF.³⁸ uncovered new biological functions that show different behaviours in the open and closed conformations of the spike protein. This explanation changed the view of glycans in biological systems and inspired new ways to analyse these systems without the need to perform full computations.

In the second example, the authors of REF.³⁹ described how molecular dynamics simulations helped them uncover fundamental patterns called glycoblocks. The systematic use of glycoblocks can be used to both understand the sequence–structure–property relationships of biomolecules and inform the design of synthetic structures with desired functions without the need for simulating the entire system.

Can computational microscopes be improved further? We discuss two vibrant directions. First, more advanced computational systems will allow the analysis of increasingly complex physical systems. Second, representing the information in a more interpretable way will facilitate the extractions of scientific insight.

Increasingly complex systems. An obvious, but nevertheless important, research direction is increasing the complexity and the accuracy of computer simulations⁴⁰. For example, increasing the size of the simulated systems, the timescale of the simulations and the number of interactions that can be modelled will significantly enhance the ability to study complex dynamic systems.

In general, such advances can be achieved by algorithmic or hardware improvements, or both. In that regard, we expect that AI technologies together with advanced hardware such as GPUs, TPUs or even OPUs^{41,42} will have an enormous impact. Furthermore, the progress in experimental quantum computing for quantum chemistry⁴³ and physics^{44–46} promises that entirely new algorithms, based on quantum mechanics, will play an important role in this area^{47,48}. Algorithmic improvements could involve adaptive and intelligent resolution during the simulation and advanced visualization methods¹³, which directly leads to the second direction of future development.

Advanced data representation. In the first dimension of computer-assisted understanding, the human scientist is supposed to generalize the new data from the computational microscope. Therefore, we believe that advances in data representation could significantly help humans to grasp the underlying structures and facilitate new scientific understanding. Scientists are currently predominantly analysing data in (potentially animated) 2D graphical representations. We believe that genuinely 3D environments (realized via either virtual or augmented reality glasses or holography) will significantly help the understanding of complex systems or complex data. Initial advances in that regard have been demonstrated in chemistry^{49–51} and astrophysics⁵², and we expect this to become a standard tool for scientists. Furthermore, the time dimensions could be used to represent even more structured data; for example, via videos (3D videos). Alternatively, sound could be used as an additional data dimension, as the human auditory sense is excellent in detecting structure or symmetries in (periodic) time-dependent data. This opportunity has been explored in high-energy physics⁵³ and in dozens of projects in astronomy⁵⁴. A powerful algorithm might be able to identify symmetries in the underlying data and project them into 3D video with sound, which might help the human to recognize and subsequently understand new properties within the data generated by the computational microscope.

Resource of inspiration

Surprising and creative ideas are the foundation of advances in science. Computer algorithms can inspire such ideas systematically, thereby significantly accelerating scientific and technological

progress. Already 70 years ago, Alan Turing noted that “Machines take me by surprise with great frequency”⁵⁵. A more recent study⁵⁶ collected stories from dozens of researchers of artificial life and evolution. These anecdotes showcase how computer algorithms can “produce surprising and creative solutions”. Accordingly, we believe that AI could be artificial muses of science in a metaphorical sense. Next, we will outline a number of ways in which computer algorithms can provide a source of inspiration for new scientific ideas.

Identifying surprises in data. Exceptional data points or unexpected regularities obtained from experiments or simulations can trigger new ideas and concepts. Our survey shows that these exceptional points are usually identified by humans, such as in the following two examples, which use high-throughput computations in chemistry⁵⁷ and quantum optics^{58,59}.

The first example deals with an unexpected phase of crystal structures in high-pressure physics. In REF.⁶⁰, the authors found an unexpected stable configuration of alternating NH₂ and NH₄ layers, rather than a dense NH₃ phase. They conceptualized this phenomenon as spontaneous ionization, a common process in acid–base chemistry, which is now a widely accepted phenomenon in the high-pressure phase diagram of NH₃. Spontaneous ionization in the high-pressure behaviour of matter has become a more general principle that can be used without performing any simulations.

In the second example, a search for new quantum experiments uncovered a solution with considerable larger quantum entanglement than expected⁵⁸. The authors of the study understood the underlying principles and thereby discovered a new concept of entanglement generation^{61,62}. The principle can be used without any computation and, for example, acts now as a new representation in more advanced AI systems for quantum physics⁶³, demonstrating the application of the computer-inspired idea in more general different contexts.

In contrast to these examples, data anomalies can manifest themselves in a more involved combination of variables, which might be very difficult for humans to grasp. Accordingly, applying advanced statistical methods and ML algorithms (for example, REF.⁶⁴) to this type of problem will be an important future research direction. Exciting works in the direction of autonomous anomaly detection have been applied to scientific data from the

Large Hadron Collider (LHC) at CERN^{65–67} (see a recent review on this topic⁶⁸). Such techniques have the potential to identify new physics signatures, which can then be conceptualized and understood by human physicists^{69,70}. An interesting technique for detecting outliers, previously used for the discovery of quantum phases, is leaving out one (or a number) of training examples and observing the consequence for the neural networks⁷¹. The introduction of influence functions, via the computation of Hessians, is one computationally efficient way to identify the impact of individual training examples⁷². Neural networks that autonomously discover symmetries could become an efficient discovery tool for outliers in scientific data in which the underlying rules might not be known beforehand^{73,74}.

Estimating the confidence of predictions will be another method to directly search for anomalies in data⁷⁵. The ability to uncover hidden regularities was demonstrated in mathematics, in which an AI hinted at relations between previously unconnected invariants in knot theory, which allowed mathematicians to conjecture and prove new theorems⁷⁶. Alternatively, an AI capable of constructing new scientific hypotheses could uncover outliers or unexpected patterns that are not discernible with standard statistical methods.

It would be truly exciting to see an AI uncover hidden patterns or irregularities in scientific data previously overlooked by humans, which could lead to new ideas and, ultimately, to new conceptual understanding. As of now, we are not aware of such cases.

Finally, we note that the data points for these analyses could be obtained from computational methods (involving those described in the previous section), with exciting opportunities for mathematics or theoretical physics⁷⁷. Alternatively, the data could be obtained directly from experiments. Here, we can imagine a closed-loop approach, in which an algorithm tries to explore the environment and steer the exploration into unexpected regions. If the data source is an experiment, a future AI system will require access to complex lab automation with large parameter spaces to explore, as demonstrated in biology⁷⁸, chemistry^{79–84} or physics^{85,86}.

Identifying surprises in the scientific literature. The number of scientific papers in essentially every scientific domain is growing enormously^{87,88}. Consequently, researchers have to specialize in narrow

subdisciplines, which makes finding new interdisciplinary ideas difficult. In the future, we believe that computers will be able to use the scientific literature in an automated way^{89–92} and identify exceptional and surprising phenomena for further investigation. Whereas the large-scale automated analysis of the scientific literature, to our knowledge, has not yet been able to induce new scientific understanding, there is significant progress in the field. One promising approach towards this goal is unsupervised word embedding of a large corpus of scientific papers. In that technique, the content of the scientific literature is transformed into a high-dimensional vector space. This approach has been used in the domain of materials science⁹³ and rediscovered central scientific concepts, such as the periodic table of the elements. Additionally, the results also suggested the existence of previously undiscovered structure–property relationships. Examples include new candidates for thermoelectric materials. Moreover, several other advanced computational techniques are being developed in materials science to extract knowledge from the scientific literature and investigate it systematically by AI technologies⁹⁴, and can lead to complex scientific conclusions, as demonstrated, for instance, on the relations between different crystal structures⁹⁵.

An alternative approach aims to build semantic knowledge networks from large bodies of scientific literature. In these networks, scientific concepts are nodes and edges carry relational information. In the simplest case, that means two scientific concepts are mentioned in the same scientific paper^{96,97}. Thus, scientific knowledge is represented as an evolving network, which can be used to identify both islands and unexplored regions of the scientific literature. This type of network was used in biochemistry to identify efficient global research strategies⁹⁶ and in quantum physics to predict and suggest future research directions⁹⁷. Advances in AI technology could improve this type of system significantly. For example, natural language processing architectures such as BERT⁹⁸, or GPT-3 (REF.⁹⁹) could help extract more scientific knowledge from research papers, and large graph-based neural networks could improve the prediction of new research topics from semantic networks¹⁰⁰.

Surprising concepts by inspecting models. We also expect considerable progress by rationalizing what AI algorithms have learned in order to solve a specific

problem, in other words, explainable or interpretable AI^{101–103}. One idea towards this goal is inspired by DeepDreaming, a method first used in computer vision^{104,105}. Put simply, the idea is to invert a neural network and probe its behaviour. This approach has been applied to rediscover thermodynamical properties³⁵ and design principles for functional molecules¹⁰⁶. An alternative and remarkable application is the ‘disentanglement of variables’ in neural networks¹⁰⁷. The goal is to understand the internal representation that the neural network has learned. Astronomical data, represented in geocentric coordinates, was used to train a neural network and the disentanglement of variables enabled the rediscovery of heliocentric coordinates via the internal representation of the model³⁴. Symmetries, or their conserved quantities, can also be autonomously extracted by using a pair of neural networks^{108,109}. The pair is then trained to identify whether two different physical situations are equivalent under some unknown symmetry. The final layer is an information bottleneck consisting of only a single neuron. In that way, the neural networks learn to identify conservation properties and compress the entire information into the output of one neuron. The output can then be easily compared and, moreover, readily interpreted by the human researcher. The authors of REF.¹⁰⁹ show how this idea can be used to rediscover conserved properties in classical mechanics (energy and angular momentum) or electromagnetism (such as the Lorentz invariants). In a related study, using gradient boosting with decision trees, feature importance has been used to explain properties of molecules and quantum optics circuits¹¹⁰. Related to this is a study in which the internal representation of an unsupervised deep generative model for quantum experiments has been inspected to understand the model’s internal worldview¹¹¹. In chemistry, counterfactual explanations for ML models have been shown to produce the rationale behind a model’s prediction. Counterfactual explanations illustrate what differences in an event or instance would generate a change in an outcome. Reference¹¹² showed how this can be achieved in a model-independent way (it has been demonstrated for random forest, sequence models and graph neural networks), indicating a great future potential for opening the black box of AI in science. Albeit not in science, a study has investigated what the chess-playing AI AlphaZero has learned about chess and how human-like knowledge is encoded in the internal

representation¹¹³. The concepts rediscovered in all of those works were not new and, thus, the most important challenge for the future is to learn how to extract previously unknown concepts.

New concepts from interpretable solutions.

Rather than getting inspiration from the AI algorithms, scientists can also be surprised by the solutions they provide. When solutions are represented in an interpretable way, they can lead to new ideas and concepts. An example of interpretable representation is a mathematical formula. Thus, scientists can inspect formulae derived by computer algorithms to solve mathematical problems directly and derive more general solution strategies. Several studies demonstrated extracting symbolic models from experimental data of mechanical systems^{36,114}, of quantum systems¹¹⁵ and in astronomy¹¹⁶; see also REF.¹¹⁷. It will be exciting to see how these approaches, for example, combined with methods such as causal inference¹¹⁸, can be improved to propose reasonable physical models of unknown systems that advance scientific understanding. Altogether, exciting advances have been achieved in the field of mathematics^{119,120}, and we foresee similar approaches making a significant impact in the physical sciences.

One recent, concrete example in astronomy is the rediscovery of Newton's law of gravitation from real-world observational data of planets and moons in our Solar System from the last 30 years (REF.³⁷). The application of graph neural networks allowed for the high-quality prediction of the object's motion. Furthermore, a symbolic regression technique called PySR (introduced in REF.¹¹⁶) was able to extract reasonable mathematical expressions for the learned behaviour. Interestingly, besides the equations of motion, the method simultaneously predicts the masses of the planetary objects correctly. The technique required the assumption of several symmetries and other physical laws. It remains to be seen whether these prerequisites can be reduced further and how related approaches can be applied to modern physics questions.

Another example of this methodology has been showcased in the field of quantum optics⁶³. There, an AI algorithm with a graph-theoretical representation of quantum optical setups designs configurations for previously unknown quantum systems. The final solutions were represented in a physically interpretable graph-theoretical representation. From there, human scientists

can quickly interpret the underlying reasons why the solutions work and apply them in other contexts without further computation. Accordingly, developing interpretable representations and methods to extract underlying concepts in other domains will be an important future research direction.

Probing the behaviour of artificial agents.

Another only rarely explored opportunity is interpreting the behaviour of machines when tasked to solving a scientific problem¹²¹. Algorithms that take action such as genetic algorithms or reinforcement learning agents adopt policies to navigate the problem space. Human scientists can observe how they navigate this space. Instead of following a strict external reward, for example, maximizing a specific property of a physical system, intrinsic rewards such as artificial curiosity can be implemented^{122,123}. Instead of maximizing directly some functions, the artificial agent tries to learn and predict the behaviour of the environment. It then chooses actions that lead to situations it cannot predict well, thus maximizing its own understanding of the environment. It has been shown using curious agents in simulated virtual universes¹²⁴ and robot agents in real laboratories⁸⁴ that curiosity is an efficient exploration strategy. Alternative intrinsic rewards for artificial agents are 'computational creativity'^{125,126} and 'surprise'¹²⁷. These intrinsic rewards can produce exceptional and unexpected solutions and agent behaviour, ultimately inspiring human scientists.

Agent of understanding

The third and final dimension we consider are algorithms that can autonomously acquire new scientific understanding, and ultimately explain these insights to humans. This feat has been described by neither the respondents of our survey nor in the scientific literature. Therefore, we will approach this class by listing the requirements of these agents, proposing tests to detect their successful realization and speculating what such systems could look like.

The idea of a machine that translates insights to humans has been discussed in Donald Michie's pioneering work of 1988 called 'Machine learning in the next five years'^{119,128,129}. Michie classified AI algorithms into three classes: weak ML, strong ML and ultrastrong ML. Weak ML achieves improved prediction quality with a larger amount of training data. Arguably, most ML approaches today fall into this category, in which the algorithm is treated as a black

box. The class of strong ML approaches requires providing a symbolic representation of their hypothesis, for example, via Boolean (logical) expressions or mathematical equations. And, finally, ultrastrong ML approaches require that the algorithm teaches the human operator such that the human performance is improved compared with the human learning from data alone. Interestingly, experiments on certain logical tasks have shown that ultrastrong ML algorithms might already exist and thereby have demonstrated the "existence of a class of relational concepts which are hard to acquire for humans, though easy to understand given an abstract explanation"¹²⁹. The idea of ultrastrong ML is related to our third dimension, the agent of understanding. Both require that the machine gets new insights and teaches them to the human. Besides our constraint to the natural sciences, there are also other differences: whereas Michie requires that the insights are transferred in a symbolic way, the agent of understanding is more flexible and allows for any teaching method, for example, via discussions in natural language, such as GPT-3. Furthermore, the agent needs to provide new scientific understanding (in the strict sense of de Regt), rather than just explanations or interpretations. Therefore, the agent of understanding is more flexible regarding the technical implementation, but stricter regarding what it teaches the human. We will go into more detail below.

First, it is important to realize that finding 'new' scientific understanding is context-dependent. What is new depends on whether we consider an individual scientist and their field of expertise, a scientific domain, the whole scientific community or even the entire scientific endeavour throughout history. Hence, true agents of understanding must be able to evaluate whether insight is new, at least in the context of a specific scientific domain that requires access to the knowledge of that scientific field.

Secondly, de Regt emphasized the importance of underlying scientific theories that allow us to recognize qualitatively characteristic consequences¹². It is not enough to simply interpolate data points or predict new ones using advanced statistical methods such as ML. Thus, even though such methods can approximate complex and expensive computations, naïve applications of neural networks cannot be agents of understanding. Scientific understanding requires more than mere calculation. To illustrate this point even further, let us consider one concrete example in quantum physics from the literature: a computational

method solved an open question about the generation of important resource states for quantum computing. Then it extracted the conceptual core of the solution in the form of a new quantum interference effect in such a fashion that human scientists can both understand the results and apply the acquired understanding in different contexts⁶³. Even if the computer itself was able to apply the conceptual core to other situations, it would not be a priori clear whether the computer truly acquired scientific understanding. What is still missing is an explanation of the discovered technique in the context of a scientific theory. In this particular example, the AI and the human scientist would need to recognize the underlying quantum interference in the context of the theory of quantum physics. Thus, we can propose the first sufficient condition for agents of understanding:

Condition (Condition for scientific understanding I).

An AI gained scientific understanding if it can recognize qualitatively characteristic consequences of a theory without performing exact computations and use them in a new context.

This condition closely follows the ideas of de Regt and Dieks¹³. Let us go one step further and imagine that there is an AI capable of explaining discoveries in the context of scientific theories. How could human scientists recognize that the machine acquired new scientific understanding? We argue that human scientists would do it in the exact same way they can recognize that other human scientists acquired new scientific understanding. That is, let the other human scientists convey the newly acquired understanding to others. This suggests the second sufficient condition for agents of understanding:

Condition (Condition for scientific understanding II).

An AI gained scientific understanding if it can transfer its understanding to a human expert.

We argue that one can only recognize indirectly whether a computer (or human) has gained scientific understanding. Therefore, finally, we propose a test in the spirit of the Turing test⁵⁵ or the Feigenbaum test¹³⁰ (or adaptations thereof in the natural sciences, such as the chemical Turing test or the Feynman test⁴).

The scientific understanding test. A human (the student) interacts with a teacher, either a human or an artificial scientist.

The teacher's goal is to explain a scientific theory and its qualitative, characteristic consequences to the student. Another human (the referee) tests both the student and the teacher independently. If the referee cannot distinguish between the qualities of their non-trivial explanations in various contexts, we argue that the teacher has scientific understanding.

In principle, there is no reason why the student, or the referee, cannot be an AI. However, to keep the test as simple as possible, we restrict the number of possible variations.

The formulation of the test implies that humans need to understand the new concepts that AI devised. If a machine truly understands something, it should be able to explain it and transfer the understanding to someone else. (We leave aside the question of whether the explanation of the AI is true or false. It has been argued that also false theories can lead to genuine understanding¹³¹.) We believe that this should always be possible, even if the understanding is far beyond what human experts know at this point. Note that, as the gap between humans' and artificial scientists' capabilities grows, it will become increasingly challenging and time-consuming for the AI to transfer its understanding to humans. We envision that computers will use advanced human-computer interaction techniques together with the tools we described for the next-level computational microscopes.

Additionally, scientific discussions between a human and a computer could be realized using advanced queries in natural language processing tools such as BERT⁹⁸ or GPT-3 (REF.⁹⁹). That way, the scientist could probe the computer with scientific questions. Suppose the scientist gains new scientific understanding by communicating with the algorithm, as judged by our scientific understanding test. In that case, they can confirm that the computer truly acquired understanding. We would like to point out that our test, like the ones originated by Turing and Feigenbaum, are not clear-cut, leaving room for situations that do not allow a clear judgement. We are optimistic that more efforts will be directed at developing the necessary technologies, which will lead to ever more convincing demonstrations of artificial scientists acting as true agents of understanding in the future.

Conclusion

Undoubtedly, advanced computational methods in general and in AI specifically will further revolutionize how scientists investigate the secrets of our world.

We outline how these new methods can directly contribute to acquiring new scientific understanding. We suspect that significant future progress in the use of AI to acquire scientific understanding will require multidisciplinary collaborations between natural scientists, computer scientists and philosophers of science. Thus, we firmly believe that these research efforts can — within our lifetimes — transform AI into true agents of understanding that will directly contribute to one of the main goals of science, namely, scientific understanding.

Mario Krenn^{1,2,3,4,✉}, Robert Pollice^{2,3}, Si Yue Guo², Matteo Aldeghj^{2,3,4}, Alba Cervera-Lierta^{2,3}, Pascal Friederich^{2,3,5}, Gabriel dos Passos Gomes^{2,3}, Florian Häse^{2,3,4,6}, Adrian Jinich⁷, AkshatKumar Nigam^{2,3}, Zhenpeng Yao^{1D 2,8,9,10} and Alán Aspuru-Guzik^{1D 2,3,4,11✉}

¹Max Planck Institute for the Science of Light (MPL), Erlangen, Germany.

²Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, Ontario, Canada.

³Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

⁴Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada.

⁵Institute of Nanotechnology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany.

⁶Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.

⁷Division of Infectious Diseases, Weill Department of Medicine, Weill Cornell Medical College, New York, USA.

⁸Center of Hydrogen Science, Shanghai Jiao Tong University, Shanghai, China.

⁹State Key Laboratory of Metal Matrix Composites, School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

¹⁰Innovation Center for Future Materials, Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, China.

¹¹Canadian Institute for Advanced Research (CIFAR) Lebovic Fellow, Toronto, Ontario, Canada.

✉e-mail: mario.krenn@mpl.mpg.de; alan@aspuru.com

<https://doi.org/10.1038/s42254-022-00518-3>

Published online 11 October 2022

- Zdeborová, L. New tool in the box. *Nat. Phys.* **13**, 420–421 (2017).
- Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement learning with neural networks for quantum feedback. *Phys. Rev. X* **8**, 031084 (2018).
- Melnikov, A. A. et al. Active learning machine learns to create new quantum experiments. *Proc. Natl Acad. Sci. USA* **115**, 1221–1226 (2018).
- Aspuru-Guzik, A., Lindh, R. & Reiher, M. The matter simulation (r)evolution. *ACS Cent. Sci.* **4**, 144–152 (2018).
- Hoffmann, R. & Malrieu, J.-P. Simulation vs. understanding: a tension, in quantum chemistry and beyond. Part A. Stage setting. *Angew. Chem.* **132**, 12690–12710 (2020).
- Hoffmann, R. & Malrieu, J.-P. Simulation vs. understanding: a tension, in quantum chemistry and beyond. Part B. The march of simulation, for better or worse. *Angew. Chem. Int. Ed.* **59**, 13156–13178 (2020).
- Hoffmann, R. & Malrieu, J.-P. Simulation vs. understanding: a tension, in quantum chemistry and

- beyond. Part C. Toward concision. *Angew. Chem. Int. Ed.* **59**, 13694–13710 (2020).
8. Marcus, G. The next decade in AI: four steps towards robust artificial intelligence. Preprint at *arXiv* 2002.06177 (2020).
 9. Thaler, J. Designing an AI physicist. *CERN Courier*, <https://cerncourier.com/a/designing-an-ai-physicist/> (2021).
 10. Potochnik, A. The diverse aims of science. *Stud. Hist. Philos. Sci. A* **53**, 71–80 (2015).
 11. Potochnik, A. *Idealization and the Aims of Science* (Univ. Chicago Press, 2017).
 12. de Regt, H. W. *Understanding Scientific Understanding* (Oxford Univ. Press, 2017).
 13. De Regt, H. W. & Dieks, D. A contextual approach to scientific understanding. *Synthese* **144**, 137–170 (2005).
 14. Boden, M. A. *Mind as Machine: A History of Cognitive Science* (Oxford Univ. Press, 2008).
 15. Doran, D., Schulz, S. & Besold, T. R. What does explainable AI really mean? A new conceptualization of perspectives. Preprint at *arXiv* 1710.00794 (2017).
 16. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4793–4813 (2020).
 17. Burkart, N. & Huber, M. F. A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **70**, 245–317 (2021).
 18. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer, 2019).
 19. Michie, D. in *Proc. 3rd European Conference on European Working Session on Learning*, 107–122 (ACM, 1988).
 20. Lavin, A. et al. Simulation intelligence: Towards a new generation of scientific methods. Preprint at *arXiv* 2112.03235 (2021).
 21. Goldman, B., Kearnes, S., Kramer, T., Riley, P. & Walters, W. P. Defining levels of automated chemical design. *J. Med. Chem.* **65**, 7073–7087 (2022).
 22. Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *NPJ Syst. Biol. Appl.* **7**, 29 (2021).
 23. Feyerabend, P. *Against Method* (Verso, 1993).
 24. De Regt, H. W. Understanding, values, and the aims of science. *Philos. Sci.* **87**, 921–932 (2020).
 25. Hempel, C. G. *Aspects of Scientific Explanation* (Free Press, 1965).
 26. Schrödinger, E. *Nature and the Greeks' and 'Science and Humanism'* (Cambridge Univ. Press, 1996).
 27. De Regt, H. W. Visualization as a tool for understanding. *Perspect. Sci.* **22**, 377–396 (2014).
 28. Friedman, M. Explanation and scientific understanding. *J. Philos.* **71**, 5–19 (1974).
 29. Kitcher, P. Explanatory unification. *Philos. Sci.* **48**, 507–531 (1981).
 30. Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Phys.* **43**, 172–198 (1927).
 31. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
 32. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
 33. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
 34. Iten, R., Metzger, T., Wilming, H., Del Rio, L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2020).
 35. Seif, A., Hafezi, M. & Jarzynski, C. Machine learning the thermodynamic arrow of time. *Nat. Phys.* **17**, 105–113 (2021).
 36. Udrescu, S.-M. & Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **6**, eaay2631 (2020).
 37. Lemos, P., Jeffrey, N., Cranmer, M., Ho, S. & Battaglia, P. Rediscovering orbital mechanics with machine learning. Preprint at *arXiv* 2202.02306 (2022).
 38. Casalino, L. et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
 39. Fogarty, C. A., Harbison, A. M., Dugdale, A. R. & Fadda, E. How and why plants and human n-glycans are different: Insight from molecular dynamics into the “glycoblacks” architecture of complex carbohydrates. *Belstein J. Org. Chem.* **16**, 2046–2056 (2020).
 40. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761 (2021).
 41. Gigan, S., Krzakala, F., Daudet, L. & Carron, I. Artificial intelligence: from electronics to optics. *Photoniques* **104**, 49–52 (2020).
 42. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
 43. Quantum, G. A. et al. Hartree-Fock on a superconducting qubit quantum computer. *Science* **369**, 1084–1089 (2020).
 44. Zhang, J. et al. Observation of a discrete time crystal. *Nature* **543**, 217–220 (2017).
 45. Schweizer, C. et al. Floquet approach to Z2 lattice gauge theories with ultracold atoms in optical lattices. *Nat. Phys.* **15**, 1168–1173 (2019).
 46. Martinez, E. A. et al. Real-time dynamics of lattice gauge theories with a few-qubit quantum computer. *Nature* **534**, 516–519 (2016).
 47. Cao, Y. et al. Quantum chemistry in the age of quantum computing. *Chem. Rev.* **119**, 10856–10915 (2019).
 48. Gross, C. & Bloch, I. Quantum simulations with ultracold atoms in optical lattices. *Science* **357**, 995–1001 (2017).
 49. O'Connor, M. et al. Sampling molecular conformations and dynamics in a multiuser virtual reality framework. *Sci. Adv.* **4**, eaat2731 (2018).
 50. Probst, D. & Reymond, J.-L. Exploring DrugBank in virtual reality chemical space. *J. Chem. Inf. Model.* **58**, 1731–1735 (2018).
 51. Schmid, J. R., Ernst, M. J. & Thiele, G. Structural chemistry 2.0: combining augmented reality and 3D online models. *J. Chem. Educ.* **97**, 4515–4519 (2020).
 52. Foley, M. et al. A 3D view of Orion: I. Barnard's Loop. *Authorea*. <https://doi.org/10.22541/au.165540488.82174026/v1> (2022).
 53. Hill, E., Cherston, J., Goldfarb, S. & Paradiso, J. A. in *Proc. 38th International Conference on High Energy Physics*, 1042 (2016).
 54. Zanella, A. et al. Sonification and sound design for astronomy research, education and public engagement. *Nat. Astron.* <https://doi.org/10.1038/s41550-022-01721-z> (2022).
 55. Turing, A. M. Computing machinery and intelligence. *Mind* **50**, 433–460 (1950).
 56. Lehman, J. et al. The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. *Artif. Life* **26**, 274–306 (2020).
 57. Pickard, C. J. & Needs, R. Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
 58. Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R. & Zeilinger, A. Automated search for new quantum experiments. *Phys. Rev. Lett.* **116**, 090405 (2016).
 59. Krenn, M., Erhard, M. & Zeilinger, A. Computer-inspired quantum experiments. *Nat. Rev. Phys.* **2**, 649–661 (2020).
 60. Pickard, C. J. & Needs, R. Highly compressed ammonia forms an ionic crystal. *Nat. Mater.* **7**, 775–779 (2008).
 61. Krenn, M., Hochrainer, A., Lahiri, M. & Zeilinger, A. Entanglement by path identity. *Phys. Rev. Lett.* **118**, 080401 (2017).
 62. Krenn, M., Gu, X. & Zeilinger, A. Quantum experiments and graphs: multiparty states as coherent superpositions of perfect matchings. *Phys. Rev. Lett.* **119**, 240403 (2017).
 63. Krenn, M., Kottmann, J., Tischler, N. & Aspuru-Guzik, A. Conceptual understanding through efficient automated design of quantum optical experiments. *Phys. Rev. X* **11**, 031044 (2021).
 64. Malhotra, P., Vig, L., Shroff, G. & Agarwal, P. in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, 89–94 (2015).
 65. Collaboration, A. Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV *pp* collisions in the ATLAS detector. *Phys. Rev. Lett.* **125**, 131801 (2020).
 66. Collaboration, C. Probing effective field theory operators in the associated production of top quarks with a Z boson in multilepton final states at $\sqrt{s} = 13$ TeV. *J. High Energy Phys.* **2021**, 83 (2021).
 67. Park, S. E., Rankin, D., Udrescu, S.-M., Yunus, M. & Harris, P. Quasi-anomalous knowledge: searching for new physics with embedded knowledge. *J. High Energy Phys.* **2021**, 30 (2021).
 68. Karagiorgi, G., Kasieczka, G., Kravitz, S., Nachman, B. & Shih, D. Machine learning in the search for new fundamental physics. *Nat. Rev. Phys.* **4**, 399–412 (2022).
 69. Schwartz, M. D. Modern machine learning and particle physics. Preprint at *arXiv* 2103.12226 (2021).
 70. Kasieczka, G. et al. The LHC Olympics 2020: a community challenge for anomaly detection in high energy physics. *Rep. Prog. Phys.* **84**, 124201 (2021).
 71. Dawid, A., Huembeli, P., Tomza, M., Lewenstein, M. & Dauphin, A. Hessian-based toolbox for reliable and interpretable machine learning in physics. *Mach. Learn. Sci. Technol.* **3**, 015002 (2021).
 72. Koh, P. W. & Liang, P. in *Proc. 34th International Conference on Machine Learning*, 1885–1894 (PMLR, 2017).
 73. Yu, H., Mineev, I. & Varshney, L. R. A group-theoretic approach to computational abstraction: Symmetry-driven hierarchical clustering. Preprint at *arXiv* 1807.11167 (2018).
 74. Dehmamy, N., Walters, R., Liu, Y., Wang, D. & Yu, R. Automatic symmetry discovery with lie algebra convolutional network. *Adv. Neural Inf. Process. Syst.* **34**, 2503–2515 (2021).
 75. Nigam, A. et al. Assigning confidence to molecular property prediction. *Expert Opin. Drug Discov.* **16**, 1009–1023 (2021).
 76. Davies, A. et al. Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
 77. Douglas, M. R. Machine learning as a tool in theoretical science. *Nat. Rev. Phys.* **4**, 145–146 (2022).
 78. King, R. D. et al. The automation of science. *Science* **324**, 85–89 (2009).
 79. Bédard, A.-C. et al. Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220–1225 (2018).
 80. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
 81. Coley, C. W. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
 82. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
 83. Chatterjee, S., Guidi, M., Seeburger, P. H. & Gilmore, K. Automated radial synthesis of organic molecules. *Nature* **579**, 379–384 (2020).
 84. Grizou, J., Points, L. J., Sharma, A. & Cronin, L. A curious formulation robot enables the discovery of a novel protocol behavior. *Sci. Adv.* **6**, eaay4237 (2020).
 85. Moon, H. et al. Machine learning enables completely automatic tuning of a quantum device faster than human experts. *Nat. Commun.* **11**, 4161 (2020).
 86. Dalgaard, M., Motzoi, F., Sørensen, J. J. & Sherson, J. Global optimization of quantum dynamics with AlphaZero deep exploration. *NPJ Quantum Inf.* **6**, 6 (2020).
 87. Larsen, P. & Von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84**, 575–603 (2010).
 88. Reisz, N. et al. Loss of sustainability in scientific work. *New J. Phys.* **24**, 053041 (2022).
 89. Evans, J. A. & Foster, J. G. Metaknowledge. *Science* **331**, 721–725 (2011).
 90. Clauset, A., Larremore, D. B. & Sinatra, R. Data-driven predictions in the science of science. *Science* **355**, 477–480 (2017).
 91. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
 92. Wang, D. & Barabási, A.-L. *The Science of Science* (Cambridge Univ. Press, 2021).
 93. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
 94. Olivetti, E. A. et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **7**, 041317 (2020).
 95. Schwalbe-Koda, D., Jensen, Z., Olivetti, E. & Gómez-Bombarelli, R. Graph similarity drives zeolite diffusionless transformations and intergrowth. *Nat. Mater.* **18**, 1177–1181 (2019).
 96. Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl Acad. Sci. USA* **112**, 14569–14574 (2015).
 97. Krenn, M. & Zeilinger, A. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proc. Natl Acad. Sci. USA* **117**, 1910–1916 (2020).

98. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at *arXiv* 1810.04805 (2018).
99. Brown, T. B. et al. Language models are few-shot learners. Preprint at *arXiv* 2005.14165 (2020).
100. Hamilton, W. L., Ying, R. & Leskovec, J. in *Proc. 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 1025–1035 (2017).
101. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1–15 (2018).
102. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **8**, 42200–42216 (2020).
103. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
104. Mahendran, A. & Vedaldi, A. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 5188–5196 (2015).
105. Mordvintsev, A., Olah, C. & Tyka, M. Inceptionism: going deeper into neural networks. <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html> (2015).
106. Shen, C., Krenn, M., Eppel, S. & Aspuru-Guzik, A. Deep molecular dreaming: Inverse machine learning for de-novo molecular design and interpretability with surjective representations. *Mach. Learn. Sci. Technol.* **2**, 03LT02 (2021).
107. Burgess, C. P. et al. Understanding disentangling in β -VAE. Preprint at *arXiv* 1804.03599 (2018).
108. Wetzel, S. J. & Scherzer, M. Machine learning of explicit order parameters: From the Ising model to SU(2) lattice gauge theory. *Phys. Rev. B* **96**, 184410 (2017).
109. Wetzel, S. J., Melko, R. G., Scott, J., Panju, M. & Ganesh, V. Discovering symmetry invariants and conserved quantities by interpreting siamese neural networks. *Phys. Rev. Res.* **2**, 033499 (2020).
110. Friederich, P., Krenn, M., Tamblyn, I. & Aspuru-Guzik, A. Scientific intuition inspired by machine learning-generated hypotheses. *Mach. Learn. Sci. Technol.* **2**, 025027 (2021).
111. Flam-Shepherd, D. et al. Learning interpretable representations of entanglement in quantum optics experiments using deep generative models. *Nat. Mach. Intell.* **4**, 544–554 (2022).
112. Wellawatte, G. P., Seshadri, A. & White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.* **13**, 3697–3705 (2022).
113. McGrath, T. et al. Acquisition of chess knowledge in AlphaZero. Preprint at *arXiv* 2111.09259 (2021).
114. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
115. Gentile, A. A. et al. Learning models of quantum systems from experiments. *Nat. Phys.* **17**, 837–843 (2021).
116. Cranmer, M. et al. Discovering symbolic models from deep learning with inductive biases. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (NeurIPS, 2020).
117. Georgescu, I. How machines could teach physicists new scientific concepts. *Nat. Rev. Phys.* <https://doi.org/10.1038/s42254-022-00497-5> (2022).
118. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proc. Natl Acad. Sci. USA* **117**, 30055–30062 (2020).
119. Raayoni, G. et al. Generating conjectures on fundamental constants with the Ramanujan Machine. *Nature* **590**, 67–73 (2021).
120. Wagner, A. Z. Constructions in combinatorics via neural networks. Preprint at *arXiv* 2104.14516 (2021).
121. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
122. Schmidhuber, J. in *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, 48–76 (Springer, 2008).
123. Pathak, D., Agrawal, P., Efron, A. A. & Darrell, T. in *Proc. 34th International Conference on Machine Learning*, 2778–2787 (PMLR, 2017).
124. Thiede, L. A., Krenn, M., Nigam, A. & Aspuru-Guzik, A. Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning. *Mach. Learn. Sci. Technol.* **3**, 035008 (2022).
125. Varshney, L. R., Rajani, N. F. & Socher, R. Explaining creative artifacts. Preprint at *arXiv* 2010.07126 (2020).
126. Varshney, L. R. et al. A big data approach to computational creativity: The curious case of Chef Watson. *IBM J. Res. Dev.* **63**, 7:1–7:18 (2019).
127. Itti, L. & Baldi, P. Bayesian surprise attracts human attention. *Vision Res.* **49**, 1295–1306 (2009).
128. Schmid, U., Zeller, C., Besold, T., Tamaddoni-Nezhad, A. & Muggleton, S. in *Proc. International Conference on Inductive Logic Programming*, 52–67 (Springer, 2016).
129. Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A. & Besold, T. Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Mach. Learn.* **107**, 1119–1140 (2018).
130. Feigenbaum, E. A. Some challenges and grand challenges for computational intelligence. *J. ACM* **50**, 32–40 (2003).
131. de Regt, H. W. & Gijsbers, V. in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* (eds Grimm, S. R., Baumberg, C. & Ammon, S.) 50–75 (2017).

Acknowledgements

The authors thank A. Alexandrova, R. Amaro, C. Berlinguette, L. Chong, C. Cisneros, A. Cooper, G. Day, F.-X. Coudert, L. Cronin, E. Fadda, R. Gomez-Bombarelli, L. Gonzalez, J. Hachmann, R. Hoffmann, J. Halborg Jensen, E. R. Johnson, L. Kamerlin, H. J. Kulik, J.-P. Malrieu, A. Milo, F. Noe, J. Kehlet Nørskov, A. Oganov, J. Perez-Mercader, C. Pickard, M. Reiher, J.-L. Reymond, D. Salahub, S. Sanvito, F. Schoenebeck, I. Siepmann, A. Sodt, I. Tamblyn, D. Truhlar, A. Tkatchenko, K. Tsuda, A. Varnek, T. Vegge, A. von Lilienfeld and E. Zurek for answering our questions on understanding, and N. Tischler and R. Fickler for their valuable comments on the manuscript. A.A.-G. and his group acknowledge generous support from the Canada 150 Research Chairs Program, the University of Toronto and Anders G. Frøseth. M.K. acknowledges support from the FWF (Austrian Science Fund) via the Erwin Schrödinger Fellowship no. J4309. R.P. acknowledges funding through a Postdoc.Mobility fellowship by the Swiss National Science Foundation (SNSF, project no. 191127).

Author contributions

The authors contributed substantially to the discussion of content. M.K., R.P. and S.Y.G. researched the data for the article. M.K. wrote the article, with substantial contributions from all authors. A.A.-G. initiated the project.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Physics thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1038/s42254-022-00518-3>.

© Springer Nature Limited 2022