Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges

Chandan K Reddy, Parshin Shojaee
Virginia Tech
reddy@cs.vt.edu, parshinshojaee@vt.edu

Abstract

Scientific discovery is a complex cognitive process that has driven human knowledge and technological progress for centuries. While artificial intelligence (AI) has made significant advances in automating aspects of scientific reasoning, simulation, and experimentation, we still lack integrated AI systems capable of performing autonomous long-term scientific research and discovery. This paper examines the current state of AI for scientific discovery, highlighting recent progress in large language models and other AI techniques applied to scientific tasks. We then outline key challenges and promising research directions toward developing more comprehensive AI systems for scientific discovery, including the need for science-focused AI agents, improved benchmarks and evaluation metrics, multimodal scientific representations, and unified frameworks combining reasoning, theorem proving, and data-driven modeling. Addressing these challenges could lead to transformative AI tools to accelerate progress across disciplines towards scientific discovery.

Introduction

Scientific discovery - the process of formulating and validating new concepts, laws, and theories to explain natural phenomena - is one of humanity's most intellectually demanding and impactful pursuits. For decades, AI researchers have sought to automate aspects of scientific reasoning and discovery. Early work focused on symbolic AI approaches to replicate the formation of scientific hypotheses and laws in symbolic forms (Segler, Preuss, and Waller 2018; MacColl 1897). More recently, deep learning and large language models (LLMs) have shown promise in tasks like literature analysis and brainstorming (Ji et al. 2024; Lu et al. 2024; Si, Yang, and Hashimoto 2024), experiment design (Boiko et al. 2023; Arlt et al. 2024), hypothesis generation (Wang et al. 2024; Ji et al. 2024), and equation discovery (Shojaee et al. 2024b; Ma et al. 2024).

Despite this progress, we still lack AI systems capable of integrating the diverse cognitive processes involved in sustained scientific research and discovery. Most work has focused on narrow aspects of scientific reasoning in isolation. Developing more comprehensive AI discovery systems capable of supporting the full cycle of scientific inquiry —from context retrieval and hypothesis generation to experiment design and evaluation (Figure 1) —could dramatically accelerate progress across scientific disciplines. This paper examines the current state and future potential of generative AI for scientific discovery. We highlight recent advances, particularly in scientific understanding and discovery frameworks, while identifying critical gaps. We then outline key research challenges and directions towards more unified AI systems for discovery, including: (i) Creating improved benchmarks and evaluation frameworks for scientific discovery; (ii) Developing science-focused AI agents that leverage scientific knowledge and reasoning capabilities; (iii) Advancing multimodal scientific representations beyond text; and (iv) Unifying automated reasoning, theorem proving, and data-driven modeling. By tackling these challenges, the AI and Science community can work towards systems that serve as collaborative partners to human scientists, accelerating the pace of discovery in science.

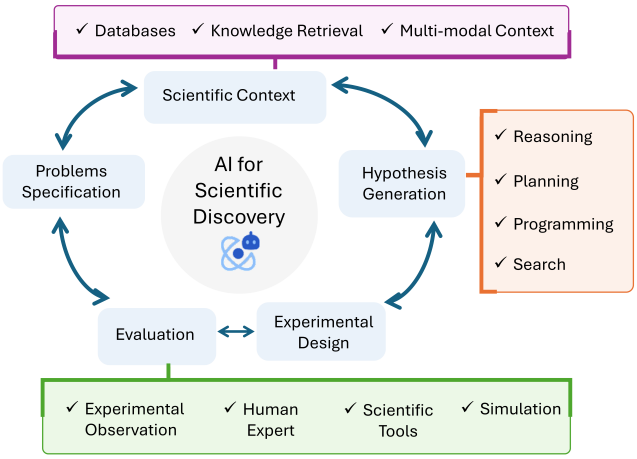

Figure 1: **Overview of the AI-driven scientific discovery framework.** The cycle illustrates the iterative process of scientific inquiry. The framework begins with user-defined problem specifications, retrieves relevant scientific context from literature and databases, and utilizes generative AI systems to produce new hypotheses and experimental designs. These AI-generated concepts are then evaluated and refined through experimental observation, expert input, and scientific tools, driving further iterations of the discovery cycle.

Recent Advances in AI for Scientific Tasks

The past decade has witnessed remarkable progress in applying AI to various scientific tasks. This section highlights some of the most significant recent advances, demonstrating AI's growing capabilities in supporting and accelerating scientific discovery across multiple disciplines.

Literature Analysis and Brainstorming

The exponential growth of scientific publications has made it increasingly challenging for researchers to stay abreast of developments in their fields. Large language models (LLMs) pre-trained on vast scientific corpora have emerged as powerful tools to address this challenge, enhancing literature analysis and interaction. Researchers have developed specialized LLMs for various scientific domains. Models like PubMedBERT (Gu et al. 2021) and BioBERT (Lee et al. 2020) focus on biomedical literature, while SciBERT (Beltagy, Lo, and Cohan 2019) covers a broader range of scientific disciplines. More recent models such as BioGPT (Luo et al. 2022) and SciGLM (Zhang et al. 2024) have further pushed the boundaries of scientific language modeling, incorporating advanced architectures and training techniques. These models, trained on sources like PubMed and arXiv, excel at literature information retrieval, summarization, and question-answering. They enable efficient navigation of scientific knowledge by quickly finding relevant papers, distilling key findings, and synthesizing information to answer complex queries.

Beyond analysis, recent works demonstrate LLMs' potential in generating novel scientific insights. For instance, SciMON (Ji et al. 2024) uses LLMs to generate new scientific ideas by analyzing patterns in the existing literature. These advancements show AI's capacity to not only aid in literature review but also contribute to identifying promising and novel research directions, potentially accelerating scientific discovery.

Theorem Proving

Automated theorem proving has recently gained attention in AI for science research due to its fundamental role in scientific reasoning. Recent years have seen remarkable progress in this field, particularly through the integration of LLMs with formal reasoning systems. The GPT-f framework (Polu and Sutskever 2020) pioneered this approach by training transformer-based language models on proof tactics, enabling navigation through complex mathematical proofs with the help of learned priors. Building on this, researchers have integrated proving techniques with LLMs and developed enhancements such as data augmentation (Han et al. 2021), retrieval augmentation (Yang et al. 2024), and novel proof search methods (Lample et al. 2022; Wang et al. 2023b). One of the key enhancements is the autoformalization approach, exemplified by the Draft-Sketch-Prove method (Jiang et al. 2023). This method uses LLMs to first draft informal proofs, translate them into formal sketches, and then complete proofs with additional proof assistant tools (Böhme and Nipkow 2010), mimicking the human process of moving from intuitive understanding to rigorous proof. As these systems become more adept at formalizing and proving complex statements, they could be applied to derive scientific theories, potentially accelerating the scientific process and leading to enhancements in fields where theoretical understanding lags behind empirical methods.

Experimental Design

Experimental design is a critical component of the scientific process, often requiring extensive domain knowledge and creative thinking. The automation of this process through generative models has the potential to accelerate scientific discovery across various fields. By leveraging LLM agents, researchers are recently developing systems that can design, plan, optimize, and even execute scientific experiments with minimal human intervention. These tools are particularly valuable in fields where experimental setup is costly, allowing researchers to explore a wider range of possibilities before physical implementation. For example, in physics, LLM-driven systems have demonstrated effectiveness in designing complex quantum experiments (Arlt et al. 2024) and optimizing parameters in high-energy physics simulations (Cai et al. 2024; Baldi, Sadowski, and Whiteson 2014). Chemistry has also recently seen advancements in automated experimentation, with LLM agent systems capable of designing and optimizing chemical reactions (M. Bran et al. 2024). Moreover, in biology and medicine, LLM-driven experimental design has shown promise in optimizing gene-editing protocols (Huang et al. 2024), and designing more effective clinical trials (Singhal et al. 2023). These AI-driven approaches to experimental design allow researchers to tackle more complex problems and explore hypotheses that might otherwise be impractical due to time or resource constraints.

Data-driven Discovery

Data-driven discovery has become a cornerstone of modern scientific research, leveraging the ever-growing volumes of experimental, observational, and synthetic data to uncover new patterns, relationships, and laws. This paradigm shift has been particularly transformative in fields where complex systems and high-dimensional data are prevalent.

In drug discovery, data-driven approaches have significantly accelerated the identification of potential therapeutic compounds. For instance, recent works employed generative (Mak, Wong, and Pichika 2023; Callaway 2024) and multimodal representation learning (Gao et al. 2024) models to discover a novel antibiotic, effective against a wide range of bacteria, by searching and screening millions of molecules in the representation space (Gao et al. 2024). These enhancements demonstrate the power of AI in exploring vast chemical spaces that would be infeasible to search manually or in the huge and infinite combinatorial space of molecules.

Equation discovery, commonly known as symbolic regression, is a data-driven task for uncovering mathematical expressions from data. Early neural methods like AI Feynman (Udrescu and Tegmark 2020) demonstrated the ability to rediscover fundamental physics laws from data alone, while later work incorporated physical constraints and structures for more interpretable models (Cranmer et al.

2020b). The advent of language modeling and representation learning brought new possibilities. Transformer-based language models, adapted for symbolic regression, treat equation discovery as a numeric-to-symbolic generation task (Biggio et al. 2021; Kamienny et al. 2022). These approaches have been enhanced with search techniques during decoding (Landajuela et al. 2022; Shojaee et al. 2024a), although challenges remain in effectively encoding and tokenizing numeric data (Golkar et al. 2023). Recent works like the SNIP model (Meidani et al. 2024) have also explored multi-modal representation learning between symbolic expressions and numeric data, moving the equation discovery search to a lower-dimensional and smoother representation space for more effective and efficient search. Recently, LLM-SR (Shojaee et al. 2024b) also demonstrated the potential of using LLMs as scientist agents in the evolutionary search for equation discovery. These advancements highlight the evolving landscape of equation discovery, with significant potential for further improvements in integrating numeric data with AI models and leveraging the mathematical reasoning capabilities of advanced LLMs.

In materials discovery, data-driven approaches have led to the prediction and subsequent synthesis of novel materials with desired properties (Pyzer-Knapp et al. 2022; Merchant et al. 2023; Miret and Krishnan 2024). Large generative models have shown remarkable success in generating novel structures. For instance, Merchant et al. (2023) introduced Graph Networks for Materials Exploration (GNoME), leading to the discovery of new stable materials. This approach represents an order-of-magnitude increase in known stable crystals, showcasing the potential of AI in expanding our materials knowledge base. LLMs have also been recently used to extract information from scientific literature in material science, generate novel material compositions, and guide experimental design (Miret and Krishnan 2024). For example, the AtomAgents (Ghafarollahi and Buehler 2024a) demonstrates how LLMs can be integrated into the material discovery pipeline, significantly improving the process in alloy design. By combining the pattern-recognition and representation learning capabilities with the reasoning and generalization abilities of advanced AI models, we are moving towards systems that can not only analyze existing data but also propose novel hypotheses for data-driven discoveries across scientific disciplines.

Key Challenges and Research Opportunities

Benchmarks for Scientific Discovery

First and foremost, evaluating AI systems for open-ended scientific discovery poses unique challenges compared to typical machine learning benchmarks. This challenge is particularly acute for large language models (LLMs) and other foundation models capable of storing and potentially "memorizing" vast amounts of scientific knowledge (Brown 2020; Bommasani et al. 2021) in their parameters. Many existing benchmarks in the field of scientific discovery only focus on rediscovering known scientific laws or solving textbook-style problems. For instance, the AI Feynman dataset consists of 120 physics equations to be rediscovered from data (Udrescu and Tegmark 2020; Udrescu et al. 2020), while datasets like SciBench (Wang et al. 2023c), ScienceQA (Lu et al. 2022), and MATH (Hendrycks et al. 2021) primarily evaluate scientific question answering and mathematical problem-solving abilities.

However, these benchmarks may not capture the entire complexity of scientific discovery processes. More critically, they may be vulnerable to reciting or memorization by large language models, potentially leading to overestimation of true discovery capabilities (Carlini et al. 2021; Shojaee et al. 2024b). As (Wu et al. 2023) points out, LLMs can often solve scientific problems by pattern matching against memorized knowledge rather than through genuine reasoning or discovery. This concern is further emphasized by studies showing that LLMs can reproduce significant portions of their training data (Carlini et al. 2022). There is a pressing need for richer benchmarks and evaluation frameworks in this research area to better understand the gap between baselines and recent methods and to identify areas for improvement. Key directions include:

- *Developing benchmark datasets focused on novel scientific discovery rather than recovery*: One promising approach is to create configurable simulated scientific domains where the underlying laws and principles can be systematically varied. This would allow testing discovery capabilities on new scenarios, mitigating the risk of models simply reciting memorized information observed in their training data. For example, (M. Bran et al. 2024) used a simulated chemistry environment to evaluate AI-driven discovery of novel chemical reactions. Similarly, (Shojaee et al. 2024b) designed simulated settings for different scientific domains such as material science, physics, and biology to evaluate AI-driven scientific equation discovery. A key challenge in this line of research is balancing the use of LLMs' prior scientific knowledge while avoiding mere recitation or memorization. This balance is crucial for advancing AI's role in scientific discovery.
- *Creating evaluation metrics for multiple facets of scientific discovery*: To comprehensively assess scientific discovery capabilities, we need a multi-faceted evaluation framework. The key metrics include: (i) *Novelty*: Measures to quantify how different a discovered hypothesis or law is from existing knowledge. This could involve comparing against a corpus of known scientific literature (Ji et al. 2024); (ii) *Generalizability*: Assessing how well discovered laws or models predict out-of-distribution unobserved data. To do so, evaluation benchmarks should be developed that test discovered laws on scenarios significantly different from the training data distribution, highlighting how scientific theories should be generalizable to new contexts; (iii) *Alignment with Scientific Principles*: Evaluating whether discovered hypotheses are consistent with fundamental laws of physics or other well-established scientific knowledge. This could involve developing formal verification methods for scientific consistency (Cornelio et al. 2023; Cranmer et al. 2020a), as well as assessing the discovered laws' compat-

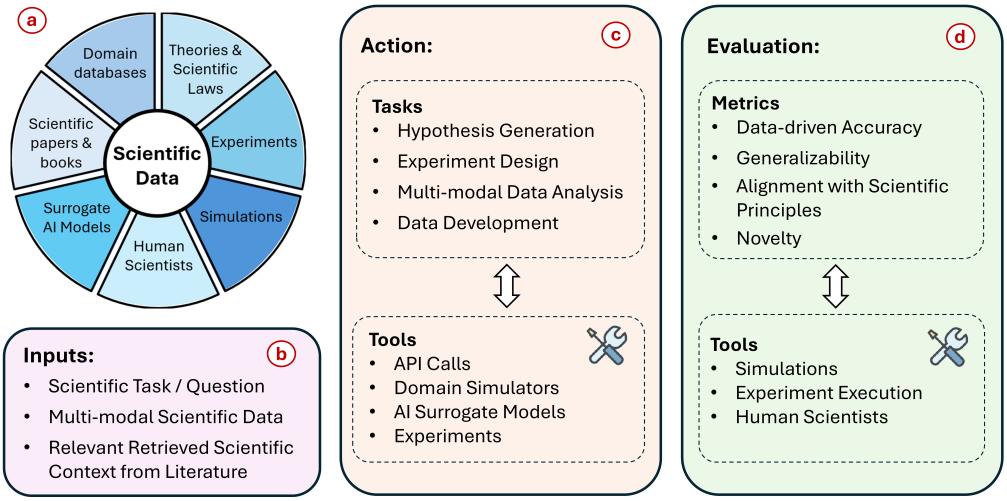

Figure 2: A comprehensive framework for **science-focused AI agents**. The diagram illustrates ⓐ the multi-modal nature of scientific data, ⓑ the inputs for scientific tasks, ⓒ the key actions performed by AI agents in scientific discovery, and ⓓ the evaluation metrics for assessing scientific outcomes. This framework highlights the integration of diverse data sources, AI-driven tools, and human experts in advancing scientific research and discovery processes.

ibility with existing scientific theories (Liu et al. 2024b).

- *Involving domain experts in benchmark design and evaluation*: The involvement of domain experts is crucial for developing meaningful benchmarks and evaluating AI-driven scientific discoveries. Experts can contribute in various aspects of the discovery process such as assessing the plausibility, novelty, and potential impact of AI-generated hypotheses; evaluating the interpretability and alignment of AI-discovered laws or models with human-understandable scientific principles; and providing feedback during the AI-driven discovery process for human-AI collaborative discovery. By integrating domain expert involvement throughout the benchmark development, discovery, and evaluation process, we can ensure that advancements in AI-driven scientific discovery are both technically sound and aligned with the needs and standards of the scientific community.

Science-Focused Agents

Current work on scientific AI often treats models as passive tools rather than active agents pursuing discovery. There is a growing need to develop science-focused AI agents (Figure 2) that can leverage broad scientific knowledge, engage in reasoning, and autonomously verify their reasoning and hypotheses. Recently, LLMs have shown impressive capabilities in knowledge retrieval and reasoning (Huang and Chang 2023), making them promising candidates for developing such agents. These agents can integrate vast amounts of scientific knowledge embedded in LLMs, generate educated hypotheses, design experiments, verify their designs, and interpret the results. Also, their ability to interface with external tools and experimental data sources with the programming execution gate allows for real-world experimentation and validation. Recent work has demonstrated the potential of LLM-based agents in scientific domains. For example, (M. Bran et al. 2024) introduced ChemCrow, an LLM-augmented system for chemistry research. ChemCrow integrates GPT-4 with domain-specific tools for tasks such as reaction prediction, retrosynthesis planning, and safety assessment. This integration allows the system to reason about chemical processes and validate the hypotheses using specialized chemical tools. Similarly, (Ghafarollahi and Buehler 2024a) developed AtomAgents, a multi-agent system for alloy design and discovery. SciAgents (Ghafarollahi and Buehler 2024b) also uses multiple AI agents, each specializing in different aspects of materials science, to collaboratively design new bio-materials. The system incorporates physics-aware constraints and can interface with simulation tools to validate its predictions. However, developing effective science-focused agents also presents several challenges:

- *Domain-specific tool integration*: Effective scientific agents require integration with specialized scientific tools and domain-specific knowledge. This challenge arises from the highly specialized nature of scientific instruments and methodologies, which are often underrepresented in LLMs' training data. (Bubeck et al. 2023) demonstrated that while LLMs like GPT-4 excel in general academic tasks, they struggle with specialized scientific reasoning, particularly in physics and chemistry. Potential research directions include developing modular architectures for integrating domain-specific knowledge bases and tool interfaces, and fine-tuning LLMs on curated scientific datasets. These approaches could enable LLMs to access domain-specific knowledge and interact effectively with specialized scientific tools, enhancing their capabilities in this setting.
- *Adaptive experimental design and hypothesis evolution*: A significant challenge in scientific-focused agents is developing systems capable of long-term, iterative scientific investigations. Such agents must design experi-

ments, interpret results, and refine hypotheses over extended periods while maintaining scientific rigor and avoiding biases. This challenge stems from the complex, multi-stage nature of scientific inquiry, which often involves repeated cycles of experimentation, analysis, and hypothesis adjustment. Potential research directions to address this challenge include meta-learning frameworks enabling agents to improve experimental design and hypothesis refinement strategies across multiple investigations; and hierarchical planning algorithms for managing both short-term experimental steps and long-term scientific discovery objectives.

- *Collaborative scientific reasoning*: Enabling collaborative scientific reasoning in AI agents is crucial for advancing scientific progress. Agents must build on their scientific knowledge, communicate hypotheses, engage in discourse, and critically judge peers' work. Current science agents struggle with deep critical analysis and identifying scientific flaws in AI-driven hypotheses and experimental designs (Birhane et al. 2023). Research opportunities include developing multi-agent systems simulating scientific communities, incorporating domain experts in the multi-agent refinement process, and creating benchmarks to enhance scientific discourse capabilities in science-focused agents.

Multi-modal Scientific Representations

The landscape of scientific data is vast and diverse, encompassing far more than just textual information. While recent advancements in language models have significantly boosted our ability to process and reason with scientific literature, we must recognize that the majority of scientific data exists in forms quite different from natural language. From microscopy images to genomic sequences, from time series sensor data to structured databases and mathematical laws, scientific knowledge is inherently multi-modal (Topol 2023; Wang et al. 2023a). This diversity presents both challenges and opportunities for AI-driven scientific discovery. The challenge lies in developing integrated representation learning techniques that can effectively capture and unify these varied scientific data types. The opportunity, however, is immense: by creating AI systems capable of reasoning across these diverse modalities, we can accelerate scientific discovery in unprecedented ways.

Representation learning offers the potential to distill complex, high-dimensional scientific data into more manageable continuous and low-dimensional forms. This is particularly crucial in scientific domains where high-quality data is limited or expensive to obtain through scientific experiments. By learning multi-modal robust representations with the help of pre-training techniques and synthetic simulation data, we can make more efficient use of limited data, potentially reducing the need for costly scientific experiments and accelerating the pace of discovery. Key directions in this line of research include:

- *Cross-modal scientific representation learning:* Recent work has shown promising results in learning pre-trained joint representations across modalities for different scientific tasks. Notable successes include DrugCLIP (Gao et al. 2024) for joint representations of molecules and protein pockets in drug discovery, Text2Mol (Edwards, Zhai, and Ji 2021) bridging natural language and molecular structures, ProtST (Xu et al. 2023) unifying protein sequences and biomedical text in proteomics, and SNIP (Meidani et al. 2024) linking mathematical expressions with numeric data. These advances demonstrate the potential of cross-modal learning to enhance scientific tasks by leveraging complementary information across modalities. Despite these promising results, significant research opportunities remain (i) *Expanding* cross-modal representation learning to diverse and new scientific domains, (ii) *Enhancing* representation quality through recent integrated self-supervised and multi-modal pre-training; and (iii) Developing unified, modality-agnostic frameworks adaptable to heterogeneous scientific data types.

- *Latent space scientific hypothesis search*: Many scientific discovery tasks involve searching through vast, combinatorial spaces of candidates. Current approaches to these problems often rely on evolutionary search or heuristic methods, which can be computationally expensive and inefficient (Sadybekov and Katritch 2023; Schmidt and Lipson 2009). Recent advances in representation learning offer a promising alternative: conducting scientific hypothesis optimization in learned latent spaces. By moving the search process into the latent space, we can potentially make the exploration of the hypothesis space more efficient and effective. This approach has shown potential across various domains, from drug discovery (Gao et al. 2024) to equation discovery (Meidani et al. 2024), molecular design (Abeer et al. 2024; Zheng, Li, and Zhang 2023), and protein engineering (Castro et al. 2022; Jumper et al. 2021). This emerging research direction has significant potential for scientific discovery. Future research avenues include (i) Integrating domain expert knowledge or feedback into the representations and discovery process, (ii) Enhancing interpretability of representations for scientific validation, and (iii) Advancing optimization techniques for nontrivial discovery objectives and more flexible hypothesis search in the latent space.

- *Multi-modal scientific reasoning frameworks:* The advancement of AI-driven scientific discovery hinges on developing systems capable of multi-modal scientific reasoning. Recent works have shown promising results in this direction. For example, multi-modal retrieval augmented generation (RAG) systems have demonstrated potential in leveraging LLMs for scientific discovery (Park et al. 2024). Models like GIT-Mol (Liu et al. 2024a) showcase the integration of visual, textual, and graph reasoning for molecular discovery. In materials science, approaches combining textual reasoning with structural data have also shown promise in predicting material properties and guiding synthesis (Miret and Krishnan 2024). However, comprehensive multi-modal scientific reasoning frameworks remain an open challenge. Such frameworks must effectively integrate rea-

soning across diverse data types. While studies like (Lu et al. 2022) have shown improved scientific question-answering through combined text and image contexts, further research is needed to explore the impact of other modalities such as numerical or tabular data, and symbolic mathematical theories on scientific discovery tasks.

- *Transfer learning in scientific domains*: Transfer learning offers great potential to accelerate scientific discovery, particularly in domains where data is limited or expensive to obtain. Recent studies have demonstrated its efficacy across various scientific fields: In drug discovery, models pre-trained on large synthetic chemical databases have shown improved performance in predicting properties of novel compounds (Gao et al. 2024). In materials science, transfer learning from simulated data to real-world experiments has also accelerated the discovery of new materials with desired properties (Chen et al. 2024). However, the application of transfer learning in scientific domains presents unique challenges due to the high specificity of scientific knowledge and potential domain shift between source and target tasks. Advancing these capabilities could unlock new avenues for cross-disciplinary discoveries and accelerate progress in data-scarce scientific domains.

Theory and Data Unification

Scientific discovery typically involves a complex interplay between theoretical reasoning, empirical observation, and mathematical modeling. However, most existing AI approaches to scientific tasks focus on just one of these aspects. There is a pressing need for unified frameworks that integrate logical and mathematical reasoning, formal theorem proving, data-driven modeling, experimental design, and causal inference. This integration is challenging but critical for capturing the full scientific discovery process. Recent advances in LLMs have shown promising results in both theorem-proving and data-driven scientific modeling. For instance, LLMs have demonstrated promising capabilities in automated theorem-proving and formal mathematical derivations from natural language problems (Yang et al. 2024; Jiang et al. 2023). On the data-driven side, (Shojaee et al. 2024b; Ma et al. 2024) have shown success in discovering equation hypotheses from data with the help of LLM-based program search. However, these approaches largely operate in isolation, and there is a significant gap in unifying these capabilities to mirror the holistic nature of scientific inquiry. Key challenges and research directions include:

- *Generating derivable hypotheses from empirical observations*: Developing methods that can not only discover patterns in data but also produce rigorous mathematical derivations of these findings is crucial for ensuring the reliability and generalizability of AI-driven scientific discoveries to out-of-distribution data. Derivable theoretical results provide a level of confidence and understanding that goes beyond mere empirical correlation. Recent work, such as the AI-Descartes system (Cornelio et al. 2023), has shown promise by combining equation discovery tools (known as symbolic regression) with automated logical reasoning. However, integrating logical reasoning and data-driven frameworks that are adaptable across scientific discovery tasks still remains an open challenge. Research opportunities exist to automate proof verification, incorporate expert feedback, and embed derivability constraints in data-driven discovery algorithms.
- *Combining symbolic and neural approaches*: How can we effectively integrate the strengths of symbolic reasoning (e.g., logical deduction, formal proofs) with the flexibility and learning capabilities of neural networks? Recent work on neuro-symbolic AI (Garcez and Lamb 2023; Sheth, Roy, and Gaur 2023) provides promising directions, but challenges remain in scaling these approaches to more complex settings and scientific tasks. Developing hybrid architectures that can transition between symbolic and neural representations is helpful in capturing the full spectrum of scientific reasoning.
- *Reasoning discovery uncertainty in formal frameworks*: Scientific discoveries often involve uncertainties and probabilities, yet formal logical frameworks struggle to incorporate these aspects. Developing frameworks that can handle probabilistic reasoning while maintaining rigorous deduction capabilities is crucial for advancing AI-driven scientific discovery. Recent work, such as probabilistic logic systems (De Raedt and Kimmig 2015; De Raedt, Kimmig, and Toivonen 2007), and neuro-symbolic programming (Ahmed et al. 2022) has made progress in this direction. However, significant challenges remain for the use of these approaches in scientific discovery, including scalability to large-scale scientific problems, and expressiveness to capture complex scientific theories in specific scientific domains.

Conclusion

Developing unified AI systems for scientific discovery is an ambitious goal, but one with substantial potential impact. Success could dramatically accelerate progress across diverse scientific disciplines. This paper has outlined current progress as well as several key research challenges and opportunities toward this vision, including developing science-focused AI agents, creating improved benchmarks, advancing multimodal representations, and unifying diverse modes of scientific reasoning. Tackling these challenges will require collaboration between AI researchers, scientists across domains, and philosophers of science. While fully autonomous AI scientists may still be far off, nearer-term progress could produce powerful AI assistants to augment human scientific capabilities. Such tools could help scientists navigate the ever-growing scientific literature, brainstorm ideas, generate novel hypotheses, design experiments, and find unexpected patterns in complex experimental data. By pursuing this research agenda, the machine learning and AI community has an opportunity to develop systems that do not just automate product-related tasks, but actively push forward the frontiers of human scientific knowledge. The path will be challenging, but the potential rewards - both scientific and technological - are immense.

Acknowledgments

This research was partially supported by the U.S. National Science Foundation (NSF) under Grant No. 2416728.

References

Abeer, A. N.; Urban, N. M.; Weil, M. R.; Alexander, F. J.; and Yoon, B.-J. 2024. Multi-objective latent space optimization of generative molecular design models. *Patterns*.

Ahmed, K.; Teso, S.; Chang, K.-W.; Van den Broeck, G.; and Vergari, A. 2022. Semantic probabilistic layers for neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 35: 29944–29959.

Arlt, S.; Duan, H.; Li, F.; Xie, S. M.; Wu, Y.; and Krenn, M. 2024. Meta-Designing Quantum Experiments with Language Models. *arXiv preprint arXiv:2406.02470*.

Baldi, P.; Sadowski, P.; and Whiteson, D. 2014. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1): 4308.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Biggio, L.; Bendinelli, T.; Neitz, A.; Lucchi, A.; and Parascandolo, G. 2021. Neural symbolic regression that scales. In *International Conference on Machine Learning*, 936–945. Pmlr.

Birhane, A.; Kasirzadeh, A.; Leslie, D.; and Wachter, S. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5(5): 277–280.

Böhme, S.; and Nipkow, T. 2010. Sledgehammer: judgement day. In *Automated Reasoning: 5th International Joint Conference, IJCAR 2010, Edinburgh, UK, July 16-19, 2010. Proceedings 5*, 107–121. Springer.

Boiko, D. A.; MacKnight, R.; Kline, B.; and Gomes, G. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992): 570–578.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cai, T.; Merz, G. W.; Charton, F.; Nolte, N.; Wilhelm, M.; Cranmer, K.; and Dixon, L. J. 2024. Transforming the bootstrap: Using transformers to compute scattering amplitudes in planar n= 4 super yang-mills theory. *Machine Learning: Science and Technology*.

Callaway, E. 2024. Major AlphaFold upgrade offers boost for drug discovery. *Nature*, 629(8012): 509–510.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Castro, E.; Godavarthi, A.; Rubinfien, J.; Givechian, K.; Bhaskar, D.; and Krishnaswamy, S. 2022. Transformer-based protein generation with regularized latent space optimization. *Nature Machine Intelligence*, 4(10): 840–851.

Chen, A.; Wang, Z.; Vidaurre, K. L. L.; Han, Y.; Ye, S.; Tao, K.; Wang, S.; Gao, J.; and Li, J. 2024. Knowledge-Reuse Transfer Learning Methods in Molecular and Material Science. *arXiv preprint arXiv:2403.12982*.

Cornelio, C.; Dash, S.; Austel, V.; Josephson, T. R.; Goncalves, J.; Clarkson, K. L.; Megiddo, N.; El Khadir, B.; and Horesh, L. 2023. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14(1): 1777.

Cranmer, M.; Greydanus, S.; Hoyer, S.; Battaglia, P.; Spergel, D.; and Ho, S. 2020a. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*.

Cranmer, M.; Sanchez Gonzalez, A.; Battaglia, P.; Xu, R.; Cranmer, K.; Spergel, D.; and Ho, S. 2020b. Discovering symbolic models from deep learning with inductive biases. *Advances in neural information processing systems*, 33: 17429–17442.

De Raedt, L.; and Kimmig, A. 2015. Probabilistic (logic) programming concepts. *Machine Learning*, 100: 5–47.

De Raedt, L.; Kimmig, A.; and Toivonen, H. 2007. ProbLog: a probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, 2468–2473. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Edwards, C.; Zhai, C.; and Ji, H. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 595–607.

Gao, B.; Qiang, B.; Tan, H.; Jia, Y.; Ren, M.; Lu, M.; Liu, J.; Ma, W.-Y.; and Lan, Y. 2024. Drugclip: Contrasive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36.

Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review*, 56(11): 12387–12406.

Ghafarollahi, A.; and Buehler, M. J. 2024a. AtomAgents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv preprint arXiv:2407.10022*.

Ghafarollahi, A.; and Buehler, M. J. 2024b. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Golkar, S.; Pettee, M.; Eickenberg, M.; Bietti, A.; Cranmer, M.; Krawezik, G.; Lanusse, F.; McCabe, M.; Ohana, R.; Parker, L.; et al. 2023. xval: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Han, J. M.; Rute, J.; Wu, Y.; Ayers, E. W.; and Polu, S. 2021. Proof artifact co-training for theorem proving with language models. *arXiv preprint arXiv:2102.06203*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Huang, J.; and Chang, K. C.-c. 2023. Towards Reasoning in Large Language Models: Survey, Implication, and Reflection. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Huang, K.; Qu, Y.; Cousins, H.; Johnson, W. A.; Yin, D.; Shah, M.; Zhou, D.; Altman, R.; Wang, M.; and Cong, L. 2024. Crispr-GPT: An LLM agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*.

Ji, H.; Wang, Q.; Downey, D.; and Hope, T. 2024. SCIMON: Scientific Inspiration Machines Optimized for Novelty. In *ACL Anthology: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 279–299*. University of Illinois Urbana-Champaign/CABBI.

Jiang, A. Q.; Welleck, S.; Zhou, J. P.; Lacroix, T.; Liu, J.; Li, W.; Jamnik, M.; Lample, G.; and Wu, Y. 2023. Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs. In *The Eleventh International Conference on Learning Representations*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kamienny, P.-A.; d'Ascoli, S.; Lample, G.; and Charton, F. 2022. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35: 10269–10281.

Lample, G.; Lacroix, T.; Lachaux, M.-A.; Rodriguez, A.; Hayat, A.; Lavril, T.; Ebner, G.; and Martinet, X. 2022. Hypertree proof search for neural theorem proving. *Advances in neural information processing systems*, 35: 26337–26349.

Landajuela, M.; Lee, C. S.; Yang, J.; Glatt, R.; Santiago, C. P.; Aravena, I.; Mundhenk, T.; Mulcahy, G.; and Petersen, B. K. 2022. A unified framework for deep symbolic regression. *Advances in Neural Information Processing Systems*, 35: 33985–33998.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Liu, P.; Ren, Y.; Tao, J.; and Ren, Z. 2024a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171: 108073.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024b. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6): bbac409.

M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1–11.

Ma, P.; Wang, T.-H.; Guo, M.; Sun, Z.; Tenenbaum, J. B.; Rus, D.; Gan, C.; and Matusik, W. 2024. LLM and Simulation as Bilevel Optimizers: A New Paradigm to Advance Physical Scientific Discovery. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 33940–33962. PMLR.

MacColl, H. 1897. Symbolic reasoning. *Mind*, 6(24): 493–510.

Mak, K.-K.; Wong, Y.-H.; and Pichika, M. R. 2023. Artificial intelligence in drug discovery and development. *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, 1–38.

Meidani, K.; Shojaee, P.; Reddy, C. K.; and Farimani, A. B. 2024. SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-training. In *The Twelfth International Conference on Learning Representations*.

Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; and Cubuk, E. D. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990): 80–85.

Miret, S.; and Krishnan, N. 2024. Are LLMs Ready for Real-World Materials Discovery? *arXiv preprint arXiv:2402.05200*.

Park, N. H.; Callahan, T. J.; Hedrick, J. L.; Erdmann, T.; and Capponi, S. 2024. Leveraging Chemistry Foundation Models to Facilitate Structure Focused Retrieval Augmented Generation in Multi-Agent Workflows for Catalyst and Materials Design. *arXiv preprint arXiv:2408.11793*.

Polu, S.; and Sutskever, I. 2020. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*.

Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; and Curioni, A. 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1): 84.

Sadybekov, A. V.; and Katritch, V. 2023. Computational approaches streamlining drug discovery. *Nature*, 616(7958): 673–685.

Schmidt, M.; and Lipson, H. 2009. Symbolic regression of implicit equations. In *Genetic programming theory and practice VII*, 73–85. Springer.

Segler, M. H.; Preuss, M.; and Waller, M. P. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698): 604–610.

Sheth, A.; Roy, K.; and Gaur, M. 2023. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intelligent Systems*, 38(3): 56–62.

Shojaee, P.; Meidani, K.; Barati Farimani, A.; and Reddy, C. 2024a. Transformer-based planning for symbolic regression. *Advances in Neural Information Processing Systems*, 36.

Shojaee, P.; Meidani, K.; Gupta, S.; Farimani, A. B.; and Reddy, C. K. 2024b. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Topol, E. J. 2023. As artificial intelligence goes multimodal, medical applications multiply.

Udrescu, S.-M.; Tan, A.; Feng, J.; Neto, O.; Wu, T.; and Tegmark, M. 2020. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems*, 33: 4860–4871.

Udrescu, S.-M.; and Tegmark, M. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16): eaay2631.

Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. 2023a. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47–60.

Wang, H.; Yuan, Y.; Liu, Z.; Shen, J.; Yin, Y.; Xiong, J.; Xie, E.; Shi, H.; Li, Y.; Li, L.; et al. 2023b. Dt-solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12632–12646.

Wang, R.; Zelikman, E.; Poesia, G.; Pu, Y.; Haber, N.; and Goodman, N. 2024. Hypothesis Search: Inductive Reasoning with Language Models. In *The Twelfth International Conference on Learning Representations*.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2023c. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Wu, Z.; Qiu, L.; Ross, A.; Akyürek, E.; Chen, B.; Wang, B.; Kim, N.; Andreas, J.; and Kim, Y. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Xu, M.; Yuan, X.; Miret, S.; and Tang, J. 2023. Protst: Multimodality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, 38749–38767. PMLR.

Yang, K.; Swope, A.; Gu, A.; Chalamala, R.; Song, P.; Yu, S.; Godil, S.; Prenger, R. J.; and Anandkumar, A. 2024. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.

Zhang, D.; Hu, Z.; Zhoubian, S.; Du, Z.; Yang, K.; Wang, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024. SciGLM: Training Scientific Language Models with Self-Reflective Instruction Annotation and Tuning. arXiv:2401.07950.

Zheng, W.; Li, J.; and Zhang, Y. 2023. Desirable molecule discovery via generative latent space exploration. *Visual Informatics*, 7(4): 13–21.