

On scientific understanding with artificial intelligence

Mario Krenn,^{1, 2, 3, 4,*} Robert Pollice,^{2, 3} Si Yue Guo,² Matteo Aldeghi,^{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 Jinich,⁷ Akshat Kumar Nigam,^{2, 3} Zhenpeng Yao,^{2, 8, 9, 10} and Alán Aspuru-Guzik^{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, Canada.

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

⁴Vector Institute for Artificial Intelligence, Toronto, Canada.

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

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

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

⁸Center of Hydrogen Science, Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai 200240, China.

⁹The State Key Laboratory of Metal Matrix Composites,
School of Materials Science and Engineering, Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai 200240, China.

¹⁰Innovation Center for Future Materials, Zhangjiang Institute for Advanced Study,
Shanghai Jiao Tong University, 429 Zhangheng Road, Shanghai 201203, China.

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

Imagine an oracle that correctly predicts the outcome of every particle physics experiment, the products of every chemical reaction, or the function of every protein. Such an oracle would revolutionize science and technology as we know them. However, as scientists, we would not be satisfied with the oracle itself. We want more. We want to comprehend how the oracle conceived these predictions. This feat, denoted as scientific understanding, has frequently been recognized as the essential aim of science. Now, the ever-growing power of computers and artificial intelligence poses one ultimate question: How can advanced artificial systems contribute to scientific understanding or achieve it autonomously?

We are convinced that this is not a mere technical question but lies at the core of science. Therefore, here we set out to answer where we are and where we can go from here. We first seek advice from the philosophy of science to *understand scientific understanding*. Then we review the current state of the art, both from literature and by collecting dozens of anecdotes from scientists about how they acquired new conceptual understanding with the help of computers. Those combined insights help us to define three dimensions of android-assisted scientific understanding: The android as a I) computational microscope, II) resource of inspiration and the ultimate, not yet existent III) agent of understanding. For each dimension, we explain new avenues to push beyond the status quo and unleash the full power of artificial intelligence's contribution to the central aim of science. We hope our perspective inspires and focuses research towards androids that get new scientific understanding and ultimately bring us closer to true artificial scientists.

I. INTRODUCTION

Artificial Intelligence (A.I.) has recently been called a “new tool in the box for scientists”[1] and that “machine learning with artificial networks is revolutionizing science”[2]. Additionally, it has been conjectured “that machines could have a significantly more creative role in future research.” [3]. For instance, it has even been postulated that “[t]he new goal of theoretical chemistry should be that of providing access to a chemical ‘oracle’: an A.I. environment which can help humans solve problems, associated with the fundamental chemical questions of the fourth industrial revolution [...], in a way such that the human cannot

distinguish between this and communicating with a human expert” [4].

However, this excitement has not been shared among all scientists. Specifically, it has been questioned whether advanced computational approaches can go beyond *numerics* [5–9] and contribute fundamentally to one of the essential aims of science, that is, gaining of new scientific understanding [10–12].

In this work, we address how artificial systems can contribute to scientific understanding – specifically, what is the state-of-the-art and how we can push further. Besides a thorough literature review, we surveyed dozens of scientists at the interface of biology, chemistry or physics on the one hand, and artificial intelligence and advanced computational methods. These personal narratives focus on the concrete discovery process of ideas and are a vital augmentation to the scientific literature. We put the literature and personal accounts in the context of a philosophi-

* mario.krenn@mpl.mpg.de
† alan@aspuru.com

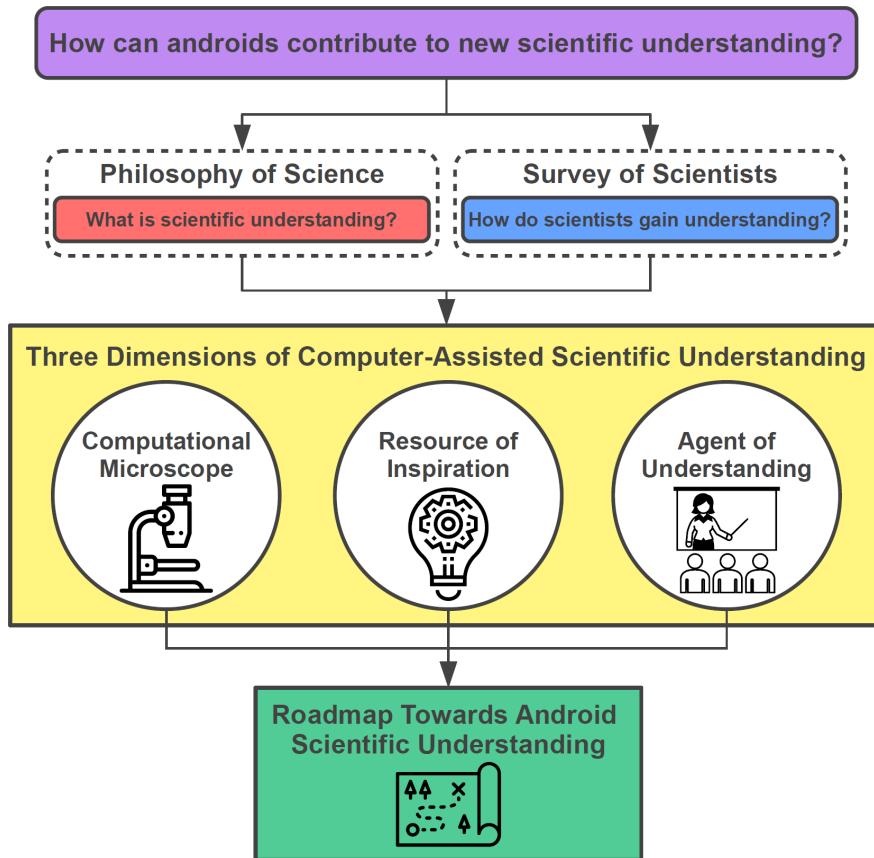


Figure 1. How can Androids contribute to new scientific understanding? In addition to scientific literature, we take inspiration from the philosophy of science and from dozens of stories provided by active computational natural scientists. Thereby we identify three fundamental dimensions of computer-assisted scientific understanding. From there, we look into the future and develop a roadmap on how to develop Androids that can contribute to understanding – the essential aim of science.

cal theory of *Scientific Understanding* recently developed by Dennis Dieks and Henk de Regt [12, 13], who was awarded the Lakatos Award in 2019 for the development of this theory. We thereby introduce three fundamental dimensions for scientific androids¹ contribution towards new scientific understanding:

- I) Androids acting as a microscope in the responses, i.e., akin to 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.
- II) Androids acting as muses, i.e., sources of inspiration for new concepts and ideas that are subsequently understood and generalized by human scientists.

¹We encapsulate all advanced artificial computational systems under *androids*, independent of their working principles. In this way, we are focusing on the operational objective rather than the methodology.

III) Lastly, in an ultimate dimension of android-assisted scientific understanding, computers are the agents of understanding. While we have not found any evidence of computers acting as true agents of understanding in science yet, we outline important characteristics of such an artificial system of the future and potential ways to achieve it.

In the first two dimensions, the android enables humans to gain new scientific understanding while in the last one the machine gains understanding itself. These classes enable us to layout a vibrant and mostly unexplored field of research, which will hopefully manifest itself as a guiding star for future developments of artificial intelligence in the natural sciences.

The goal of this perspective is to put *Scientific Understanding* back to the limelight – where we are convinced it belongs. We hope to inspire physicists, chemists and biologists and A.I. researchers to go beyond the status quo, focus on these central aims of science, and revolutionize computer-assisted scientific understanding. In that way, we believe that androids

will become true agents of understanding that contribute to science in a fundamental and creative way.

II. SCIENTIFIC UNDERSTANDING

Let us imagine an oracle providing non-trivial predictions that are always true. While such a hypothetical system would have a very significant scientific impact, scientists would not be satisfied. We 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 that really mean? Can we find criteria for scientific understanding? To do that, we seek guidance from the field of philosophy of science. Notably, while hardly any scientist would argue against “understanding” as an essential aim of science (next to explanation, description and prediction [14]), this view was not always accepted by philosophers. Specifically, 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 science [15]. Other philosophers criticized these rather unsatisfying conclusions, and they tried to formalize what *scientific understanding* means. Proposals include that understanding is connected to the ability to build causal models (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[16, 17]) or that understanding corresponds to providing unification [18, 19].

In recent years, Henk de Regt and Dennis Dieks have developed a new theory of scientific understanding, which is both contextual and pragmatic [12–14]. Importantly, they find that techniques such as visualization or unification are “tools for understanding”, thereby unifying previous ideas in one general framework. Their theory is agnostic to the specific “tool” being used, making it particularly useful for application in scientific disciplines. They extend crucial insights by Werner Heisenberg [20] and rather than introducing mere 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*”. Put simply, 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].

Concretely, de Regt and Dieks define two inter-linked criteria:

1. **Criterion of Understanding Phenomena:** A phenomenon P can be understood if a theory T of P exists that is intelligible.
2. **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 of one particular strength: We can use it experimentally to 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 also coincides with Angelika Potocznik’s argument that “*understanding requires successful mastery, in some sense, of the target of understanding*”[11]. We will follow this approach and, consequently, here explore its relationship to the role of A.I. in science. Accordingly, we believe we can significantly advance A.I.’s contribution to this central aim of Science if we have a clear picture of how scientists gain conceptual understanding, and instil it to artificial systems afterwards. We approach this goal by applying ideas of de Regt and Dieks directly to android assisted science (and ultimately, to android scientists themselves).

III. WHAT IS NEXT?

A. Beyond Re-Discovery

In recent years, scientists at the interface between A.I. and the natural sciences tried to rediscover scientific laws or concepts with machines. The question is, however, whether an android is capable of contributing to new scientific understanding if it can *rediscover* physical laws and concepts, such as the heliocentric world view [21], the arrow of time [22] or mechanical equations of motions [23]? We believe that this is not guaranteed. The human creators of these androids know what they are looking for in these case studies. Therefore, it is unclear how both conscious and unconscious biases (in the broadest sense, e.g., by choosing particular representations) in the code or the data analysis can be prevented. Consequently, even if an algorithm can rediscover interesting physical phenomena, we cannot know whether and how they can be used to advance Science by helping to uncover new scientific understanding.

Hence, we believe we need to go beyond rediscovery tasks. Therefore we focus explicitly on the question of how to get *new* scientific understanding.

B. Beyond Discovery

Importantly, other central aims of science such as prediction and discovery can lead to scientific and technological disruptions while not directly contributing to scientific understanding as discussed above [11, 14]. For instance, imagine the hypothetical discovery of the *hitherto* best material for energy storage that could revolutionize batteries. However, this game-changing discovery would not qualify as understanding if chemists could not use the underlying principles fruitfully in other contexts (without computation).

Similarly, the recent breakthrough in protein folding will undoubtedly change the landscape of biochemistry. However, so far, AlphaFold is a black box – an oracle[24, 25]. As such it does not directly provide new scientific understanding in the sense of de Regt and Dieks (but could of course in the future enable humans to gain new scientific understanding). Hence, we believe we must go beyond artificial discoveries in science.

C. Where to go from here?

The ultimate goal is to get new *understanding* from androids. Loosely speaking, we want to find new ideas or concepts that we can apply and use in different situations without (complete) computations.

This article aims to explain precisely what such a goal requires, what previous approaches have achieved, and how we can go further. We want to clearly lay out this underappreciated but essential research question and thereby give a clear goal for the future of A.I. in the natural sciences.

IV. THREE DIMENSIONS OF COMPUTER-ASSISTED UNDERSTANDING

We use scientific literature and personal anecdotes of dozens of scientists, and the context of the philosophy of science, to introduce a new classification of androids contribution to scientific understanding². It helps to see diverse unexplored journeys that can be investigated in the future.

An android can act

- I) as a **computational microscope**, providing information not (yet) attainable by experiment
- II) as a **resource of inspiration** or an *artificial muse*, expanding the scope of human imagination and creativity.

²We call the classification *dimensions*, as they are independent and non-exclusive.

In those two classes, the human scientist is essential to take the new insight and inspiration and develop it to full understanding. Finally, an android can be

- III) an **agent of understanding**, replacing the human in generalizing observations and transferring scientific concepts to new phenomena.

We stress that these three classes should not be understood dogmatically but rather guide future possibilities. In the following sections, based on concrete examples, we discuss each class in more detail and propose avenues for pushing the boundaries of the current computational faculties.

A. Computational microscope for scientific understanding

Microscopes are devices that enable us to investigate objects and phenomena imperceptible to the naked eye. In a similar way, *computational microscopes* enable the investigation of objects or processes that we cannot visualize or probe in any other way. One main objective is to simulate biological, chemical or physical processes that happen at length and time scales not perceivable by experiment.

As we are interested in understanding, the new computer-generated data needs to be generalized to other contexts without complete computation[13]. We show now two concrete examples.

The first example is molecular dynamics simulations of the SARS-CoV-2. The authors uncovered new biological functions that show different behaviours in the open and closed conformations of the spike protein. This explanation changed the view upon glycans in biological systems and inspired new ways to analyze these systems without the need to perform full computations [26].

In the second example, the authors describe how molecular dynamics simulations helped to uncover fundamental patterns called *glycoblcks*. The systematic use of glycoblcks can both be used to understand sequence-structure-property relationships of biomolecules and can also inform the design of synthetic structures with desired functions without the need for simulating the entire system [27].

The next computational microscope

A computational microscope aims to provide data via computation that are not (yet) accessible by experiments that humans can understand. How could we make computational microscopes even more insightful and make it easier for human scientists to use this data to gain scientific understanding? There are two vibrant directions going forwards. First, more advanced computational systems will allow to analyze of

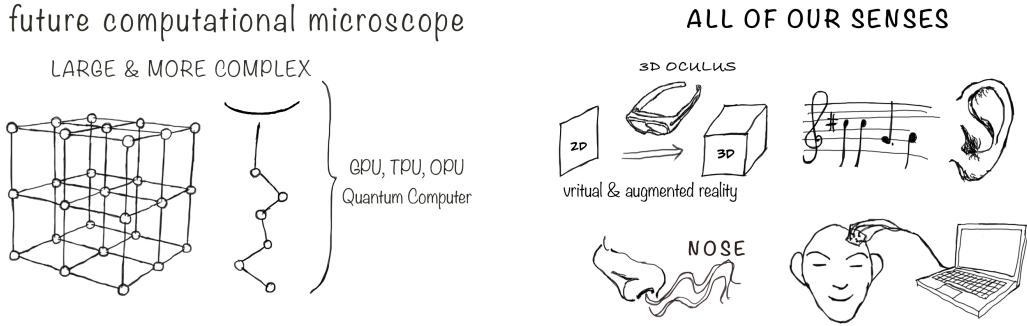


Figure 2. The future computational microscope. We envision two types of advances in the next-generation computational microscopes which aim to advance genuine scientific understanding. First (left), larger and more complex computations will allow the computational observation of phenomena not accessible so far. There, new computational paradigms will play a significant role, such as Graphical Processor Units (GPU), Tensor Processor Units (TPU), Optical Processor Units (OPU) and – ultimately – quantum computers. Second (right), new ways to represent the highly complex data will advance our ability to sense structure and recognize underlying patterns. The involvement of all our senses could, for sensing computer-generated data, be an exciting pathway to advanced understanding.

more complex physical systems. Second, representing the information in a more interpretable way will help to lift the indications from computers to true scientific understanding.

More Complex Systems – One obvious but nevertheless important research direction is increasing the complexity as well as the accuracy of computer simulations[28]. For example, increasing the size of the systems, the time-scale of the simulations, the number of interactions that can be modelled will significantly increase the applicability in complex dynamic systems. In general, this can be achieved by either algorithmic improvements or hardware improvements, or both. In that regard, we expect that modern neural network technologies together with advanced hardware such as GPUs, TPUs or even OPUs [29, 30] will have an enormous impact. Furthermore, the recent progress in experimental quantum computing for quantum chemistry [31] and physics[32–34] promises that entirely new algorithms, based on quantum mechanics itself, will play an important role in this area [35, 36]. Algorithmic improvements could involve adaptive and intelligent resolution during simulation and advanced visualization methods[13], which directly leads to the second future techniques:

Full spectrum of senses – We believe that human scientists can get more out of data if the full capabilities of all our senses are addressed. At the moment, we analyze data largely in (potentially animated) 2-dimensional pictures. As a first step, we believe that real 3D environments (realized either via virtual or augmented reality glasses, or holography) will significantly help in understanding complex systems or complex data. Initial advances in that regard have been demonstrated in the domain of chemistry [37–39], and we expect this to become a standard tool for scientists

to advance scientific understanding. In addition, we expect that going beyond the visual sense can open entirely new ways to experience scientific data. For example, the auditory sense is excellent in detecting structure or symmetries in (periodic) time-dependent data [40]. Furthermore, including the sense of touch, smell and taste could further expand the horizon of experiences. We expect that in order to realize that, physical scientists need to work closely together with psychologists and neurologists (and potentially even with artists), to develop suitable data representations that can be efficiently recognized by scientists with all their senses. An ultimate, admittedly futuristic version of a computational microscope could circumvent the receptors of human senses and instead use a computer-brain interface to enhance further experiencing computed data.

B. Resource of inspiration for scientific understanding

Surprising and creative ideas are the foundation of Science. Computer algorithms are a means to provoke such ideas systematically, thereby significantly accelerating scientific and technological progress. Already 70 years ago, Alan Turing realized that computers could surprise their human creators: “*Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks.*” and “*Naturally I am often wrong, and the result is a surprise for me for by the time the experiment is done these assumptions have been forgotten*”[41].

A much more recent study provides stories by

dozens of researchers of artificial life and evolution. They demonstrate in an impressive way how computer algorithms can surprise their human creators and lead to behaviour that the authors *would denote as creative*[42]. Accordingly, we believe that androids can be artificial muses of Science in a metaphorical sense.

Those examples demonstrate that computers can indeed be used as a source of surprises. But what are the most general ways to get inspirations from computers? And how can they be lifted by humans to true scientific understanding? We will outline a number of ways to develop ways to provoke surprising behaviours of algorithms and use their solutions, internal or external states as a source of inspiration for new scientific ideas.

The future resource of inspiration

Identifying surprises in data – Exceptional data points or unexpected regularities obtained from experiments or simulations can surprise human scientists and inspire new ideas and concepts. Our survey shows that these exceptional points are usually identified by humans, such as the following two examples, which use high-throughput computations in chemistry [43] and quantum optics [44, 45].

The first example deals with a surprising phase of crystal structures in high-pressure physics. There, the authors found an unexpected stable configurations of alternating NH₂ and NH₄ layers, rather than a dense NH₃ phase. The authors 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 [46].

In the second example, a search for new quantum experiments uncovered a solution with considerable larger quantum entanglement than expected. The authors understood the underlying principles and thereby discovered a new concept of entanglement generation [47, 48]. The principle can be used without any computation and, for example, acts now as a new representation in more advanced artificial intelligence systems for quantum physics [49], demonstrating the application of the computer-inspired idea in more general different contexts.

In contrast to these examples and many others from literature and from personal accounts, the anomalies could manifest themselves in a more involved combination of variables, which might be very difficult for humans to grasp. Accordingly, applying advanced statistical methods and machine learning algorithms (e.g., see reference [50]) to this type of problem will be an important future research direction. Exciting

works into the direction of autonomous anomaly detection have been applied on scientific data from the Large Hadron Collider (LHC) at CERN [51–53]. Such techniques have the potential to identify new physics signatures, which can then be conceptualized and understood by human physicists [54, 55]. Neural networks that autonomously discover symmetries could become an efficient discovery tool for outliers in scientific data where the underlying rules might not be known beforehand [56, 57].

Estimating the confidence of predictions will be one method to directly search for anomalies in data [58]. The ability to uncover hidden regularities was very recently demonstrated in mathematics, where an A.I. hinted on relations between previously unconnected invariants in knot theory, which allowed mathematicians to conjecture and prove new theorems [59]. Alternatively, an A.I. 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 A.I. uncover hidden patterns or irregularities in scientific data previously overlooked by humans, which leads to new ideas and, ultimately, to new conceptual understanding. As of now, we are not aware of cases like that.

The data points for these systems could be obtained from computational methods (involving those described in section IV A), with exciting opportunities for mathematics or theoretical physics [60]. Alternatively, the data could be obtained directly from experiments. Here we can imagine a closed-loop approach where an algorithm tries to explore the environment and steer the exploration into unexpected regions. If the data-source is an experiment, this future system will require access to complex lab automation with large parameter spaces to explore, as demonstrated recently in biology [61], chemistry [62–67] or physics [68, 69].

Identifying surprises in the scientific literature – The number of scientific papers in essentially every scientific domain is growing enormously[70]. Consequently, researchers have to specialise in narrow sub-disciplines, 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[71–74] and identify exceptional and surprising phenomena for further investigation. While 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. Recently, this technique has been applied in the domain of material science [75] and rediscovered central scientific con-

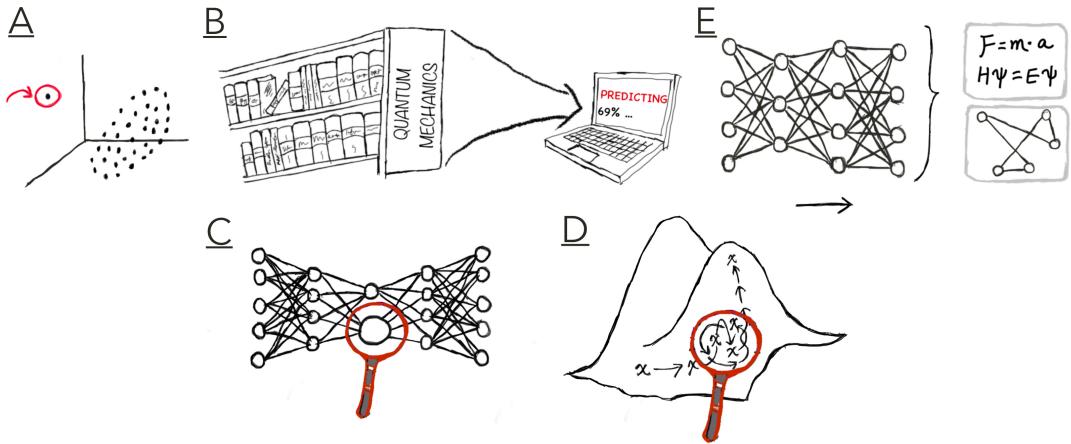


Figure 3. The future re-source of inspiration: An android could act as a computational muse and inspire the human scientist by (A) identifying surprises in data, (B) identifying surprises in the scientific literature, (C) finding surprising concepts by inspecting models or (D) by probing the behaviour of artificial agents or (E) my finding new concepts from interpretable solutions.

cepts 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 material science to extract knowledge from the scientific literature and investigate it systematically by A.I. technologies[76], and can lead to complex scientific conclusions as demonstrated for instance on zeolite transformations [77].

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 [78, 79]. 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 strategies [78] and in quantum physics to predict and suggest future research directions [79]. Advances in A.I. technology could improve this type of system significantly. For example, natural language processing architectures such as BERT [80], or GPT3 [81] 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 networks [82].

Surprising concepts by inspecting models – We also expect considerable progress by rationalising what A.I. algorithms have learned in order to solve a specific problem, i.e., explainable or interpretable A.I. [83–86]. One idea towards this goal is inspired by DeepDreaming, a method first used in computer vision [87, 88]. Put simply, the idea is to invert a neural network

and probe its behaviour. Recently, this approach has been applied to rediscover thermodynamical properties [22], and design principles for functional molecules [89]. An alternative and remarkable application is the *disentanglement of variables* in neural networks [90]. The goal is to understand the internal representation the neural network has learned. Recently, astronomical data, represented in geocentric coordinates, was used to train a neural network and disentanglement of variables enabled the rediscovery of heliocentric coordinates via the internal representation of the model [21]. In a related study, using gradient boosting with decision trees, feature importance has been used to explain properties of molecules, and quantum optics circuits [91]. Related to this is a study where the internal representation of an unsupervised deep generative model for quantum experiments has been inspected to understand the model’s internal worldview[92]. In the chemical domain, counterfactual explanations for machine learning models have been demonstrated to produce rationale behind a model’s prediction. Counterfactual explanations illustrate what differences to an event or instance would generate a change in an outcome. Wellawatte et al. [93] 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 great future potentials for opening the black-box of AI in science. Albeit not in science, recent work has investigated what the chess-playing A.I AlphaZero has learned about chess and how human-like knowledge is encoded in the internal representation [94]. 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. Progress towards resolving that challenge will be essential in the near future to inspire new scientific ideas.

New concepts from interpretable solutions – Rather than getting inspirations from the A.I. algorithms themselves, scientists can also be surprised by the corresponding solutions. When solutions are represented in an interpretable way, they can provoke new ideas and lead to new concepts. An example of interpretable representations 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 publications demonstrated extracting symbolic models from experimental data of mechanical systems [23, 95], of quantum systems [96] and in astronomy [97]. It will be exciting to see how these approaches, e.g., combined with methods such as causal inference [98], 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[99, 100], and we foresee similar approaches making a significant impact in the physical sciences as well.

One concrete, recent 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 [101]. 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 [97]) was able to extract reasonable mathematical expressions for the learned behaviour. Interestingly, besides the equations of motions, the method simultaneously predicts the masses of the planetary objects correctly. The technique required the assumption of several symmetries and other physical laws. It will be interesting to see whether these prerequisites can be reduced further and how related approaches can be applied to modern physics questions.

Another concrete example of this methodology has been showcased in the field of quantum optics [49]. There, an A.I. 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 it 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 solve a scientific problem [102]. Algorithms that take actions 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, e.g., maximise a specific property of a physical system, intrinsic rewards such as artificial curiosity can be implemented [103, 104]. 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 universes [105] and robot agents in real laboratories [67] that curiosity is an efficient exploration strategy. Alternative intrinsic rewards for artificial agents are *computational creativity*[106, 107] and *surprise* [108]. These intrinsic rewards can produce exceptional and unexpected solutions, ultimately inspiring human scientists.

C. Agent of Understanding

The third and final class we consider are algorithms that can autonomously acquire new scientific understanding, a feat that has neither been described by 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 computer programs could look like.

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 an insight is new, at least in the context of a specific scientific domain that requires access to the knowledge of a scientific field.

Secondly, de Regt emphasized the importance of underlying scientific theories that allow us to recognize qualitatively characteristic consequences [12]. It is not enough to simply interpolate data points or predict new ones using advanced statistical methods such as machine learning. 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 contexts [49]. 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 android 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 for Scientific Understanding I:

An android 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 Dieks [13]. Let us go one step further and imagine that there is an android 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 transfer the newly acquired understanding to themselves. This allows us to propose the second sufficient condition for agents of understanding:

Condition for Scientific Understanding II:

An android 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 test [41] or the Feigenbaum test[109] (or adaptations thereof in the natural sciences such as the Chemical Turing Test or the Feynman Test [4]):

The Scientific Understanding Test:

A human (the student) interacts with a teacher, either a human or an android 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 independently³. If the referee cannot distinguish between the qualities of their non-trivial explanations in various contexts, we argue that the teacher has scientific understanding.

³In principle, there is no reason for the student or the referee not to be androids. However, to keep the test as simple as possible, we want to keep the number of possible variations small.

This implies that *humans* need to understand the new concepts that androids devised. If a machine truly understands something, it will be able to explain it and transfer the understanding to someone else.⁴ We believe that this should always be possible, even if the understanding is far beyond what human experts know at this point. We envision that computers will use advanced human-computer interaction techniques together with the tools we described for (future) computational microscopes.

Additionally, scientific discussions between a human and a computer could be realized using advanced queries in natural language processing tools such as BIRD [80] or GPT-3 [81]. 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.⁵ We are optimistic that more efforts will be directed at developing the necessary technologies, which will lead to ever more convincing demonstrations of android scientists acting as true agents of understanding in the future.

V. CONCLUSION

Undoubtedly, advanced computational methods in general and artificial intelligence specifically will further revolutionize how scientists investigate the secrets of our world. We outline how these new methods can directly contribute to one of the main aims of science, namely acquiring new scientific understanding. We suspect that significant future progress in the use of androids 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 androids into true agents of understanding that will directly contribute to one of the most essential aims of science, namely Scientific Understanding.

ACKNOWLEDGEMENTS

The authors thank Anastassia Alexandrova, Rommie Amaro, Curtis Berlinguette, Lillian Chong, Gerardo Cisneros, Andy Cooper, Graeme Day,

⁴We leave aside the question whether the explanation of the android is true or false. It has been argued that also false theories can lead to genuine understanding [110].

⁵We 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.

Francois-Xavier Coudert, Lee Cronin, Elisa Fadda, Rafael Gomez-Bombarelli, Leticia Gonzalez, Johannes Hachmann, Roald Hoffmann, Jan Halborg Jensen, Erin R. Johnson, Lynn Kamerlin, Heather J. Kulik, Jean-Paul Malrieu, Anat Milo, Frank Noe, Jens Kehlet Nørskov, Artem Oganov, Juan Perez-Mercader, Chris Pickard, Markus Reiher, Jean-Louis Reymond, Dennis Salahub, Stefano Sanvito, Franziska Schoenebeck, Ilja Siepmann, Alex Sodt, Isaac Tamblyn, Donald Truhlar, Alexandre Tkatchenko, Koji Tsuda, Alexandre Varnek, Tejs Vegge, Anatole von

Lilienfeld and Eva Zurek for answering our questions on understanding, Xuemei Gu for Figure 2 and 3, and Nora Tischler and Robert Fickler for 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).

-
- [1] Lenka Zdeborová, “New tool in the box,” *Nature Physics* **13**, 420–421 (2017).
- [2] Thomas Fösel, Petru Tighineanu, Talitha Weiss, and Florian Marquardt, “Reinforcement learning with neural networks for quantum feedback,” *Physical Review X* **8**, 031084 (2018).
- [3] Alexey A Melnikov, Hendrik Poulsen Nautrup, Mario Krenn, Vedran Dunjko, Markus Tiersch, Anton Zeilinger, and Hans J Briegel, “Active learning machine learns to create new quantum experiments,” *Proceedings of the National Academy of Sciences* **115**, 1221–1226 (2018).
- [4] Alán Aspuru-Guzik, Roland Lindh, and Markus Reiher, “The matter simulation (r)evolution,” *ACS central science* **4**, 144–152 (2018).
- [5] Roald Hoffmann and Jean-Paul Malrieu, “Simulation vs. understanding: A tension, in quantum chemistry and beyond. part a. stage setting,” *Angewandte Chemie* **132**, 12690–12710 (2020).
- [6] Roald Hoffmann and Jean-Paul Malrieu, “Simulation vs. understanding: A tension, in quantum chemistry and beyond. part b. the march of simulation, for better or worse,” *Angewandte Chemie International Edition* **59**, 13156–13178 (2020).
- [7] Roald Hoffmann and Jean-Paul Malrieu, “Simulation vs. understanding: A tension, in quantum chemistry and beyond. part c. toward consilience,” *Angewandte Chemie International Edition* **59**, 13694–13710 (2020).
- [8] Gary Marcus, “The next decade in ai: four steps towards robust artificial intelligence,” arXiv:2002.06177 (2020).
- [9] Jesse Thaler, “Designing an ai physicist,” *CERN Courier* **9–10**, 49 (2021).
- [10] Angela Potochnik, “The diverse aims of science,” *Studies in History and Philosophy of Science Part A* **53**, 71–80 (2015).
- [11] Angela Potochnik, *Idealization and the Aims of Science* (University of Chicago Press, 2017).
- [12] Henk W De Regt, *Understanding scientific understanding* (Oxford University Press, 2017).
- [13] Henk W De Regt and Dennis Dieks, “A contextual approach to scientific understanding,” *Synthese* **144**, 137–170 (2005).
- [14] Henk W De Regt, “Understanding, values, and the aims of science,” *Philosophy of Science* **87**, 921–932 (2020).
- [15] Carl G Hempel, *Aspects of scientific explanation* (Free Press New York, 1965).
- [16] Erwin Schrödinger, ‘*Nature and the Greeks’ and ‘Science and Humanism*’ (Cambridge University Press, 1996).
- [17] Henk W De Regt, “Visualization as a tool for understanding,” *Perspectives on science* **22**, 377–396 (2014).
- [18] Michael Friedman, “Explanation and scientific understanding,” *The Journal of Philosophy* **71**, 5–19 (1974).
- [19] Philip Kitcher, “Explanatory unification,” *Philosophy of science* **48**, 507–531 (1981).
- [20] Werner Heisenberg, “über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik,” *Zeitschrift für Physik* **43**, 172–198 (1927).
- [21] Raban Iten, Tony Metger, Henrik Wilming, Lídia Del Rio, and Renato Renner, “Discovering physical concepts with neural networks,” *Physical Review Letters* **124**, 010508 (2020).
- [22] Alireza Seif, Mohammad Hafezi, and Christopher Jarzynski, “Machine learning the thermodynamic arrow of time,” *Nature Physics* **17**, 105–113 (2021).
- [23] Silviu-Marian Udrescu and Max Tegmark, “Ai feynman: A physics-inspired method for symbolic regression,” *Science Advances* **6**, eaay2631 (2020).
- [24] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, *et al.*, “Highly accurate protein structure prediction with alphafold,” *Nature* **596**, 583–589 (2021).
- [25] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, *et al.*, “Highly accurate protein structure prediction for the human proteome,” *Nature* **596**, 590–596 (2021).
- [26] Lorenzo Casalino, Zied Gaieb, Jory A Goldsmith, Christy K Hjorth, Abigail C Dommer, Aoife M Harbison, Carl A Fogarty, Emilia P Barros, Bryn C Taylor, Jason S McLellan, *et al.*, “Beyond shielding: the roles of glycans in the sars-cov-2 spike protein,” *ACS Central Science* **6**, 1722–1734 (2020).
- [27] Carl A Fogarty, Aoife M Harbison, Amy R Dugdale, and Elisa Fadda, “How and why plants and human n-glycans are different: Insight from molecu-

- lar dynamics into the "glycoblocks" architecture of complex carbohydrates," *Beilstein journal of organic chemistry* **16**, 2046–2056 (2020).
- [28] Pascal Friederich, Florian Häse, Jonny Proppe, and Alán Aspuru-Guzik, "Machine-learned potentials for next-generation matter simulations," *Nature Materials* **20**, 750–761 (2021).
- [29] Sylvain Gigan, Florent Krzakala, Laurent Daudet, and Igor Carron, "Artificial intelligence: From electronics to optics," *Photoniques* **104**, 49–52 (2020).
- [30] Xingyuan Xu, Mengxi Tan, Bill Corcoran, Jiayang Wu, Andreas Boes, Thach G Nguyen, Sai T Chu, Brent E Little, Damien G Hicks, Roberto Morandotti, *et al.*, "11 tops photonic convolutional accelerator for optical neural networks," *Nature* **589**, 44–51 (2021).
- [31] Google AI Quantum *et al.*, "Hartree-fock on a superconducting qubit quantum computer," *Science* **369**, 1084–1089 (2020).
- [32] Jiehang Zhang, PW Hess, A Kyriyanidis, P Becker, A Lee, J Smith, G Pagano, I-D Potirniche, Andrew C Potter, A Vishwanath, *et al.*, "Observation of a discrete time crystal," *Nature* **543**, 217–220 (2017).
- [33] Christian Schweizer, Fabian Grusdt, Moritz Berngruber, Luca Barbiero, Eugene Demler, Nathan Goldman, Immanuel Bloch, and Monika Aidelsburger, "Floquet approach to z2 lattice gauge theories with ultracold atoms in optical lattices," *Nature Physics* **15**, 1168–1173 (2019).
- [34] Esteban A Martinez, Christine A Muschik, Philipp Schindler, Daniel Nigg, Alexander Erhard, Markus Heyl, Philipp Hauke, Marcello Dalmonte, Thomas Monz, Peter Zoller, *et al.*, "Real-time dynamics of lattice gauge theories with a few-qubit quantum computer," *Nature* **534**, 516–519 (2016).
- [35] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, *et al.*, "Quantum chemistry in the age of quantum computing," *Chemical reviews* **119**, 10856–10915 (2019).
- [36] Christian Gross and Immanuel Bloch, "Quantum simulations with ultracold atoms in optical lattices," *Science* **357**, 995–1001 (2017).
- [37] Michael O'Connor, Helen M Deeks, Edward Dawn, Oussama Metatla, Anne Roudaut, Matthew Sutton, Lisa May Thomas, Becca Rose Glowacki, Rebecca Sage, Philip Tew, *et al.*, "Sampling molecular conformations and dynamics in a multiuser virtual reality framework," *Science advances* **4**, eaat2731 (2018).
- [38] Daniel Probst and Jean-Louis Reymond, "Exploring drugbank in virtual reality chemical space," *Journal of chemical information and modeling* **58**, 1731–1735 (2018).
- [39] Jonas R Schmid, Moritz J Ernst, and Günther Thiele, "Structural chemistry 2.0: Combining augmented reality and 3d online models," (2020).
- [40] Davide Castelvecchi, "Using sound to explore events of the universe," *Nature* **597**, 144 (2021).
- [41] Alan M Turing, "Computing machinery and intelligence," *Mind* **50**, 433–460 (1950).
- [42] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, *et al.*, "The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities," *Artificial life* **26**, 274–306 (2020).
- [43] Chris J Pickard and RJ Needs, "Ab initio random structure searching," *Journal of Physics: Condensed Matter* **23**, 053201 (2011).
- [44] Mario Krenn, Mehul Malik, Robert Fickler, Radek Lapkiewicz, and Anton Zeilinger, "Automated search for new quantum experiments," *Physical review letters* **116**, 090405 (2016).
- [45] Mario Krenn, Manuel Erhard, and Anton Zeilinger, "Computer-inspired quantum experiments," *Nature Reviews Physics* **2**, 649–661 (2020).
- [46] Chris J Pickard and RJ Needs, "Highly compressed ammonia forms an ionic crystal," *Nature materials* **7**, 775–779 (2008).
- [47] Mario Krenn, Armin Hochrainer, Mayukh Lahiri, and Anton Zeilinger, "Entanglement by path identity," *Physical review letters* **118**, 080401 (2017).
- [48] Mario Krenn, Xuemei Gu, and Anton Zeilinger, "Quantum experiments and graphs: Multiparty states as coherent superpositions of perfect matchings," *Physical review letters* **119**, 240403 (2017).
- [49] Mario Krenn, Jakob Kottmann, Nora Tischler, and Alán Aspuru-Guzik, "Conceptual understanding through efficient automated design of quantum optical experiments," *Physical Review X* **11**, 031044 (2021).
- [50] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal, "Long short term memory networks for anomaly detection in time series," *Proceedings on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* **89**, 89–94 (2015).
- [51] ATLAS Collaboration, "Dijet resonance search with weak supervision using $s = 13$ tev p p collisions in the atlas detector," *Physical review letters* **125**, 131801 (2020).
- [52] CMS Collaboration, "Probing effective field theory operators in the associated production of top quarks with a z boson in multilepton final states at $s=13$ tev," *Journal of High Energy Physics* **2021**, 1–55 (2021).
- [53] Sang Eon Park, Dylan Rankin, Silviu-Marian Udrescu, Mikaeel Yunus, and Philip Harris, "Quasi anomalous knowledge: searching for new physics with embedded knowledge," *Journal of High Energy Physics* **2021**, 1–26 (2021).
- [54] Matthew D Schwartz, "Modern machine learning and particle physics," arXiv:2103.12226 (2021).
- [55] Gregor Kasieczka, Benjamin Nachman, David Shih, Oz Amram, Anders Andreassen, Kees Benkendorfer, Blaz Bortolato, Gustaaf Brooijmans, Florencio Canelli, Jack Collins, *et al.*, "The lhc olympics 2020: a community challenge for anomaly detection in high energy physics," *Reports on Progress in Physics* (2021).
- [56] Haizi Yu, Igor Mineyev, and Lav R Varshney, "A group-theoretic approach to computational abstraction: Symmetry-driven hierarchical clustering," arXiv:1807.11167 (2018).
- [57] Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu, "Automatic symmetry

- discovery with lie algebra convolutional network,” Advances in Neural Information Processing Systems **34** (2021).
- [58] Akshat Kumar Nigam, Robert Pollice, Matthew FD Hurley, Riley J Hickman, Matteo Aldeghi, Naruki Yoshikawa, Seyone Chithrananda, Vincent A Voelz, and Alán Aspuru-Guzik, “Assigning confidence to molecular property prediction,” Expert Opinion on Drug Discovery , 1–15 (2021).
- [59] A Davies, P Velickovic, L Buesing, S Blackwell, D Zheng, N Tomasev, R Tamburn, P Battaglia, C Blundell, A Juhasz, *et al.*, “Advancing mathematics by guiding human intuition with ai,” Nature (2021).
- [60] Michael R Douglas, “Machine learning as a tool in theoretical science,” Nature Reviews Physics , 1–2 (2022).
- [61] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, *et al.*, “The automation of science,” Science **324**, 85–89 (2009).
- [62] Anne-Catherine Bédard, Andrea Adamo, Kosi C Aroh, M Grace Russell, Aaron A Bedermann, Jeremy Torosian, Brian Yue, Klavs F Jensen, and Timothy F Jamison, “Reconfigurable system for automated optimization of diverse chemical reactions,” Science **361**, 1220–1225 (2018).
- [63] Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J Kitson, Davide Angelone, *et al.*, “Organic synthesis in a modular robotic system driven by a chemical programming language,” Science **363** (2019).
- [64] Connor W Coley, Dale A Thomas, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, *et al.*, “A robotic platform for flow synthesis of organic compounds informed by ai planning,” Science **365** (2019).
- [65] Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaoobo Li, Ben M Alston, Buyi Li, Rob Clowes, *et al.*, “A mobile robotic chemist,” Nature **583**, 237–241 (2020).
- [66] Sourav Chatterjee, Mara Guidi, Peter H Seeberger, and Kerry Gilmore, “Automated radial synthesis of organic molecules,” Nature **579**, 379–384 (2020).
- [67] Jonathan Grizou, Laurie J Points, Abhishek Sharma, and Leroy Cronin, “A curious formulation robot enables the discovery of a novel protocell behavior,” Science advances **6**, eaay4237 (2020).
- [68] Hyungil Moon, Dominic T Lennon, James Kirkpatrick, Nina M van Esbroeck, Leon C Camenzind, Liuqi Yu, Florian Vigneau, Dominik M Zumbühl, G Andrew D Briggs, Michael A Osborne, *et al.*, “Machine learning enables completely automatic tuning of a quantum device faster than human experts,” Nature communications **11**, 1–10 (2020).
- [69] Mogens Dalgaard, Felix Motzoi, Jens Jakob Sørensen, and Jacob Sherson, “Global optimization of quantum dynamics with alphazero deep exploration,” npj Quantum Information **6**, 1–9 (2020).
- [70] Peder Larsen and Markus Von Ins, “The rate of growth in scientific publication and the decline in coverage provided by science citation index,” Scientometrics **84**, 575–603 (2010).
- [71] James A Evans and Jacob G Foster, “Metaknowledge,” Science **331**, 721–725 (2011).
- [72] Aaron Clauset, Daniel B Larremore, and Roberta Sinatra, “Data-driven predictions in the science of science,” Science **355**, 477–480 (2017).
- [73] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, *et al.*, “Science of science,” Science **359** (2018).
- [74] Dashun Wang and Albert-László Barabási, *The science of science* (Cambridge University Press, 2021).
- [75] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain, “Unsupervised word embeddings capture latent knowledge from materials science literature,” Nature **571**, 95–98 (2019).
- [76] Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski, “Data-driven materials research enabled by natural language processing and information extraction,” Applied Physics Reviews **7**, 041317 (2020).
- [77] Daniel Schwalbe-Koda, Zach Jensen, Elsa Olivetti, and Rafael Gómez-Bombarelli, “Graph similarity drives zeolite diffusionless transformations and intergrowth,” Nature materials **18**, 1177–1181 (2019).
- [78] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans, “Choosing experiments to accelerate collective discovery,” Proceedings of the National Academy of Sciences **112**, 14569–14574 (2015).
- [79] Mario Krenn and Anton Zeilinger, “Predicting research trends with semantic and neural networks with an application in quantum physics,” Proceedings of the National Academy of Sciences **117**, 1910–1916 (2020).
- [80] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv:1810.04805 (2018).
- [81] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, *et al.*, “Language models are few-shot learners,” arXiv:2005.14165 (2020).
- [82] William L Hamilton, Rex Ying, and Jure Leskovec, “Inductive representation learning on large graphs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017) pp. 1025–1035.
- [83] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, “Methods for interpreting and understanding deep neural networks,” Digital Signal Processing **73**, 1–15 (2018).
- [84] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*, Vol. 11700 (Springer Nature, 2019).

- [85] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke, “Explainable machine learning for scientific insights and discoveries,” *IEEE Access* **8**, 42200–42216 (2020).
- [86] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence* **2**, 56–67 (2020).
- [87] Aravindh Mahendran and Andrea Vedaldi, “Understanding deep image representations by inverting them,” in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 5188–5196.
- [88] Alexander Mordvintsev, Christopher Olah, and Mike Tyka, “Inceptionism: Going deeper into neural networks,” <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html> (2015).
- [89] Cynthia Shen, Mario Krenn, Sagi Eppel, and Alan Aspuru-Guzik, “Deep molecular dreaming: Inverse machine learning for de-novo molecular design and interpretability with surjective representations,” *Machine Learning: Science and Technology* (2021).
- [90] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner, “Understanding disentangling in beta-vae,” arXiv:1804.03599 (2018).
- [91] Pascal Friederich, Mario Krenn, Isaac Tamblyn, and Alán Aspuru-Guzik, “Scientific intuition inspired by machine learning-generated hypotheses,” *Machine Learning: Science and Technology* **2**, 025027 (2021).
- [92] Daniel Flam-Shepherd, Tony Wu, Xuemei Gu, Alba Cervera-Lierta, Mario Krenn, and Alan Aspuru-Guzik, “Learning interpretable representations of entanglement in quantum optics experiments using deep generative models,” arXiv:2109.02490 (2021).
- [93] Geemi P Wellawatte, Aditi Seshadri, and Andrew D White, “Model agnostic generation of counterfactual explanations for molecules,” *Chemical Science* (2022).
- [94] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik, “Acquisition of chess knowledge in alphazero,” arXiv:2111.09259 (2021).
- [95] Michael Schmidt and Hod Lipson, “Distilling free-form natural laws from experimental data,” *Science* **324**, 81–85 (2009).
- [96] Antonio A Gentile, Brian Flynn, Sebastian Knauer, Nathan Wiebe, Stefano Paesani, Christopher E Granade, John G Rarity, Raffaele Santagati, and Anthony Laing, “Learning models of quantum systems from experiments,” *Nature Physics* **17**, 837–843 (2021).
- [97] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho, “Discovering symbolic models from deep learning with inductive biases,” arXiv:2006.11287 (2020).
- [98] Kyle Cranmer, Johann Brehmer, and Gilles Louppe, “The frontier of simulation-based inference,” *Proceedings of the National Academy of Sciences* **117**, 30055–30062 (2020).
- [99] Gal Raayoni, Shahar Gottlieb, Yahel Manor, George Pisha, Yoav Harris, Uri Mendlovic, Doron Haviv, Yaron Hadad, and Ido Kaminer, “Generating conjectures on fundamental constants with the ramanujan machine,” *Nature* **590**, 67–73 (2021).
- [100] Adam Zsolt Wagner, “Constructions in combinatorics via neural networks,” arXiv:2104.14516 (2021).
- [101] Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia, “Rediscovering orbital mechanics with machine learning,” arXiv:2202.02306 (2022).
- [102] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al., “Machine behaviour,” *Nature* **568**, 477–486 (2019).
- [103] Jürgen Schmidhuber, “Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes,” *Workshop on anticipatory behavior in adaptive learning systems*, 48–76 (2008).
- [104] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell, “Curiosity-driven exploration by self-supervised prediction,” *International Conference on Machine Learning*, 2778–2787 (2017).
- [105] Luca A Thiede, Mario Krenn, AkshatKumar Nigam, and Alan Aspuru-Guzik, “Curiosity in exploring chemical space: Intrinsic rewards for deep molecular reinforcement learning,” arXiv:2012.11293 (2020).
- [106] Lav R Varshney, Nazneen Fatema Rajani, and Richard Socher, “Explaining creative artifacts,” arXiv:2010.07126 (2020).
- [107] Lav R Varshney, Florian Pinel, Kush R Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Y-M Chee, “A big data approach to computational creativity: The curious case of chef watson,” *IBM Journal of Research and Development* **63**, 7–1 (2019).
- [108] Laurent Itti and Pierre Baldi, “Bayesian surprise attracts human attention,” *Vision research* **49**, 1295–1306 (2009).
- [109] Edward A Feigenbaum, “Some challenges and grand challenges for computational intelligence,” *Journal of the ACM (JACM)* **50**, 32–40 (2003).
- [110] Henk W De Regt and Victor Gijsbers, “How false theories can yield genuine understanding,” *Explaining understanding: New perspectives from epistemology and philosophy of science*, 50–75 (2017).