A Survey on Artificial Intelligence for Scientific Research: New Paradigm and Prospect

Project of AI3605 Lectures on Frontiers of Artificial Intelligence, 2023 Fall, SJTU

Xiangyuan Xue (521030910387)

School of Electronic Information and Electrical Engineering

Abstract

The integration of AI in scientific research promises to redefine the process of scientific discovery across fields. This survey seeks to deliver a panoramic view of this exciting intersection. Our discussion covers three paradigms of AI-empowered scientific research including representation learning, hypothesis generation and surrogate modeling. We also present some representative applications of AI for scientific research including AlphaTensor in mathematics, AlphaFold in biology and FourCastNet in meteorology. Finally, we discuss the challenges that must be addressed in the future development of AI for scientific research. We hope this survey can provide a comprehensive overview and inspire more researchers to explore the potential of AI for scientific research.

1 Introduction

Artificial Intelligence (AI) [18] has undergone a remarkable evolution since its inception. This field of study is based on the foundational belief that human intelligence can be so precisely described that a machine can be made to simulate it. In the early years, AI research was dominated by symbolic approaches, expert systems, and logic-based models. The focus pivoted on creating systems that could perform logical operations and solve problems by following predefined rules. However, these systems were limited by their reliance on hand-coded knowledge and lacked the ability to learn from data. With the advent of powerful computational hardware and the growing availability of large datasets, machine learning [8] has taken the center stage. This branch of AI focuses on the development of algorithms that can improve automatically through experience. The breakthrough moment came with the application of deep neural networks (DNNs) [17], which were inspired by the structure and function of the human brain. These neural networks are composed of layers of interconnected nodes, which can learn complex patterns in data.

In the 2010s, AI made significant leaps forward with advancements in deep learning architectures like convolutional neural networks (CNNs) [11] and recurrent neural networks (RNNs) [4], solving complex tasks such as image classification and speech recognition. In recent years, transformers [22] have emerged as a novel architecture for sequence modeling, achieving state-of-the-art results in multiple domains, especially in natural language processing (NLP). Transformers show strong capabilities in capturing long-range dependencies and modeling complex relationships, which makes learning from tremendous amounts of data feasible. Nowadays, generative pretrained transformers (GPTs) [14] have become the mainstream in large language models (LLMs), which brings about a new paradigm for AI research and has become one of the most useful productivity tools in daily life.

As technology rapidly evolves, AI is being applied more widely across various fields, including autonomous driving (AD), electronic design automation (EDA) [6], aritificial intelligence generated content (AIGC) [1], etc. Under the impetus of this trend, researchers in basic science also begins to explore the use of AI to assist in scientific research, publishing more and more impressive works.

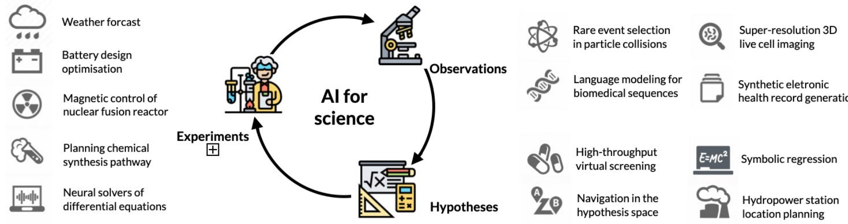

Figure 1: Scientific discovery is a multifaceted process including observations, hypotheses and experiments. AI can augment and accelerate research at any stage of this process. Some of the critical contributions that enhance scientific research are highlighted in this figure.

NeurIPS, one of the most prestigious conferences in AI, has a special track on AI for scientific research, which is especially interested in structural biology, particle physics, and drug discovery. Hundreds of papers are submitted to this track every year, and the number is rapidly growing. Many researchers believe that AI holds tremendous promise in having an impact on the way scientific discovery is performed today. In this survey, we will focus on the new paradigm of AI for scientific research, review the recent progress in this field and discuss the challenges and prospects. We aim to provide a comprehensive overview and inspire more future research in this area.

2 Paradigm

In the long river of human civilization, science discovery has undergone four major paradigms: empirical science, theoretical science, computational science and data-driven science. With the advent of the information age, the data-driven science paradigm has become the mainstream, while the emergence of AI significantly empowers this paradigm [23]. Method accumulation in machine learning and data mining provides all kinds of tools for data-driven science, while the rapid growth of computing power and data volume makes it possible to apply these tools to scientific research. In this section, we will introduce three main paradigms of AI for scientific research: representation learning, hypothesis generation and surrogate modeling, thus providing a glance at the potential of AI in scientific discovery.

2.1 Representation Learning

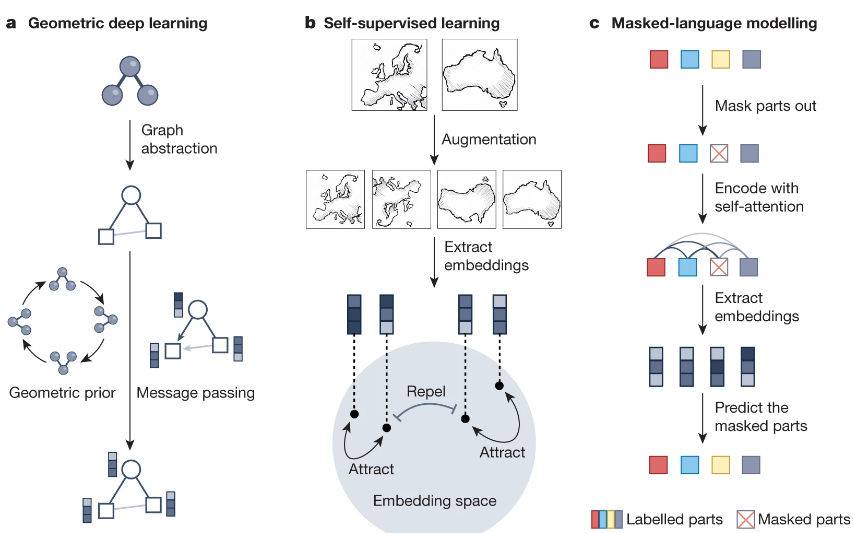

Figure 2: Learning meaningful representations of scientific data is a crucial step in scientific discovery. Representation learning can be achieved by mainly three methods: geometric deep learning, self-supervised learning and mask-language modeling, as is illustrated in the figure.

Learning representation is a cornerstone in the field of machine learning, serving as a fundamental process for understanding complex data. The representation of data entails the transformation of raw data into a form that AI systems can interpret and manipulate while retaining essential information. Diverse methods such as Word2Vec [13], VAE [10] and CLIP [16] exemplify the strides made in developing algorithms that can learn efficient and meaningful data representations. These techniques have proven instrumental in capturing the underlying structure of data in various forms, no matter it is text, images, or more abstract entities.

Figure 2 illustrates three methods for representation learning: geometric deep learning, self-supervised learning and mask-language modeling. Geometric deep learning is a general framework for learning on non-Euclidean domains, which can describe abstract data with complex structures. It is mainly implemented by graph neural networks (GNNs) [19]. Self-supervised learning [12] is a branch of unsupervised learning, which aims to learn representations from unlabeled data by solving pretext tasks. Mask-language modeling [2] is a pre-training technique in NLP, which can learn contextualized representations of words by masking some of them and predicting them from the remaining ones. It has been widely used in training LLMs.

In scientific research, the quality of data representation is paramount. High-quality representations are those that retain critical information necessary for scientific inquiry. They must embody simplicity, allowing for easy interpretation and manipulation by researchers. Moreover, these representations should be compact, reducing dimensionality without sacrificing relevant features, and should be adaptable to incorporate new findings and data seamlessly. It is also crucial that high-quality representations maintain the fidelity of the original data.

Representation learning can significantly advance scientific research by enabling the handling of complex and high-dimensional data. To be exact, it can distill essential patterns and relationships within the data, which may be non-trivial for human researchers to comprehend or discover. In fields such as bioinformatics, cheminformatics, and materials science, where the data can be exceedingly intricate, representation learning offers tools to extract meaningful features that can lead to the discovery of new drugs, materials, or decipher complex biological processes.

2.2 Hypothesis Generation

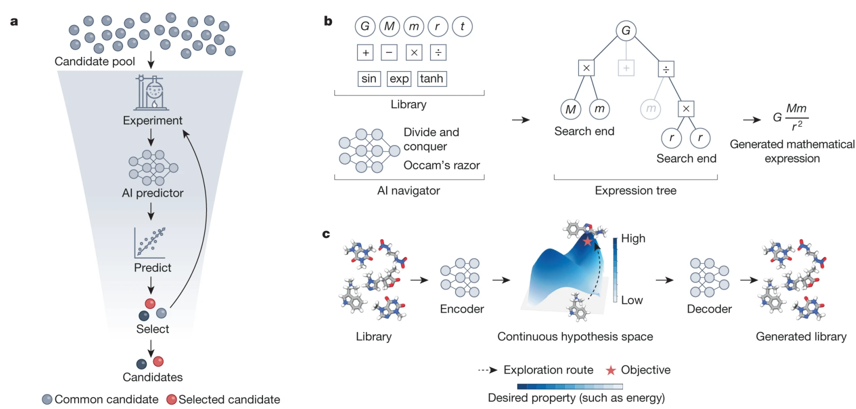

Figure 3: Formulation of scientific hypotheses can be time-consuming, while utilizing AI can significantly accelerate this process. Three possible methods for AI-guided hypothesis generation are shown in the figure: black-box prediction, law searching and hypothesis optimization.

One of the most crucial aspects of scientific inquiry is the generation of testable hypotheses. These hypotheses serve as the foundation upon which empirical studies are constructed and knowledge is accrued. Hypotheses take on various forms across the breadth of scientific disciplines, ranging from symbolic expressions in fields such as mathematics, to specific molecular structures in chemistry, and to the identification of genetic variants in biology. Each of these forms represents a prediction or an explanation that can be subjected to empirical testing.

Traditional formulation of meaningful scientific hypotheses is an endeavor that often involves significant human ingenuity and deep domain expertise. Scientists spend considerable amounts of time analyzing data, with the aim of uncovering patterns or inconsistencies that may suggest new lines of enquiry. The advent of advanced computational techniques, particularly those categorized as black-box predictors, has the potential to significantly accelerate the fitting process. These machine learning algorithms are capable of identifying patterns within large and complex datasets with greater speed and efficiency than humanly possible. They can rapidly analyze data points to derive correlations and causal relations, thereby suggesting potential hypotheses for further investigation. Generative AI represents a further evolution in hypothesis generation, in that it has the capability to autonomously develop new hypotheses without explicit human guidance. By leveraging powerful models such as LLMs, AI systems can propose novel conjectures that might not be immediately obvious to human researchers. With adequate computational resources, generative AI can produce a plurality of hypotheses that can subsequently be prioritized and verified through automated systems.

Figure 3 shows three possible methods for AI-guided hypothesis generation: black-box prediction, law searching and hypothesis optimization. Both discrete and continuous hypotheis spaces can be explored by these methods. By integrating AI into the scientific process, the horizon of discovery can be widened considerably. It enables the scientific community to pose a greater number of educated guesses about the workings of natural phenomena, thus opening up new opportunities for breakthroughs. Scientific fields that are data-intensive, such as genomics or materials science, stand to benefit immensely from these advancements in AI hypothesis generation, as the rate of experimental iteration can vastly exceed traditional methods.

2.3 Surrogate Modeling

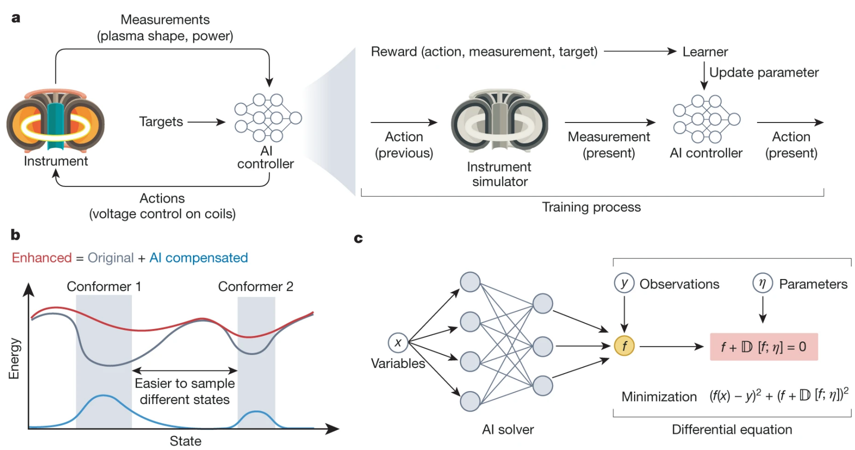

Figure 4: Computer simulations can be significantly augmented by surrogate modeling. The figure illustrates three possible methods for AI-driven experimentation and simulation: reinforcement learning for experimental pipeline control, supervised learning for error compensation and supervised learning and deep learning for solving partial differential equations.

The core of scientific enquiry is underpinned by the rigorous evaluation of hypotheses through experimentation. Traditional laboratory experiments can often present challenges in terms of resource allocation, financial expenditure and practical feasibility. In cases where traditional experiments are either too hazardous to carry out or require resources that are scarce or costly, computer simulations stand out as an alluring and promising alternative.

Computer simulations can reduce the need for physical resources, diminish risks, and significantly lower costs associated with scientific experimentation. However, one of the primary challenges that emerge with the use of simulations is the balance between computational speed and the accuracy of the results. Simulations that are highly detailed and accurate can be so computationally intensive that they become impracticable, whereas faster simulations might compromise on the fidelity of the results, potentially leading to inaccuracies in the hypothesis being tested.

Advancements in artificial intelligence, particularly deep learning models, have begun to offer solutions to this conundrum. Deep learning can aid in creating simulations that are both efficient and of high accuracy. By learning from vast datasets obtained from previous experiments and results, these models can predict outcomes with a significantly reduced computational overhead, therefore enabling a more expedient yet reliable analysis.

As is shown in Figure 4, one of the salient features of employing AI in scientific experimentation and simulation is the versatility of methods it brings to the table. Reinforcement Learning (RL) [21], for instance, can be adeptly applied to the control of experimental pipelines. Supervised learning excels at error compensation tasks. It can be trained on datasets where the true values are known and can learn to predict and adjust for errors in new data. Furthermore, AI-driven optimization techniques can solve complex equations that are often encountered in physics and engineering, providing solutions that would be infeasible using traditional numerical methods due to time constraints or computational complexity.

3 Applications

Researchers have been exploring the potential of AI for scientific research for decades, producing a wealth of literature on the subject. Previous works have completely reviewed the advances in AI for quantum, atomistic and continuum systems [24]. Recent advances in AI for biotechnology and drug discovery have also been reviewed in [5] and [7]. AI methods have been currently applied to a wide range of scientific fields, including but not limited to physics, chemistry, biology, astronomy and material science. In this section, we will introduce three representative applications of AI for scientific research: AlphaTensor [3] in mathematics, AlphaFold [9] in biology and FourCastNet [15] in meteorology. By introducing these applications, we hope to provide an insight about how AI methods overturn the traditional scientific research paradigm and what we can expect from the future of AI for scientific research.

3.1 AlphaTensor

Matrix multiplication is a critical operation with applications spanning across artificial intelligence, scientific computing and beyond. The search for more efficient matrix multiplication algorithms stands to revolutionize computational speeds on a grand scale. In this vein, the AlphaTensor framework emerges for the discovery of innovative matrix multiplication algorithms.

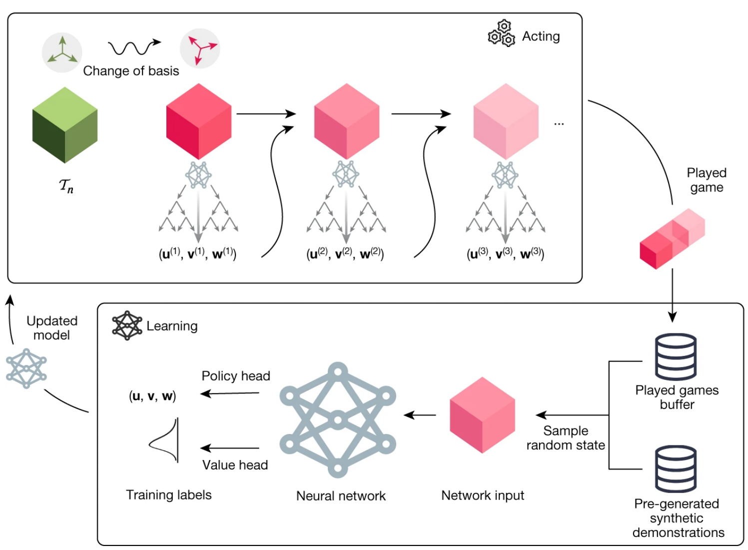

Figure 5: The overview of the AlphaTensor architecture. The model outputs decomposition policies with Monte Carlo tree search (MCTS) and learns from the reward signals. It iteratively evaluates the value function and improves the policy to discover better decomposition strategies.

AlphaTensor is a sophisticated algorithmic invention, drawing upon deep reinforcement learning principles similar to those of AlphaZero [20]. It functions by contextualizing the search for efficient matrix multiplication methods as a single-player game. The player navigates a grand expanse of potential tensor decompositions constrained within a finite factor space. The model is guided and enhanced iteratively via reinforcement learning mechanisms, essentially learning to identify superior multiplication strategies over time. AlphaTensor achieves outstanding performance, including the formulation of matrix multiplication algorithms that transcend the efficiency of those designed by human experts. This milestone underscores not just the potential but the tangible success of AlphaTensor as a transformative tool in algorithm design.

The success of AlphaTensor represents an evolutionary step in the continual quest for algorithmic efficiency. By optimizing fundamental operations such as matrix multiplication, AlphaTensor is setting the stage for significant accelerations in computational processes across numerous applications. Its ability to innovate and optimize simultaneously demonstrates the potential for machine learning to spearhead scientific discoveries and advancements.

3.2 AlphaFold

The structure of a protein determines its function. Hence, the prediction of protein structures is a critical aspect of molecular biology and has significant implications for various fields including drug discovery, understanding genetic diseases, and developing new biochemical processes. However, determining these structures experimentally using techniques like X-ray crystallography and cryo-electron microscopy is an incredibly time-consuming and expensive process.

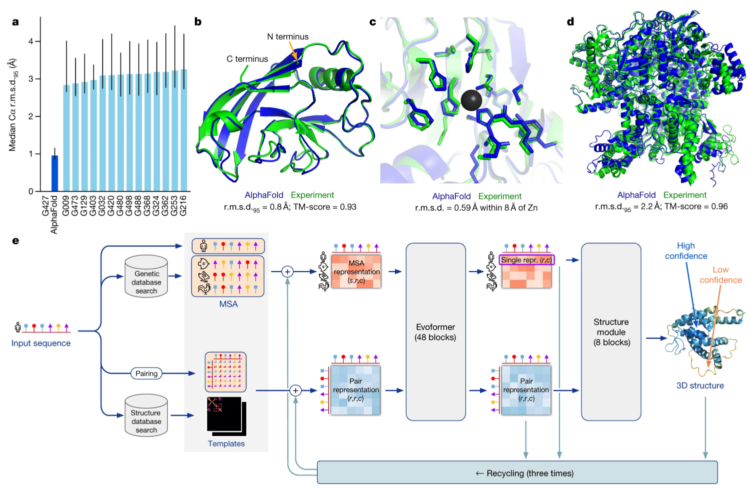

Figure 6: The overview of the AlphaFold architecture. The model retrieves the sequences of proteins from the database and predicts the spatial structures with a deep transformer network, which shows outstanding accuracy and extremely high efficiency in practice.

AlphaFold, a cutting-edge artificial intelligence program, is able to predict protein structures with a level of accuracy that is comparable to experimental methods. It comprises a neural network trained on a massive dataset of known protein structures. The system makes use of an attention-based mechanism that allows it to focus on different parts of the protein sequence as it predicts how the amino acids interact to fold the protein into its three-dimensional shape. This methodology enables AlphaFold to make predictions based on the context of protein sequences.

One of the primary advantages of AlphaFold is its speed and efficiency. It can predict the structure of a protein in several days, a process that could take traditional experimental methods several years to complete. This dramatic reduction in time allows for swift progression in understanding the molecular foundations of diseases and accelerates the drug discovery process by aiding in the identification of therapeutic targets. The accuracy of AlphaFold is another significant advantage. Its predictions have been found to be remarkably precise, which extends not only to familiar proteins but also to those that do not have known structures.

AlphaFold is a transformative tool in the field of molecular biology, which has completely overturned the traditional scientific research paradigm in this field. Many scientists believe that AlphaFold will win the Nobel Prize in the near future. The advent of AlphaFold indicates a new era in which artificial intelligence converges with scientific inquiry to yield significant breakthroughs that were once thought to be out of reach.

3.3 FourCastNet

The profound significance of accurate weather forecasting cannot be overstated. It has a direct impact on safety, economy, and the day-to-day operations of society. From mitigating disasters to planning agricultural activities, forecasting serves as a navigational system for human interaction with the natural world. However, traditional forecasting methods require large amounts of computational resources and time due to the complexity of the atmosphere.

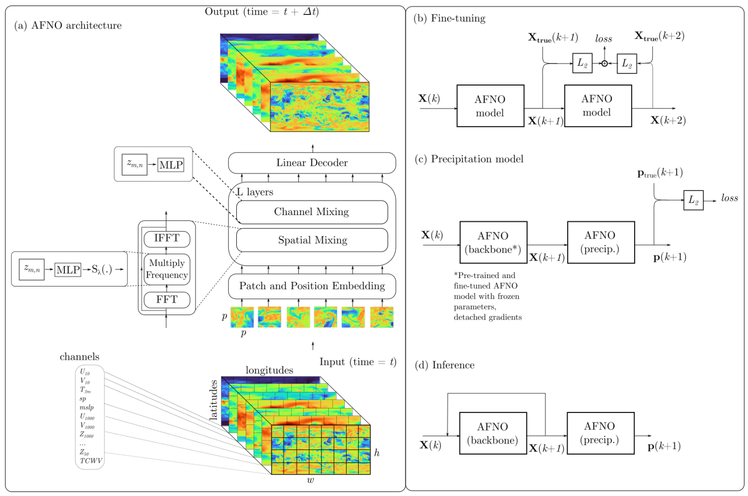

Figure 7: The overview of the FourCastNet architecture. The model utilizes the multi-layer transformer architecture to model the spatial and temporal dependencies of weather patterns. It is able to autorregressively predict the weather conditions in the next few steps.

FourCastNet stands out as a pioneering solution in the domain of AI-powered forecasting. Functioning as a Fourier ForeCasting Neural Network, FourCastNet revolutionizes the field with its remarkable capacity to deliver precise short to medium-range global predictions. Its methodology is grounded in the sophisticated interpretation of large-scale weather patterns, including surface wind speed, precipitation and atmospheric water vapor.

FourCastNet operates on the synthesis of data-driven deep learning models that are trained to simulate and anticipate weather dynamics. This AI model exhibits comparable accuracy to top-tier Numerical Weather Prediction (NWP) systems such as the ECMWF Integrated Forecasting System (IFS) when forecasting large-scale atmospheric conditions over short lead times. Moreover, FourCastNet surpasses its NWP counterparts in predicting small-scale variables like precipitation, which are typically more challenging to forecast due to their variability and finer granularity.

The superior performance of FourCastNet is not solely restricted to its predictive accuracy. It also boasts an unprecedented computational efficiency. FourCastNet can produce a comprehensive week-long forecast within several seconds, which enables meteorologists to execute rapid, cost-effective and extensive ensemble forecasts. The application of AI to weather forecasting is absolutely transformative, bringing huge benefits to economy and human society in the long term.

4 Challenges

Although the integration of AI in scientific research has the potential to revolutionize the way we approach scientific inquiry, there are still many challenges that must be addressed before this potential can be fully realized. In this section, we discuss three key challenges that must be overcome in the future development of AI for scientific research.

Adapting to novel paradigms. One significant challenge facing researchers in traditional fields is the necessity to adapt to the rapidly evolving landscape of artificial intelligence. The integration of AI in scientific research represents a paradigm shift, fundamentally altering methodologies and the interpretation of data. Traditional researchers must acquire new skills and knowledge to harness the power of AI effectively. This often requires a mindset change to include AI-driven techniques and principles. The challenge is amplified by the fact that the pace of AI development may outstrip the speed at which researchers can acclimate to these changes.

Assessing AI applicability. Another challenge lies in the potential misuse of AI tools, which can lead to significantly negative effects in research outcomes. The application of artificial intelligence is not universally beneficial or appropriate for every research scenario. Overreliance on AI without proper understanding can lead to misinterpretation of data and faulty conclusions. There is a pressing need to critically assess the applicability of AI methods in various scientific fields, to ensure that these powerful tools are used to enhance, rather than obscure, the accuracy and reliability of scientific inquiry. Researchers must develop a keen understanding of the strengths and limitations of AI in order to effectively integrate it into their work.

Cultivating interdisciplinary talents. Education systems face the challenge of keeping pace with advances in AI and its growing role in scientific research. There is a clear need for interdisciplinary talents who are well-versed in both domain-specific knowledge and AI. Educational institutions should foster environments where learning about AI is integrated with traditional scientific disciplines to prepare students for the demands of a workforce that increasingly relies on AI skills. This implies not only the introduction of AI components into existing courses but also the development of entirely new programs centered around the intersection of AI and scientific research.

5 Conclusion

In this survey, we presented a comprehensive overview of the integration of artificial intelligence in scientific research. We first described some new paradigms of AI-driven scientific research, including representation learning, hypothesis generation and surrogate modeling, which can significantly accelerate the scientific discovery process. Then we introduced several representative applications of AI for scientific research, including AlphaTensor in mathematics, AlphaFold in biology and FourCastNet in meteorology. These applications have completely overturned the traditional scientific research paradigm in their respective fields. Finally, we discussed some challenges that must be addressed in the future development of AI for scientific research, including adapting to novel paradigms, assessing AI applicability and cultivating interdisciplinary talents.

We are optimistic about the future of AI for scientific research. The integration of AI in scientific research has the potential to revolutionize the way we approach scientific inquiry. We believe that the development of AI for scientific research will continue to accelerate in the future, and we look forward to witnessing the emergence of more AI-driven scientific discoveries.

Acknowledgements. The structure of this survey is mainly referenced from [23]. The figures are cited from the original papers including [23, 3, 9, 15]. Generative LLMs are utilized including new Bing for literature retrieval and GPT-4 for article polishing, but it is necessary to emphasize that the main arguments of this survey are original instead of generated by LLMs.

References

[1] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Andreas Holzinger, Katharina Keiblinger, Petr Holub, Kurt Zatloukal, and Heimo Müller. Ai for life: Trends in artificial intelligence for biotechnology. *New Biotechnology*, 74:16–24, 2023.

[6] Guyue Huang, Jingbo Hu, Yifan He, Jialong Liu, Mingyuan Ma, Zhaoyang Shen, Juejian Wu, Yuanfan Xu, Hengrui Zhang, Kai Zhong, et al. Machine learning for electronic design automation: A survey. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 26(5):1–46, 2021.

[7] José Jiménez-Luna, Francesca Grisoni, Nils Weskamp, and Gisbert Schneider. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert opinion on drug discovery*, 16(9):949–959, 2021.

[8] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

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

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[12] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[15] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[18] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010.

[19] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[20] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[21] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

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

[24] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.