

Position Paper: Unlocking the Potential of AI Researchers in Scientific Discovery—What Is Missing?

Hengjie Yu

*School of Engineering
Westlake University*

Hangzhou, Zhejiang 310030, China
yuhengjie@westlake.edu.cn

Yuping Yan

*School of Engineering
Westlake University*

Hangzhou, Zhejiang 310030, China

Shuya Liu

*School of Engineering
Westlake University*

Hangzhou, Zhejiang 310030, China

Maozhen Qu

*Department of Chemistry
National University of Singapore*
Singapore 117543, Singapore

Haiyun Yang

*School of Engineering
Westlake University*

Hangzhou, Zhejiang 310030, China

Yaochu Jin

*School of Engineering
Westlake University*

Hangzhou, Zhejiang 310030, China
jinyaochu@westlake.edu.cn

Abstract—The potential of AI researchers in scientific discovery remains largely untapped. Over the past decade, AI for Science (AI4Science) publications in 145 Nature Index journals have increased fifteen-fold, yet they still account for less than 3% of the total publications. Drawing upon the Diffusion of Innovation theory, we project AI4Science’s share of total publications to rise from 2.72% in 2024 to approximately 20% by 2050. Achieving this shift requires fully harnessing the potential of AI researchers, as nearly 95% of AI-driven research in these journals is led by experimental scientists. To facilitate this, we propose structured workflows and strategic interventions to position AI researchers at the forefront of scientific discovery. Specifically, we identify three critical pathways: equipping experimental scientists with accessible AI tools to amplify the impact of AI researchers, bridging cognitive and methodological gaps to enable more direct involvement in scientific discovery, and proactively fostering a thriving AI-driven scientific ecosystem. By addressing these challenges, we aim to empower AI researchers as key drivers of future scientific breakthroughs.

Keywords—AI for Science, Scientific discovery, Diffusion of Innovation, Research paradigm, AI researcher

I. INTRODUCTION

Over the past decade, AI has become a powerful catalyst for scientific discovery [1], [2], [3]. Since 2015, its adoption and impact have expanded rapidly across scientific disciplines, driving unprecedented growth [4]. This surge has fueled the emergence of AI for Science (AI4Science), a field that leverages AI methods to tackle complex data challenges and uncover insights that were previously beyond reach. The impact of AI4Science is exemplified by the success of deep learning models in structural biology, chemistry, and biomedical research. For instance, AlphaFold [5] and ESMFold [6] have revolutionized protein structure prediction, dramatically improving the accuracy and efficiency of computational modeling compared to traditional physics-based methods. Similarly, in molecular representation learning, models like MolCLR [7] leverage contrastive learning to enhance the predictive power of molecular property predictions, facilitating drug discovery and materials science applications. Beyond structural and molecular biology, AI, especially large language models (LLMs), has also demonstrated remarkable utility in biomedical data analysis. GPTCelltype [8], an R package utilizing GPT-4, automates accurate cell type annotation from single-cell RNA sequencing data, greatly reducing the

effort and expertise needed for this task. PathCha [9], a vision-language AI assistant tailored for human pathology, excels in diagnostic accuracy and user preference by integrating a specialized vision encoder with a pretrained LLM. AI has been recognized as a transformative technology and powerful paradigm in various scientific fields [3], such as health and medicine [10], materials [11], biology [12], chemistry [13], industry [14], and the environment [15].

Transformative technologies have periodically reshaped scientific discovery. Computational modeling and numerical simulation, such as density functional theory [16], revolutionized research in physics, chemistry, and engineering by enabling large-scale predictive simulations. The 1990s brought high-throughput experimental techniques like next-generation sequencing [17] and mass spectrometry [18], shifting biological science toward data-driven exploration. The early 2000s saw the rise of statistical learning and data science [19], enhancing pattern recognition and predictive modeling across disciplines. At the same time, AI is reaching a point of saturation in its traditional core areas, making it increasingly difficult for new researchers to find opportunities. Expanding AI into scientific discovery is therefore not only essential for accelerating progress across disciplines, but also offers a critical new frontier for AI researchers. In this sense, science needs AI to unlock complex discoveries, and AI researchers need science as a rich and sustainable direction for future innovation. With AI playing an expanding role in scientific discovery, a careful assessment of its current contributions and remaining challenges is both timely and necessary. This article explores two fundamental questions: To what extent has AI advanced scientific discovery, and what potential remains for further expansion and innovation? Additionally, what gaps still exist in the engagement of AI researchers with scientific discovery, and how can these be bridged?

To answer these questions, this research examines AI-related research papers published in leading natural and health science journals over the past decade. By analyzing these publications and their author affiliations, we aim to understand the current progress of AI4Science within high-impact research and identify pathways for AI researchers to engage in this field. Furthermore, we offer practical guidance and a structured workflow to support AI researchers in embarking on scientific discovery. By highlighting key entry points and methodological approaches, we aim to lower the barriers for AI researchers seeking to contribute to scientific

advancements. Ultimately, we hope this survey serves as both an inspiration and a resource, facilitating broader integration of AI into scientific discovery while simultaneously expanding the frontiers of human knowledge, fostering mutual benefits for both AI researchers and the broader scientific community.

II. METHODS

A. Data collection

The 145 Nature Index journals, selected by an expert panel of active scientists for their reputation and impact, have long been central venues for groundbreaking research in the natural and health sciences. We retrieved AI-related research articles published between 2015 and 2024 in the 145 journals covered by the Nature Index from the Web of Science Core Collection. The search query combined publication type (“Article”), publication years (2015-2024), and a comprehensive list of 50 AI-related terms in the topic field (e.g., “machine learning,” “deep learning,” “transformer model,” “reinforcement learning,” “computer vision,” etc.), yielding 20,603 articles. Retrieved metadata included titles, abstracts, publication years, author affiliations, and journal names. The retrieved data included information such as titles, abstracts, publication years, author affiliations, and journal names. To obtain the total number of research articles published in these journals during the same period, we used the search query (SO=“Journal name” AND PY=2015-2024 AND DT=Article), which yielded a total of 1.16 million research articles.

B. Research type classification

Although the search terms targeted AI-related studies, not all retrieved articles were classified as AI for Science, necessitating a classification of research types. We defined four categories of research: *Typical AI Research*, *AI for Science*, *Science for AI*, and *Not Applicable*. *Typical AI Research* refers to studies focused on the development or analysis of AI methods, such as machine learning, neural networks, deep learning, algorithms, or hardware, without direct application to natural or health sciences; *AI for Science* encompasses studies that apply AI techniques to address scientific challenges in fields such as materials science, chemistry, physics, biology, and medicine, where AI is explicitly used to enhance discovery, prediction, data analysis, or scientific understanding; *Science for AI* denotes research that leverages findings from natural or cognitive sciences to inspire or improve AI models, architectures, or hardware, with the goal of advancing AI itself rather than solving scientific problems; *Not Applicable* covers studies outside these domains, including those unrelated to AI, those applying AI in non-scientific contexts, or those where AI is mentioned only peripherally.

Given the scale of the dataset, manual classification was impractical. Therefore, we employed a hybrid workflow that combines automated reasoning models with human verification (Fig. 1). Large and small reasoning models equipped with carefully designed prompts, a voting mechanism, and optimization of search terms, prompts, and model usage were integrated to ensure classification accuracy. The process begins with three state-of-the-art 20-32B reasoning models, Qwen3-32B [20], GPT-OSS-20B, and DeepSeek R1-32B [21], released in 2025. Articles (title and abstract) unanimously classified as *AI for Science* by all three models were directly assigned. In cases of disagreement, a larger 120B model (GPT-OSS-120B) was introduced. If the

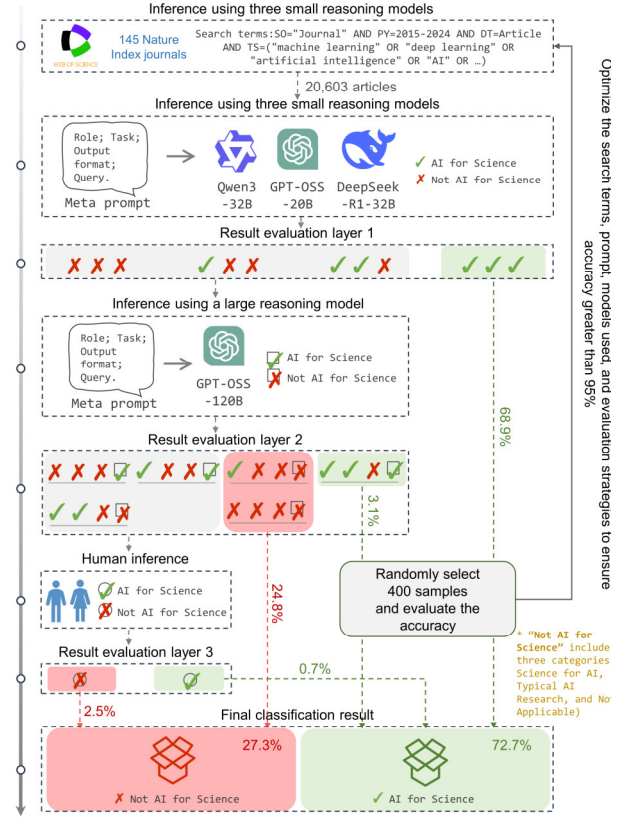


Fig. 1. Workflow for classifying research articles in collaboration with reasoning models and human oversight.

large model’s output aligned with the majority decision of the smaller models, the article was categorized accordingly. To address LLMs’ discrepancies in classification, human verification was applied. A random sample of 400 LLM classifications was assessed for accuracy, leading to four rounds of optimization, which resulted in a classification accuracy exceeding 95%.

C. Author affiliation classification

To classify author affiliations, we applied the similar hybrid workflow used in Research Type Classification, utilizing both large and small reasoning models. Affiliations were categorized into three groups: *AI Institute*, *Science Institute*, and *Not Applicable*. *AI Institute* includes institutions explicitly focused on AI, machine learning, or related fields. *Science Institute* refers to institutions dedicated to research in natural or health sciences, like biology, chemistry, or medicine, without reference to AI. *Not Applicable* includes non-research organizations, comprehensive research institutions, or institutions where AI or scientific research is not the primary focus.

D. Scientific entity extraction and curation

We used GPT-OSS-120B with carefully designed prompts to extract terms from article titles and abstracts. The extracted terms were processed through a multi-step curation pipeline: first, inconsistent delimiters were corrected and malformed or abnormally short entries removed; then, terms were split into individual entities, stripped of whitespace and punctuation, and standardized to first-letter capitalization. To reduce redundancy, we normalized terms by ignoring case, punctuation, and hyphens, and clustered those with shared prefixes. Finally, we filtered terms by length and word count,

grouped similar expressions, and ranked the clusters by total frequency for downstream analysis.

E. Scientific entity classification and word cloud plotting

Using the LLM-human collaboration workflow, each identified term was classified into one of three categories based on a structured prompt: *AI* (terms directly related to artificial intelligence technologies, such as machine learning, neural networks, and deep learning), *Science* (terms associated with natural or health sciences, such as physics, biology, or medicine), or *Not Applicable* (general or unrelated terms not specific to either domain). High-frequency terms underwent minor manual refinement, such as merging obvious synonyms and standardizing capitalization, to improve clarity and consistency for visualization. We then used the *wordcloud* package to generate a word cloud, clearly depicting the most frequently mentioned entities in the dataset.

III. RESULTS AND DISCUSSION

A. AI4Science: A Rapidly Growing, Yet Relatively Minor Component of Scientific Discovery

We compiled a dataset of 20,603 AI-related research articles published in these journals over the past decade. Although our search terms were designed to capture AI-related studies, not all retrieved articles qualified as “*AI for Science*.” Using a classification framework based on a hybrid system of large and small reasoning models, augmented with carefully designed prompts, and human verification, we ultimately categorized 72.7% of the articles as “*AI for Science*.”

Further analysis reveals that the number of AI-related articles exhibits a rapid increase over the last ten years, with both the absolute number and proportion rising across 145 Nature Index journals (Fig. 2A). The growth in AI-related research is especially evident in recent years, with the number of such articles in 2024 being fifteen times higher than in 2015, and the relative proportion of AI-related articles has also surged, reaching twelve times the level observed in 2015. This significant rise highlights the escalating adoption of AI methodologies across various scientific disciplines. Notably, 21 journals each published over 50 AI-related articles in 2024 alone, indicating a broad integration of AI. Furthermore, 16 journals had more than 5% of their total publications focused on AI, with five surpassing 9%. Among these, *Nature Methods* stands out, with AI-related articles comprised 15.87% of its total output in 2024, reflecting the pivotal role of AI in advancing specific fields. However, they still represent only 2.72% of total publications, indicating that AI is still in the early stages of widespread adoption in science.

The Diffusion of Innovation theory [22], developed by Everett Rogers, describes how new ideas and technologies spread through a population over time. It posits that adoption follows an S-curve, with innovations initially embraced by a small group of innovators and early adopters, followed by the early majority, late majority, and finally laggards. Based on this pattern, we anticipate that AI-related research will continue to grow rapidly over the next few decades, with its expansion gradually slowing as it reaches maturity (Fig. 2B). As the innovation matures, its adoption rate increases rapidly, then slows as it becomes widely accepted and integrated into everyday practices. AI’s application in scientific discovery is likely following a similar pattern. Early adopters in fields like computational biology, materials science, and data-driven

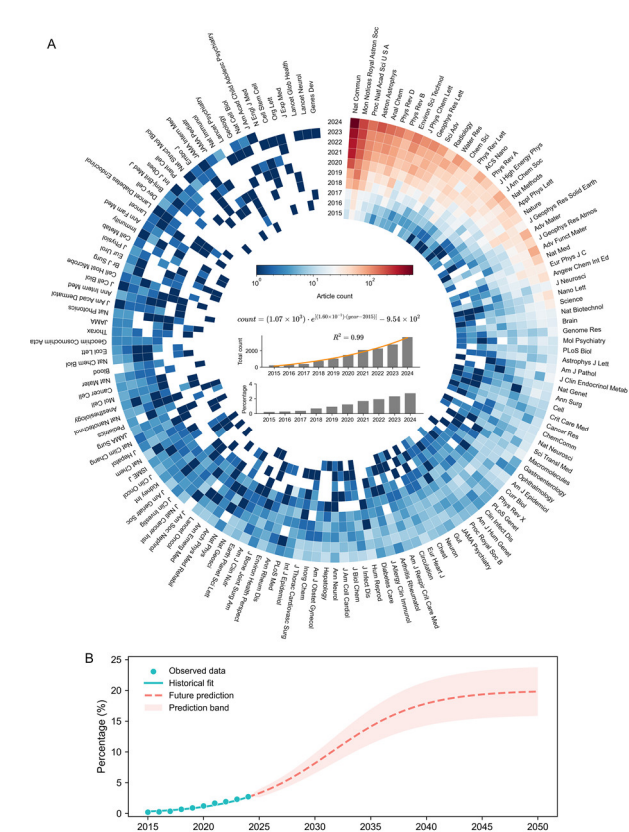


Fig.2. Publication and projected growth trends of AI-related research articles in 145 Nature Index journals. (A) Publication trends of 20,603 AI-related research articles in Nature Index journals (2015-2024). (B) Projected growth trend of AI-related research in Nature Index journals based on a logistic growth model. The shaded area represents the upper and lower bounds of this 20% projection.

research have led the charge, demonstrating the transformative potential of AI. As more researchers integrate AI into their work and the technology becomes more accessible, adoption will spread to the early majority and late majority, solidifying its place as a central tool in scientific research. By 2050, we project that AI-related research will stabilize at around 20% of the total Nature Index journal publications, reflecting AI’s fully integrated role in modern research.

B. AI Researchers Increasingly Contribute to Scientific Discovery in a Supporting Role

To elucidate the roles of AI researchers in the current AI for Science wave, we analyzed 61,759 author affiliations from 20,603 AI-related articles. These affiliations were categorized into three groups: *AI Institute*, *Science Institute*, or *Not Applicable* (for those not clearly affiliated with either AI or Science). The trends in author affiliations are illustrated in Fig. 3. AI institutions are increasingly involved in scientific discovery, yet they largely maintain a supportive role despite their growing significance. The average number of AI institutions per paper has risen from 0.16 in 2015 to 0.38 in 2024 (Fig. 3A). In contrast, the average number of institutions per research article and the total number of scientific institutions have remained relatively stable, experiencing a decline between 2018 and 2022, followed by a slight upward trend in recent years. The proportion of papers with AI institutions as author affiliations has grown from 13.33% to 25.30% (Fig. 3B). Additionally, AI institute-led research saw a notable rise in 2016, followed by a gradual

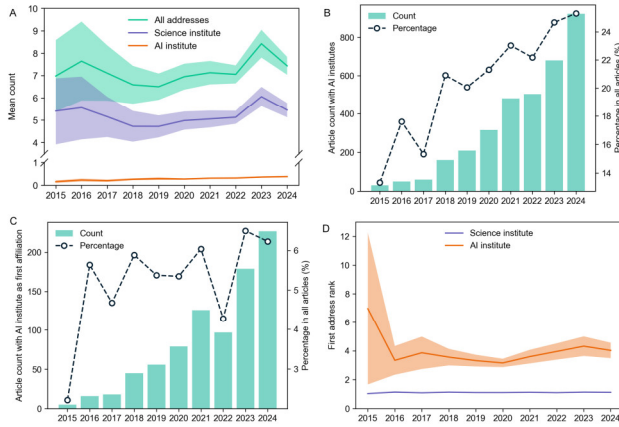


Fig. 3. Trends in author affiliations of 20,603 AI-related research articles (2015-2024). (A) Number of institutions per article by institution type. (B) Number and proportion of articles involving AI-related institutions. (C) Number and proportion of articles with AI-related institutions as the first affiliation. (D) Rank of the first-affiliated AI and Science institution. The shaded area represents the 95% confidence interval.

and uneven upward trend in the subsequent years (Fig. 3C), with the average percentage over the past three years at 5.66%. Similarly, the average ranking of the first-listed AI institution has improved from 6.97 in the first year to 3.68 over the past nine years. However, scientific institutions continue to dominate AI-driven scientific discovery, with the average ranking of the first-listed science institution remaining stable at around 1.1, shifting slightly from 1.09 in the first three years to 1.12 in the latest three years (Fig. 3D).

Given that many Nature Index journals may be less familiar to AI researchers, we also analyzed participation in five prominent interdisciplinary journals widely recognized in the AI community: Nature, Nature Communications, Science, Science Advances, and PNAS. These five journals account for 3,262 (21.7%) of all AI-for-Science articles. AI engagement here has consistently outpaced the broader Nature Index: in 2015, AI institutions appeared more frequently, as co-authors and first affiliations, than the overall average by roughly twofold. By 2024, 36.3% of articles included an AI affiliation, though only 10.7% listed one as the first. Notably, AI participation plateaued after 2021, suggesting a shift from rapid expansion to consolidation in these venues.

C. Unlocking The Potential of AI Researchers in Scientific Discovery: Three Key Directions

AI is rapidly accelerating scientific discovery. However, despite the widespread use of AI tools by experimental scientists, the role of AI researchers remains underexplored (Fig. 3). Achieving the projected growth, from 2.72% to 20%, in AI-driven scientific contributions requires stronger engagement from AI researchers. We identify three key directions, which are both ongoing and yet to be fully defined, that are essential to realizing this transformation, as illustrated in Fig. 4.

Developing user-friendly AI tools for scientific discovery offers major opportunities for AI researchers by increasing the real-world impact of their work. Providing experimental scientists with AI tools for data collection [23], modeling and analysis [24], and experiment design and conduction [25] can significantly accelerate scientific discovery. across a range of scientific fields. In fact, early adopters among experimental

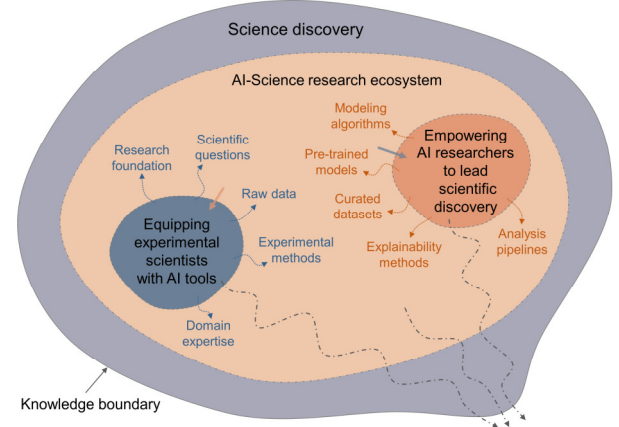


Fig. 4. Schematic representation of three key strategies for unlocking the potential of AI researchers in scientific discovery: (1) equipping experimental scientists with user-friendly AI tools, (2) enabling AI researchers to take a more direct and active role in scientific discovery, and (3) fostering a thriving AI-driven scientific ecosystem to sustain long-term innovation.

scientists have already begun utilizing AI tools in their research. As shown in Fig. 3B and Fig. 3C, 75% of AI-related articles are currently independently authored by experimental scientists. The next key step in advancing the application of AI is the development of more user-friendly tools and platforms, which will be discussed below. This is crucial for further driving the adoption of AI among experimental scientists.

The success of AI4Science depends not only on experimental scientists using AI tools but also on AI researchers actively participating in scientific discovery. By combining AI expertise with domain knowledge, these researchers can develop specialized methods to address scientific challenges. Although the number of AI researchers leading AI-driven discoveries has increased, this growth has slowed and remains limited (Fig. 3C). We attribute this limitation to two main gaps: a cognitive gap (regarding what scientific discoveries AI can be applied to) and a methodological gap (in terms of the workflow in which AI researchers lead scientific discoveries). Discussions on bridging these gaps will be presented in later sections.

Collaboration is central to AI4Science and extends beyond individual projects to the broader research ecosystem. A representative example is SHapley Additive exPlanations (SHAP) [26], [27], an interpretability method developed by AI researchers and widely adopted across many scientific fields. For instance, SHAP has been used to quantify the contributions of individual metabolites to disease risk, identifying key metabolites influencing the risk of 24 investigated diseases [28]. It has also been applied to interpret nanomaterial-plant-environment interactions [29], [30], providing insights into the factors affecting the root uptake of metal-oxide nanoparticles and their interactions within the soil. Additionally, SHAP was employed to analyze the contributions of raw ingredients and their interactions, offering valuable insights into the factors influencing the compressive strength of alkali-activated materials [31]. Its widespread use is largely enabled by an accessible open-source package, making the method easy for experimental scientists to apply. Importantly, experimental scientists often value model interpretability more than small performance gains, as explanations help generate new scientific

hypotheses. Building such collaborative ecosystems, supported by shared tools and standards, can accelerate AI4Science by improving transparency, reproducibility, and adoption.

D. Four Approaches to Equipping Experimental Scientists with AI

AI will not replace scientists because scientific discovery requires human intuition, creativity, and the ability to navigate uncertainty—capabilities that AI lacks [32], [33]. However, scientists who leverage AI can process vast amounts of data, automate complex analyses, and generate insights more efficiently, giving them a significant advantage in accelerating breakthroughs and expanding the frontiers of knowledge [34], [35]. Here, we discuss four key approaches, both established and emerging, that enable researchers to harness AI effectively, lowering barriers to adoption while accelerating scientific discovery (Fig. 5). These approaches not only illustrate how experimental scientists can harness AI tools but also present significant opportunities for AI and software researchers. Building user-friendly platforms of this kind, akin to existing statistical analysis software but enhanced with AI capabilities, can bridge the gap between advanced machine learning and practical scientific applications, enabling broader adoption and innovation.

User-friendly AI platforms can streamline data-driven research by automating model selection, training, and model explanation (Fig. 5A). Scientists only need to provide data, define prediction targets, and set resource constraints, after which the platform generates models, analyses, and interpretable reports. Model explanation and the resulting reports are critical components as they directly contribute to scientific understanding, which is one of the primary objectives of scientific research [36], [37]. Besides, many scientific fields lack sufficient experimental data, yet AI can still provide meaningful insights by integrating domain knowledge with LLMs (Fig. 5B). Although LLMs face challenges such as hallucinations and limited reasoning, retrieval-augmented generation can improve accuracy by incorporating external knowledge sources [38]. By building domain-specific knowledge bases, these platforms enable zero- or few-shot predictions, allowing scientists to explore

hypotheses even with limited data. As experimental scientists increasingly apply AI in both data-rich [39], [40] and data-scarce [41], [42] settings, user-friendly platforms will further promote adoption and enhance the role of AI in scientific discovery.

High-quality data is fundamental to machine learning-driven scientific discovery [43], yet much of this data remains embedded in unstructured scientific literature. Text, figures, and supplementary materials are not readily accessible to existing informatics systems, making data extraction a major bottleneck for AI-enabled science [44]. Consequently, converting published literature into usable, structured datasets is critical for advancing AI-assisted scientific discovery. Although LLMs have improved literature-based data extraction [45], their ability to handle complex textual content, figures, and supplementary files remains limited. With the emergence of more advanced LLMs, particularly those fine-tuned for data extraction tasks [46], and the rapid development of multimodal LLMs, substantial improvements in extraction accuracy and coverage are expected. In this workflow, researchers specify target articles and variables of interest, after which AI systems automatically retrieve, curate, and structure relevant data, enabling more efficient meta-analyses, comparative studies, and data-driven hypothesis generation (Fig. 5C).

Beyond data extraction, AI can also support scientific discovery by autonomously designing and executing experiments (Fig. 5D). Researchers define hypotheses, control variables, and desired outputs, while an AI-powered platform, enhanced by knowledge-augmented reasoning, develops an optimized experimental plan. Robotics then execute experiments, and a validation module assesses the reliability and quality of the data. Although AI- and robot-driven automated laboratories are still in their early stages, initial applications have already emerged [47], [48], demonstrating remarkable potential to transform scientific research [49]. This approach minimizes manual labor, increases reproducibility, and enables high-throughput hypothesis testing, ultimately accelerating scientific progress.

E. Scientific Domain "Ocean" with AI Application Potential

AI offers immense potential in scientific discovery, yet many opportunities remain underexplored by AI researchers. Currently, the application of AI in science is largely concentrated in fields with abundant structured datasets, such as foundational life sciences and medical research. Prominent examples include protein structure and function prediction, single-cell annotation, and drug discovery (Fig. 6A). The success of AI in these areas is largely driven by the availability of well-curated, large-scale datasets and clearly defined machine learning tasks. However, scientific discovery extends far beyond these domains, and numerous critical challenges in other fields remain largely untapped by AI researchers.

AI has strong potential to transform a wide range of scientific

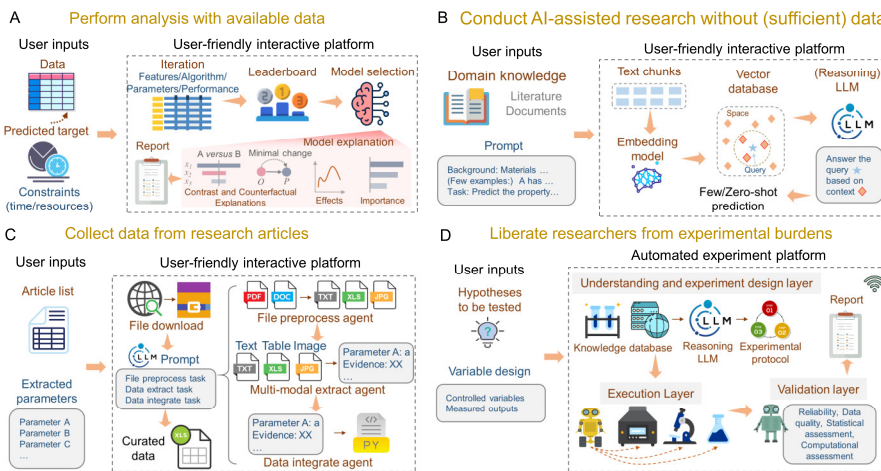


Fig. 5. Schematic diagram of four approaches to equipping experimental scientists with AI. (A) User-friendly platforms facilitate modeling and explanation when researchers have access to sufficient data. (B) Domain-enhanced LLMs enable zero-shot and few-shot inference in data-scarce scenarios. (C) The integration of LLMs, multimodal systems, and multi-task frameworks makes automated data extraction from literature increasingly feasible. (D) Advances in LLMs and robotics are driving the development of autonomous experimental platforms, freeing researchers from labor-intensive experiments.

LLMs and robotics are expected to further lower the barrier to experimental validation [49], [59], making it increasingly accessible.

IV. CONCLUSION

AI4Science has grown rapidly over the past decade, yet its full potential, particularly the role of AI researchers, remains underutilized. Although AI is increasingly integrated into scientific research, its application is still largely led by experimental scientists, with AI researchers often serving in supporting roles. This limits the depth of AI's impact on discovery. To accelerate progress, it is essential to position AI researchers as active contributors rather than mere tool developers. By fostering deeper collaboration and bridging cognitive and methodological gaps, they can drive more transformative advances in science. Looking ahead, the future of AI4Science depends on a well-defined human-AI collaboration paradigm that leverages AI's analytical strengths while keeping human researchers in control of scientific reasoning and validation. Addressing challenges such as hallucinations and defining clear boundaries for AI use will be critical. Moreover, experimental validation remains the ultimate safeguard of AI-driven discoveries, ensuring scientific rigor across disciplines. With these shifts, AI4Science could expand from its current 2.72% share of Nature Index publications to 20% by 2050, unlocking new frontiers in scientific discovery.

ACKNOWLEDGMENT

This work is supported by Westlake Education Foundation (Grant No. 103110846022301) and China Postdoctoral Science Foundation (Grant No. 2024M762941). The authors acknowledge the use of ChatGPT solely for language polishing and assume full accountability for the manuscript's accuracy and integrity.

REFERENCES

- [1] Y. Gil, M. Greaves, J. Hendler, and H. Hirsh, "Amplify scientific discovery with artificial intelligence," *Science*, vol. 346, no. 6206, pp. 171–172, Oct. 2014, doi: 10.1126/science.1259439.
- [2] H. Wang *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, Aug. 2023, doi: 10.1038/s41586-023-06221-2.
- [3] Y. Xu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation*, vol. 2, no. 4, p. 100179, Nov. 2021, doi: 10.1016/j.xinn.2021.100179.
- [4] J. Gao and D. Wang, "Quantifying the use and potential benefits of artificial intelligence in scientific research," *Nat Hum Behav*, vol. 8, no. 12, pp. 2281–2292, Oct. 2024, doi: 10.1038/s41562-024-02020-5.
- [5] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
- [6] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023, doi: 10.1126/science.ade2574.
- [7] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, "Molecular contrastive learning of representations via graph neural networks," *Nat Mach Intell*, vol. 4, no. 3, pp. 279–287, Mar. 2022, doi: 10.1038/s42256-022-00447-x.
- [8] W. Hou and Z. Ji, "Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis," *Nat Methods*, vol. 21, no. 8, pp. 1462–1465, Aug. 2024, doi: 10.1038/s41592-024-02235-4.
- [9] M. Y. Lu *et al.*, "A multimodal generative AI copilot for human pathology," *Nature*, vol. 634, no. 8033, pp. 466–473, Oct. 2024, doi: 10.1038/s41586-024-07618-3.
- [10] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat Med*, vol. 28, no. 1, pp. 31–38, Jan. 2022, doi: 10.1038/s41591-021-01614-0.
- [11] K. Choudhary *et al.*, "Recent advances and applications of deep learning methods in materials science," *npj Comput Mater*, vol. 8, no. 1, p. 59, Apr. 2022, doi: 10.1038/s41524-022-00734-6.
- [12] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, no. 7, p. 878, Jul. 2016, doi: 10.15252/msb.20156651.
- [13] Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins, and Q. Zhou, "Artificial Intelligence in Chemistry: Current Trends and Future Directions," *J. Chem. Inf. Model.*, vol. 61, no. 7, pp. 3197–3212, Jul. 2021, doi: 10.1021/acs.jcim.1c00619.
- [14] Z.-L. Zhao, H.-J. Yu, and F. Cheng, "An Analysis of Factors Affecting Agricultural Tractors' Reliability Using Random Survival Forests Based on Warranty Data," *IEEE Access*, vol. 10, pp. 50183–50194, 2022, doi: 10.1109/ACCESS.2022.3172348.
- [15] H. Yu, S. Tang, E. M. Hamed, S. F. Y. Li, Y. Jin, and F. Cheng, "Optimizing the benefit–risk trade-off in nano-agrochemicals through explainable machine learning: beyond concentration," *Environ. Sci.: Nano*, vol. 11, no. 8, pp. 3374–3389, 2024, doi: 10.1039/D4EN00213J.
- [16] B. Huang, G. F. Von Rudorff, and O. A. Von Lilienfeld, "The central role of density functional theory in the AI age," *Science*, vol. 381, no. 6654, pp. 170–175, Jul. 2023, doi: 10.1126/science.abn3445.
- [17] H. Satam *et al.*, "Next-Generation Sequencing Technology: Current Trends and Advancements," *Biology*, vol. 12, no. 7, p. 997, Jul. 2023, doi: 10.3390/biology12070997.
- [18] S. Heuckeroth *et al.*, "Reproducible mass spectrometry data processing and compound annotation in MZmine 3," *Nat Protoc*, vol. 19, no. 9, pp. 2597–2641, Sep. 2024, doi: 10.1038/s41596-024-00996-y.
- [19] K. Edfeldt *et al.*, "A data science roadmap for open science organizations engaged in early-stage drug discovery," *Nat Commun*, vol. 15, no. 1, p. 5640, Jul. 2024, doi: 10.1038/s41467-024-49777-x.
- [20] A. Yang *et al.*, "Qwen3 Technical Report," 2025, *arXiv*. doi: 10.48550/ARXIV.2505.09388.
- [21] D. Guo *et al.*, "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, Sep. 2025, doi: 10.1038/s41586-025-09422-z.
- [22] E. M. Rogers, A. Singhal, and M. M. Quinlan, "Diffusion of innovations," in *An integrated approach to communication theory and research*, Routledge, 2014, pp. 432–448.
- [23] M. P. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," *Nat Commun*, vol. 15, no. 1, p. 1569, Feb. 2024, doi: 10.1038/s41467-024-45914-8.
- [24] M. Krenn *et al.*, "On scientific understanding with artificial intelligence," *Nat Rev Phys*, vol. 4, no. 12, pp. 761–769, Oct. 2022, doi: 10.1038/s42254-022-00518-3.
- [25] J. Gottweis *et al.*, "Towards an AI co-scientist," 2025, *arXiv*. doi: 10.48550/ARXIV.2502.18864.
- [26] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, *arXiv*. doi: 10.48550/ARXIV.1705.07874.
- [27] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

- [28] T. Buerger *et al.*, “Metabolomic profiles predict individual multidisease outcomes,” *Nat Med*, vol. 28, no. 11, pp. 2309–2320, Nov. 2022, doi: 10.1038/s41591-022-01980-3.
- [29] H. Yu, Z. Zhao, D. Luo, and F. Cheng, “Interpretable machine learning for investigating complex nanomaterial–plant–soil interactions,” *Environ. Sci.: Nano*, vol. 9, no. 11, pp. 4305–4316, 2022, doi: 10.1039/D2EN00181K.
- [30] H. Yu, S. Tang, S. F. Y. Li, and F. Cheng, “Averaging strategy for interpretable machine learning on small datasets to understand nanoparticle uptake after seed nano-treatment,” *Environmental Science & Technology*, vol. 57, no. 34, pp. 12760–12770, 2023, doi: 10.1021/acs.est.3c01878.
- [31] X. Zheng *et al.*, “A data-driven approach to predict the compressive strength of alkali-activated materials and correlation of influencing parameters using SHapley Additive exPlanations (SHAP) analysis,” *Journal of Materials Research and Technology*, vol. 25, pp. 4074–4093, Jul. 2023, doi: 10.1016/j.jmrt.2023.06.207.
- [32] V. A. Makarov, T. Stouch, B. Allgood, C. D. Willis, and N. Lynch, “Best practices for artificial intelligence in life sciences research,” *Drug Discovery Today*, vol. 26, no. 5, pp. 1107–1110, May 2021, doi: 10.1016/j.drudis.2021.01.017.
- [33] L. Messeri and M. J. Crockett, “Artificial intelligence and illusions of understanding in scientific research,” *Nature*, vol. 627, no. 8002, pp. 49–58, Mar. 2024, doi: 10.1038/s41586-024-07146-0.
- [34] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, “Interpretable and Explainable Machine Learning for Materials Science and Chemistry,” *Acc. Mater. Res.*, vol. 3, no. 6, pp. 597–607, Jun. 2022, doi: 10.1021/accountsnmr.1c00244.
- [35] S. Kapoor *et al.*, “REFORMS: Consensus-based Recommendations for Machine-learning-based Science,” *Sci. Adv.*, vol. 10, no. 18, p. eadk3452, May 2024, doi: 10.1126/sciadv.adk3452.
- [36] M. Krenn *et al.*, “On scientific understanding with artificial intelligence,” *Nat Rev Phys*, vol. 4, no. 12, pp. 761–769, Oct. 2022, doi: 10.1038/s42254-022-00518-3.
- [37] H. Yu *et al.*, “Empowering scientific discovery with explainable small domain-specific and large language models,” *Artif Intell Rev*, vol. 58, no. 12, p. 371, Oct. 2025, doi: 10.1007/s10462-025-11365-w.
- [38] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [39] E. O. Pyzer-Knapp *et al.*, “Accelerating materials discovery using artificial intelligence, high performance computing and robotics,” *npj Comput Mater*, vol. 8, no. 1, p. 84, Apr. 2022, doi: 10.1038/s41524-022-00765-z.
- [40] F. Wong *et al.*, “Discovery of a structural class of antibiotics with explainable deep learning,” *Nature*, vol. 626, no. 7997, pp. 177–185, Feb. 2024, doi: 10.1038/s41586-023-06887-8.
- [41] Z. Chen *et al.*, “ProBID-Net: a deep learning model for protein–protein binding interface design,” *Chem. Sci.*, vol. 15, no. 47, pp. 19977–19990, 2024, doi: 10.1039/D4SC02233E.
- [42] S. Liu, T. Wen, A. S. L. S. Pattamatta, and D. J. Srolovitz, “A prompt-engineered large language model, deep learning workflow for materials classification,” *Materials Today*, vol. 80, pp. 240–249, Nov. 2024, doi: 10.1016/j.mattod.2024.08.028.
- [43] Y. Liu *et al.*, “Data quantity governance for machine learning in materials science,” *National Science Review*, vol. 10, no. 7, p. nwad125, May 2023, doi: 10.1093/nsr/nwad125.
- [44] S. Gupta, A. Mahmood, P. Shetty, A. Adeboye, and R. Ramprasad, “Data extraction from polymer literature using large language models,” *Commun Mater*, vol. 5, no. 1, p. 269, Dec. 2024, doi: 10.1038/s43246-024-00708-9.
- [45] M. P. Polak *et al.*, “Flexible, model-agnostic method for materials data extraction from text using general purpose language models,” *Digital Discovery*, vol. 3, no. 6, pp. 1221–1235, 2024, doi: 10.1039/D4DD00016A.
- [46] Q. Ai, F. Meng, J. Shi, B. Pelkie, and C. W. Coley, “Extracting structured data from organic synthesis procedures using a fine-tuned large language model,” *Digital Discovery*, vol. 3, no. 9, pp. 1822–1831, 2024, doi: 10.1039/D4DD00091A.
- [47] N. J. Szymanski *et al.*, “An autonomous laboratory for the accelerated synthesis of novel materials,” *Nature*, vol. 624, no. 7990, pp. 86–91, Dec. 2023, doi: 10.1038/s41586-023-06734-w.
- [48] N. H. Angello *et al.*, “Closed-loop transfer enables artificial intelligence to yield chemical knowledge,” *Nature*, vol. 633, no. 8029, pp. 351–358, Sep. 2024, doi: 10.1038/s41586-024-07892-1.
- [49] A. Angelopoulos, J. F. Cahoon, and R. Alterovitz, “Transforming science labs into automated factories of discovery,” *Science Robotics*, vol. 9, no. 95, p. eadm6991, doi: 10.1126/scirobotics.adm6991.
- [50] M. Mann, C. Kumar, W.-F. Zeng, and M. T. Strauss, “Artificial intelligence for proteomics and biomarker discovery,” *Cell Systems*, vol. 12, no. 8, pp. 759–770, Aug. 2021, doi: 10.1016/j.cels.2021.06.006.
- [51] S. Gao *et al.*, “Empowering biomedical discovery with AI agents,” *Cell*, vol. 187, no. 22, pp. 6125–6151, Oct. 2024, doi: 10.1016/j.cell.2024.09.022.
- [52] R. Najjar, “Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging,” *Diagnostics*, vol. 13, no. 17, p. 2760, Aug. 2023, doi: 10.3390/diagnostics13172760.
- [53] S. Schmidgall *et al.*, “Agent Laboratory: Using LLM Agents as Research Assistants,” 2025, *arXiv*. doi: 10.48550/ARXIV.2501.04227.
- [54] X. Hu *et al.*, “Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas,” 2024, *arXiv*. doi: 10.48550/ARXIV.2410.14255.
- [55] Z. Yang *et al.*, “MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses,” 2024, *arXiv*. doi: 10.48550/ARXIV.2410.07076.
- [56] A. Novikov *et al.*, “AlphaEvolve: A coding agent for scientific and algorithmic discovery,” 2025, *arXiv*. doi: 10.48550/ARXIV.2506.13131.
- [57] K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, and J. Zou, “The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies,” *Nature*, Jul. 2025, doi: 10.1038/s41586-025-09442-9.
- [58] H. Yu, K. A. Dawson, H. Yang, S. Liu, Y. Yan, and Y. Jin, “A million-scale dataset and generalizable foundation model for nanomaterial-protein interactions,” 2025, *arXiv*. doi: 10.48550/ARXIV.2507.14245.
- [59] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “Augmenting large language models with chemistry tools,” *Nat Mach Intell*, vol. 6, no. 5, pp. 525–535, May 2024, doi: 10.1038/s42256-024-00832-8.