https://doi.org/10.1038/s44387-025-00018-6

AI-enabled scientific revolution in the age of generative AI: second NSF workshop report

Anuj Karpatne[1,6], Aryan Deshwal[2,6], Xiaowei Jia[3,6], Wei Ding[4], Michael Steinbach[2], Aidong Zhang[5] & Vipin Kumar[2] ✉

Artificial intelligence (AI) is transforming science and, in turn, being advanced by scientific challenges. A 2023 NSF-sponsored workshop at NSF Headquarters launched a national dialogue on this synergy. A second NSF-sponsored workshop, held in August 2024 at the University of Minnesota, revisited those insights in light of the GenAI revolution. This report captures key discussions and recommendations to guide the development of GenAI aligned with scientific discovery.

Artificial intelligence (AI) is poised to transform science by introducing groundbreaking approaches with far-reaching societal implications. At the same time, the unique challenges posed by scientific problems create opportunities to advance AI itself, fostering a virtuous cycle: fundamental AI research drives scientific discovery, while science-inspired challenges push the boundaries of AI development. In March 2023, a workshop[1] was held at the NSF Headquarters to identify key challenges and actionable steps needed to enable the next AI revolution in the sciences. The discussions and recommendations from this first workshop are summarized in this report[2].

Soon after the first workshop, the Generative AI (GenAI) revolution—exemplified by the remarkable success of foundation models[3,4] and applications like ChatGPT[5] that leverage these models—began to see widespread adoption across science and engineering. Specifically, foundation models (i.e., pre-trained models capable of being fine-tuned on a wide variety of tasks) are now being used across nearly every scientific discipline[6,7]. In some cases, these models already outperform traditional methods employed by their respective scientific communities[8], highlighting their transformative potential in driving scientific discovery.

Objectives

The second workshop[9], held on 6 August 2024, at the University of Minnesota, aimed to revisit the recommendations of the first workshop in light of recent advancements in AI. A key objective was to evaluate the potential of emerging technologies, such as generative AI, to drive scientific discovery while assessing their limitations (e.g., hallucinations[10,11]) and exploring approaches to enhance their reliability and effectiveness (e.g., by incorporating scientific theories). The workshop further focused on identifying gaps in the current capabilities of generative AI for addressing grand scientific challenges and outlining the advancements needed to foster synergistic research across scientific disciplines. Additionally, it sought to identify actionable strategies for effectively integrating these advancements into the practice of scientific discovery.

Workshop structure

The workshop was designed to enable an open discussion on the promise of GenAI for accelerating scientific discoveries and driving the next frontiers in GenAI that are shaped by the unique needs of a diverse set of science and engineering disciplines. It brought together 31 leading experts from AI and various scientific and engineering fields, including computational biology, health sciences, neuroscience, materials informatics, ecology, climate science, hydrology, limnology, and physics. Participants represented a broad mix of institutions—spanning academia, industry, and federal mission-driven agencies—with about half having also attended the first workshop. Appendix A contains the list of participants.

The workshop program consisted of four sessions that alternated between lightning talks and panel discussions. Lightning talks addressed two key questions: (i) How is GenAI being used in scientific disciplines, and (ii) what new opportunities in science can be enabled by advances in generative AI. Panel discussions focused on: (i) what is missing in the current state of generative AI to meet the needs of scientific problems, and (ii) how to build the next generation ecosystem for generative AI innovations that is driven by the needs of the scientific community. Each session was assigned a pair of scribes, who captured the discussion and kept notes in a Google Doc accessible to all attendees. Workshop attendees were also given the option of providing a 1-page position paper addressing the four questions prior to the workshop. Nearly two-thirds of the attendees took this option, and these position papers enabled framing of the discussion at the workshop. The workshop opened with welcome remarks from NSF and UMN leadership. This was followed by a quick recap of the report of the first workshop and a summary of the position papers submitted by the attendees.

[1]Virginia Tech, Blacksburg, VA, USA. [2]University of Minnesota, Minneapolis, MN, USA. [3]University of Pittsburgh, Pittsburgh, PA, USA. [4]University of Massachusetts Boston, Boston, MA, USA. [5]University of Virginia, Charlottesville, VA, USA. [6]These authors contributed equally: Anuj Karpatne, Aryan Deshwal, Xiaowei Jia. ✉e-mail: kumar001@umn.edu

Outline of this perspective

This perspective article attempts to capture the wide-ranging discussions at the workshop as well as some concrete recommendations that the workshop participants considered critical for advancing the state of the art in GenAI for accelerating scientific discovery. Specifically, "Current and potential applications of generative AI in science" provides an overview of how GenAI is beginning to be used in scientific disciplines and what new opportunities in science can be enabled by advances in GenAI. "Gaps in the current state of generative AI to meet the needs of scientific problems" focuses on some of the major gaps in current frameworks of GenAI that limit them from meeting the unique needs of current and envisioned scientific problems. "Potential solutions to fix the current gaps in generative AI" discusses potential solutions for fixing these gaps. "Building the next generation ecosystem for generative AI in science" outlines steps that can be taken to build the next-generation ecosystem for GenAI innovations that are driven by the needs of the scientific community. Figure 1 provides an overview of the workshop report summarizing key topics described in each of the four sections.

Current and potential applications of generative AI in science

GenAI is powered by two major advances in deep learning: (i) self-supervised learning[12], which has enabled us to learn useful feature representations outside the traditional confines of supervised datasets and (ii) context-sensitive AI architectures (e.g., attention-based Transformers[13]) that are able to adapt their outputs based on the context-specific needs of a data sample, allowing a high degree of controllability in the generated samples based on input conditionings. These advances, combined with the availability of large-scale and diverse data and massive computing capabilities, have brought us to a new era in AI, where we can generate data samples of remarkable complexity across a wide range of applications, including natural language generation[14], code completion[15], and text-to-image synthesis[16]. They power conversational AI systems capable of producing human-like dialogue, assist programmers by suggesting and even writing functional code, and enable artists and designers to create high-quality visual content from textual descriptions. Beyond these mainstream applications, Generative AI is beginning to make inroads into scientific domains such as protein structure prediction[17], drug discovery[18], materials design[19], and climate modeling[20], where it can rapidly explore vast chemical spaces, predict molecular properties, and simulate complex physical systems. By automating intricate processes, uncovering hidden patterns, and enabling large-scale data-driven experimentation, these models are poised to transform the way scientific research is conducted, accelerating discovery and innovation in ways previously unimaginable.

There are two broad approaches to using Generative AI in scientific research. The first involves leveraging pre-trained foundation models, such as large language models (LLMs)[14] and vision-language models (VLMs)[21], which are trained on Internet-scale text and image corpora. The vast number of parameters in these models enables emergent capabilities, such as in-context learning[22], where the model can incorporate new information simply through natural language prompts. Specifically, in-context learning can be used to guide the model's responses and improve alignment with specific tasks. Additionally, we can also fine-tune on domain-specific data by updating the model parameters. This approach is particularly useful in domains where scientific data is naturally represented as text and/or images, making pre-trained LLMs and VLMs effective starting points. A growing number of scientific applications are exploring the use of pre-trained foundation models, including drug discovery[18], theorem proving[23], organismal biology[24], and medical image diagnosis[25]. The second approach is to develop domain-specific foundation models from scratch for scientific applications that require processing richer data formats beyond text and images. This is particularly relevant in fields such as environmental science[26], climate science[27], chemistry[28], materials science[29], biological sciences[30,31], agricultural science[32], and transportation[33], where data is often structured as graphs, 3D structures, time-series, spectral/hyperspectral images, or spatio-temporal fields. Training specialized foundation models in these domains enables deeper integration with scientific data and enhances their ability to generate meaningful insights tailored to complex real-world problems. Note that both approaches may be applicable to a specific domain depending on the data available. Next, we describe some broad classes of

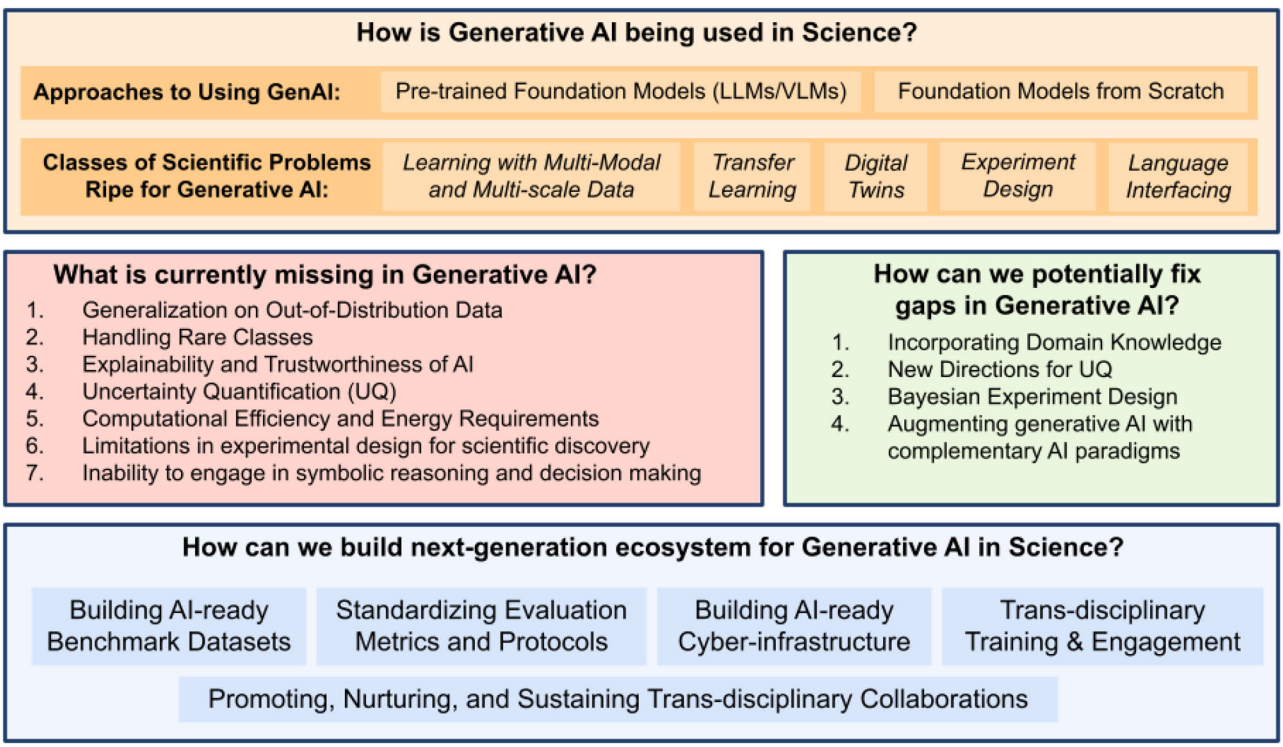

Fig. 1 | Overview of the workshop report summarized in a Figure. This report cover current and potential applications of generative AI in science, key gaps in generative AI for science, potential solutions to fix these gaps, and recommendations to build the next-generation ecosystem of generative AI in science.

scientific problems that are particularly well-suited for harnessing the power of generative AI.

Learning with multi-modal and multi-scale data

Many scientific phenomena arise from complex interactions among natural processes governed by diverse scientific principles across multiple disciplines (e.g., physical, chemical, biological, social). These processes operate and interact at varying spatial and temporal scales—for example, atmospheric dynamics unfolding over days and thousands of kilometers, versus microbial activity occurring within hours and microns. Observations of such phenomena are typically collected through diverse instruments and platforms, ranging from satellite-based remote sensors and ground-based monitoring stations to laboratory assays and textual records in scientific literature. This results in data that differ not only in resolution and fidelity but also in modality, including time series, images, tabular measurements, geospatial grids, and unstructured text.

Foundation models such as LLMs and VLMs offer a compelling solution to this complexity. Their ability to learn generalizable feature representations from large, heterogeneous datasets enables the integration —or fusion—of information across these varied sources. This is particularly valuable for scientific discovery, where insight often depends on drawing connections across data types and scales. For instance, foundation models can combine satellite imagery with meteorological time series and topographic maps to improve environmental modeling[34], or synthesize experimental results with scientific literature to guide new materials design[35]. Their multi-modal and multi-scale learning capabilities position them as foundational tools for advancing interdisciplinary science.

Transfer learning from highly observed to sparsely observed systems

A persistent challenge in many scientific domains is the scarcity of labeled or high-quality data. While some systems—such as well-monitored urban environments or heavily studied ecosystems—generate abundant observational data, many others remain undersampled due to cost, inaccessibility, or technical limitations. Foundation models, by virtue of their pretraining on large, diverse datasets, provide a powerful mechanism for addressing this imbalance through transfer learning. Once trained on richly annotated or data-dense systems, these models can be fine-tuned or adapted to new related tasks and environments where data is sparse. This allows scientific insights and predictive capabilities to be extended from data-rich contexts to less well-observed ones. For example, a foundation model trained on high-resolution water quality data from hundreds of extensively monitored lakes can be adapted to infer pollutant levels in remote or data-poor lakes using only a handful of measurements. Similarly, models trained on weather records from well-instrumented regions can support forecasting in emerging economies where historical data is limited. This ability to transfer knowledge across domains, scales, and data regimes makes foundation models especially attractive for accelerating discovery in the face of limited observational resources.

Building digital twins

There is a growing interest in developing digital twins of scientific and environmental systems that are capable of simulating complex processes under different conditions, and assimilating new data and human feedback. These digital twins hold transformative potential across disciplines. For example, they can simulate social interactions and contact dynamics to better understand vaccine adoption. In social science research, generative AI can be used to create digital personas that augment or even replace human participants in studies. Similarly, traffic control policies—often too costly or ethically complex to test in real life—can be explored through generative simulations.

Generative AI is enabling a new generation of powerful and flexible simulators, significantly enhancing the design and capabilities of digital twins for complex scientific systems. Traditionally, digital twins have relied on hybrid approaches that integrate process-based models, data-driven algorithms, and visualization tools to meet diverse modeling and analysis needs. Recent advances in generative AI allow for the creation of more expressive and adaptive simulators that can better capture system complexity and support scenario exploration. AI-powered simulators can be continuously updated and refined through human interaction and policy feedback. This adaptability stems from several capabilities of generative AI: (1) Generative AI models can ingest and integrate diverse data sources—such as remote sensing and ground-based observations—to capture multiple facets of a system and extract spatio-temporal patterns for more accurate simulations. (2)Through prompt engineering, generative models can simulate system behaviors over large spatial and temporal scales, allowing researchers to explore outcomes under specific scenarios, such as new policies for traffic control or natural resource management. (3) These models can assimilate new data and incorporate human feedback, enabling dynamic adjustments to the simulated processes as conditions evolve. Together, these developments not only improve the fidelity and functionality of digital twins but also broaden their accessibility and applicability across scientific domains.

Candidate generation for experimental design

Many scientific discoveries and engineering design problems involve searching over exponentially large design spaces. For example, in drug discovery, the number of potential small molecules is estimated to exceed 10^{60}. Similarly, hardware design requires searching over a large combinatorial space of candidate chips and circuits. In additive manufacturing, we require searching over many candidate 3D printing designs. This challenge is exacerbated by the fact that evaluation of each candidate design is typically extremely expensive in terms of consumed resources (e.g., a wet-lab experiment). Unlike previous generations of deep generative models like variational autoencoders, modern methods like diffusion models are more expressive and generate higher-quality samples. This was first observed in the context of image generation but is also now seen in scientific domains (e.g., molecule or protein generation). Additionally, this improved generation capability also allows researchers to specify precise requirements, such as incorporating specific functional groups in molecule design, enabling more targeted exploration of the chemical space.

LLM-based natural language interfaces for scientific discovery

Generative AI, particularly large language models, is enabling researchers to investigate previously challenging scientific questions through several key mechanisms. One such mechanism is their natural language interface, which provides a flexible framework for encoding domain knowledge, integrating constraints, and formulating scientific queries in an intuitive way. This makes LLMs especially useful for facilitating interdisciplinary research. For example, in materials science, LLMs can act as translators for domain-specific concepts, improving communication between chemists, physicists, and computer scientists, as well as synthesizing insights from diverse scientific literature that spans these fields. Furthermore, LLMs may enable scientists to incorporate high-level side information that is easily expressible in natural language. For example, using LLMs for modeling stochastic processes (e.g., temperature, precipitation) can allow conditioning with scenarios described in natural language, like the location of a place or data[36]. Another emerging capability is program synthesis: LLMs can generate executable code to explore new hypotheses or automate scientific workflows. For instance, they have recently been used to generate heuristic functions for solving problems in combinatorics[37].

Gaps in the current state of generative AI to meet the needs of scientific problems

Generative AI methods have shown transformative capabilities in areas such as natural language processing and computer vision, but there are multiple technical challenges that need to be addressed in realizing the full potential of these techniques for scientific problems. This section examines six key challenges that represent critical areas for future research and development: 1) Robustness and Generalization to Out-of-Distribution Data, 2)

Challenges in Handling Rare Classes, 3) Explainability and Trustworthiness Concerns, 4) Uncertainty Quantification (UQ), 5) Computational and Energy Efficiency Constraints, 6) Limitations in Reasoning and Decision-Making under Uncertainty, and 7) Inability to Engage in Symbolic Reasoning and Decision Making. In the subsequent subsections, we specifically analyze how these challenges arise within scientific contexts and differentiate them from similar issues in more traditional AI methods.

Challenges of hallucination, robustness, and generalization in out-of-distribution settings

Since modern generative AI models fundamentally operate by learning from data, their performance is inherently bounded by the quality and diversity of their training data. These models often struggle with brittleness when presented with out-of-distribution inputs during inference, producing incorrect, incoherent, and inconsistent outputs. For example, large-scale generative models in vision and language frequently exhibit "hallucinations" when presented with novel prompts or testing scenarios not present in their training data. An analogous problem of hallucination is encountered in climate/weather applications where AI models display inconsistencies such as fine-scale features disappearing during longer runs, patch or mesh artifacts appearing at higher spatial resolutions, and divergent errors accumulating over time.

These challenges in extrapolating to out-of-distribution data are particularly concerning in scientific applications for two main reasons. First, available data in many scientific applications often under-represent the full range of possible data distributions encountered in novel real-world testing scenarios. This limitation stems from the substantial costs associated with collecting ground-truth annotated data in scientific applications, typically including domain expert human labor, time, and expensive materials. Consequently, generative models may struggle to generalize when tested outside the temporal periods, spatial regions, scales, or operating regimes/parameters of scientific systems used during training. Second, extrapolating to out-of-distribution data is a fundamental requirement in many scientific applications where the goal is to extract generalizable scientific patterns from data that can explain, simulate, and predict scientific phenomena in unseen regimes, potentially leading to the discovery of new materials and drugs or forecasting climate change impacts for hypothesized forcings of CO_2 emissions. Addressing these challenges requires advances in generative AI frameworks that not only excel at general-purpose tasks but also possess the capability to address specialized requirements of specific scientific domains using limited data.

Handling rare classes

While current generative AI models excel at modeling "typical" patterns in data observed across the majority of samples, they struggle to capture "atypical" patterns that occur infrequently in rare classes or heavy tails of data distributions. This can be attributed to a variety of reasons, including the choice of standard loss functions that are not designed to be sensitive to performance on rare classes, and inherent biases in data collection and annotation efforts favoring certain classes over others. In societal real-world applications, the ability to model rare classes or heavy tails of distributions is critical for operational and decision-making settings. For instance, in designing flood warning systems, accurately modeling extreme precipitation events is essential despite their infrequent occurrence. Challenges in modeling rare classes are further exacerbated by the poor quality of data in many real-world applications. Real-world data often contains noise and missing values at varying spatial and temporal resolutions, making it difficult to build reliable models that can handle both typical and atypical cases effectively.

Explainability and trustworthiness of AI models

While current generative AI frameworks have achieved widespread adoption due to their exceptional performance in vision and language domains, they remain black-box systems that cannot reliably explain their reasoning process. This makes it challenging to diagnose error sources or understand why models generate specific outputs, undermining their trustworthiness even when they produce correct results. Furthermore, while commercial applications often prioritize predictive accuracy, scientific applications care about a fundamentally different end-goal: discovering and understanding the causal pathways relating inputs to outputs, which can serve as building blocks for advancing scientific knowledge. The current black-box nature of AI models limits their utility in developing explainable theories and identifying relationships found in data, limiting their scope in applications aimed at accelerating scientific discovery.

Research in explainable AI is challenging for a number of reasons. First, evaluating AI explanations involves inherently qualitative and subjective aspects, unlike the well-defined metrics used for traditional tasks like image classification or multiple-choice questions. Various tools in explainable AI (e.g., saliency maps, attention maps, class activation maps, and prototypical part networks) employ different approaches and serve different explanatory goals, making it challenging to establish a unified evaluation framework. Second, modern AI models have an extremely large number of parameters (in the order of billions to trillions), which makes it difficult to associate them with semantic meaning expressed using a limited vocabulary of abstract features.

Beyond the explainability of AI models, developing trustworthy AI requires addressing several other concerns. This includes addressing issues related to the privacy concerns of generative models, to ascertain whether and how sensitive information from the training data may inadvertently be revealed through model outputs. For example, generative models might reproduce verbatim text from private training data, creating serious privacy vulnerabilities. These datasets may include unpublished experimental results, proprietary chemical formulations, patient-specific biomedical data, confidential genomic sequences, classified environmental impact assessments, trade-secret engineering designs, or secure government research findings. Ensuring that these systems can be audited and monitored for such behavior is critical for protecting individual and institutional privacy. The lack of transparency in generative AI models also creates significant security vulnerabilities, particularly to adversarial attacks that can induce the generation of harmful or misleading content. These models often inherit and amplify biases present in their training data, potentially perpetuating societal inequalities through their outputs.

Uncertainty quantification

UQ[38] represents a critical outstanding challenge for generative AI systems, particularly for high-stakes scientific applications. For example, UQ is critical for decision-making in scientific contexts like hurricane prediction, which can have severe consequences for both false positives and false negatives. Hence, in addition to the output label of True or False, what these applications need is the confidence associated with these labels. Concretely, this means that we want the uncertainty provided by the models to be well-calibrated[23,24]. This contrasts with traditional domains such as e-commerce, where high false positive and false negative rates are more tolerable as long as a few recommendations are successful.

The ability to quantify uncertainty is especially important in self-driving automated experiments, which leverage Bayesian Optimization (BO) strategies to make informed decisions when selecting experiments. Accurate uncertainty estimates allow these systems to explore efficiently while operating under tight resource constraints. Traditional approaches for quantifying uncertainty (e.g., ensembles, MC-Dropout, and posterior sampling) that are commonly employed for small-scale models are not necessarily well-calibrated. In addition, they are computationally intractable for large-scale generative AI models due to their requirement for multiple forward passes or maintaining multiple copies of the network. While LLMs inherently produce probability distributions via their autoregressive formulation, these outputs are often poorly calibrated. More critically, LLMs tend to conflate syntactic fluency with semantic correctness—a limitation that undermines reliable UQ in natural language generation[39]. Furthermore, these probability estimates degrade with sequence length due to the multiplicative

accumulation of token probabilities, leading to poor calibration for longer texts.

Computational efficiency and energy requirements

Generative AI models face computational efficiency challenges for both training and inference. Canonical pre-training of such models typically requires a large amount of computational resources. During post-training, although a single-step inference can be expensive, multi-step inference settings present even greater challenges. This is particularly evident in diffusion models, where generating high-quality samples requires iterating through hundreds or thousands of denoising steps. These computational challenges become especially problematic for scientific applications that require extensive sampling to capture the full spectrum of variations possible. For instance, modeling biodiversity of terrestrial ecosystems requires generating diverse vegetation patterns to accurately represent land surface dynamics, which capture energy fluxes as well as carbon and water cycles. The energy and computational requirements for both training and inference present a significant barrier, particularly in resource-constrained academic settings where access to high-performance computing infrastructure may be limited.

Limitations in experimental design for scientific discovery

Many scientific discoveries and engineering design problems (e.g., molecule design, materials discovery, protein engineering, hardware design, additive manufacturing) can be instantiated as expensive experimental design tasks. In a typical experimental design setup for scientific discovery, we need to explore large combinatorial spaces in order to reach a desired goal (e.g., optimize an objective of interest) guided by expensive experiments in a resource-efficient manner. Recently, generative AI models (e.g., diffusion models) have shown great promise in modeling distributions of complex high-dimensional scientific objects. These include molecules, chips and circuits represented as graphs, materials as periodic 3D structures, and proteins as sequences. However, one significant limitation of such models in experimental design is that they require generating a large number of candidates to reach the design goal, which becomes impractical for scientific discovery, where each experimental validation is expensive in terms of resources (e.g., synthesizing a material with a physical wet-lab experiment).

Inability to engage in symbolic reasoning and make decisions over long horizons

The reasoning capabilities of Large Language Models and other generative AI models lie on a wide spectrum, with notable strengths in handling pattern recognition and knowledge synthesis. For instance, these models excel at tasks requiring broad knowledge integration, such as scientific literature synthesis[40]. However, these models struggle with several other reasoning tasks. In particular, these models face significant limitations in long-horizon tasks that require long chains of reasoning and decision-making under uncertain environments. For example, while an LLM might generate text describing a chemical synthesis, it cannot reliably reason through the steps to design a novel synthesis pathway, especially if it involves reaction conditions or molecules outside its training data. Another such example is a problem that requires deciding the long-term consequences of interventions in complex systems from domains such as ecology, biology, climate science, e.g., studying perturbations in cells, which are complex biological systems, arising from a large number of molecular interactions[41]. Additionally, they struggle with precise symbolic reasoning, tasks requiring formal logic, and systematic manipulation of symbols. The current generation of generative models also shows limited reliability at higher levels of abstraction, such as deriving new mathematical proofs or discovering new physical principles.

Potential solutions to fix the current gaps in generative AI

This section discusses a number of solutions to address the limitations of current GenAI approaches in the context of scientific domains. Specifically, we explore potential solutions from three key perspectives: (1) integrating domain-specific knowledge, (2) performing general UQ for generative AI, and (3) leveraging complementary AI methodologies, such as reinforcement learning, planning, and symbolic reasoning, to improve reasoning and decision-making.

Incorporating domain knowledge

Domain knowledge can be incorporated in generative AI for science in both approaches: foundation models trained from scratch, and pre-trained LLMs/VLMs, as described in the following.

Foundation models from scratch. One promising direction when building foundation models is to integrate scientific knowledge during model selection to constrain the hypothesis space to be consistent with established physical theories or known relationships to improve the model's robustness and generalizability. While the incorporation of knowledge benefits general AI models, it is particularly crucial for large generative AI models due to (1) their larger parameter space, which increases the risks of hallucination, and (2) the need for interpretability when combining complex multiple data sources for modeling interacting processes. Scientific knowledge can take different forms, including physical equations, causal modular structures, and knowledge graphs. The knowledge can also be infused into different components of generative AI models, including the self-supervised pre-training process, model architecture, and the model adaptation process[42–44].

In particular, designing pre-text tasks could involve minimizing the violation of specific physical laws or partial differential equations, emulating physics-based simulations[26,45], or curating triplet samples in contrastive learning based on the similarity of physical properties (e.g., based on knowledge graphs)[46]. For example, the incorporation of physics-guided loss functions and pre-training via synthetic models has demonstrated data/label efficiency and out-of-sample generalization[47]. Such methods can also be useful for modeling rare classes with sparse data, e.g., BioClip[30] enhances image classification of rare species by leveraging knowledge guidance of taxonomic groupings.

The model architecture could also be enhanced with compositional structures, which can provide transparency for intermediate processes and reflect their causal dependencies[48,49]. Specifically, knowledge-guided model architectures can facilitate the simulation over many intermediate processes and provide estimates of their contributions, which also enhances the model's interpretability[48,49]. Guided by scientific knowledge, we can reduce trainable model components without compromising performance in capturing complex dynamics or adapting to downstream tasks. Hence, the integration of knowledge is also helpful for improving the computational efficiency of complex generative AI models.

Pre-trained large-language/vision foundation models for science. Additionally, scientific knowledge can be incorporated during the post-training stage of LLMs and VLMs through prompt engineering and in-context learning to efficiently adapt the model to the target problem without the need for extensive fine-tuning. For example, the dependencies among multiple intermediate processes and their associated equations can be embedded as rationale within the prompt, which provides contextual guidance to support the inference of the generative AI model[50]. One could also feed observations of water quality measures (e.g., water temperature, nutrient concentration) for a target lake, along with relevant physical equations, into an LLM, to improve the generation of future predictions of water quality for this lake.

Another promising direction is leveraging retrieval-augmented generation (RAG) to enable generative AI models to access external knowledge during inference, and thus make the model more robust and trustworthy[51,52]. By incorporating up-to-date information, e.g., satellite imagery or climate reports, and scientific tools, e.g., physical simulators, RAG ensures that these models remain informed of up-to-date scientific knowledge. This approach significantly enhances the models' ability to provide relevant, accurate, and timely insights in dynamic contexts.

While most existing RAG-based studies focus on retrieving data of the same modality, e.g., retrieving text from other documents to answer questions, scientific problems often involve multi-source data of different modalities to capture diverse complex processes. Hence, there is a significant potential for building a cross-modal RAG to retrieve scientific information of different modalities. For example, one could generate text descriptions of target farmlands with relevant satellite imagery, weather time series, crop management reports, and existing physical crop models. As another example, radiologists could diagnose conditions and write reports combining medical images and historical health records. Such cross-modal retrieval ability allows for the effective utilization of heterogeneous and multi-modal data sources, thus improving the model's robustness and trustworthiness.

New directions for UQ

Effective UQ is essential for building trust in generative AI models, particularly in scientific domains where data is often imbalanced, noisy, or sparse. However, UQ remains challenging for large generative models due to the high cost of training and inference, and the complexity of modifying model architectures. A promising yet understudied direction for UQ in generative AI is deterministic uncertainty-based methods[53,54], which offer an appealing solution by enabling uncertainty estimation through single forward passes. Their computational efficiency makes them particularly attractive for large-scale generative AI models where traditional sampling-based approaches become intractable. The key behind this approach is to combine the complementary strengths of two established paradigms - Gaussian processes (GPs), which excel at UQ but struggle to scale with high-dimensional inputs and large datasets common in scientific applications, and neural networks, which handle complex settings well but provide poor uncertainty estimates. Complementing these methods, conformal prediction[55,56] provides a distribution-free framework for UQ that relies solely on the exchangeability assumption, circumventing the need for specific distributional assumptions about the underlying data or model. Another promising avenue that has been explored in recent work[57–59] involves using the language model itself to assess its own uncertainty through direct verbalization (for example, by fine tuning the language models to verbally provide confidence). While such verbalizations may seem naive, they hint at a broader opportunity: framing UQ as a learnable task for LLMs themselves[60], potentially enabling models to express calibrated uncertainty in a more human-interpretable form[60].

Sample-efficient experimental design with Bayesian principles

Bayesian experimental design[61] principles offer valuable frameworks to augment deep generative models with sample-efficient decision-making policies[62,63]. These policies can quantify the information value of candidate experiments for discovering scientific objects (e.g., molecules, materials, proteins, hardware design, 3D printing designs)[64]. We can leverage recent advances in this field that have been developed for various design goals like multiobjective optimization[65], integration of multi-fidelity evaluations[66], and level set estimation[67]. These developments can help develop strategies for general tasks of active data acquisition and targeted generation, potentially enabling broad access to optimization and design procedures across science and engineering while reducing costs, freeing up human time, and accelerating scientific progress.

Augmenting generative AI with complementary AI paradigms

Several promising directions exist for enhancing the decision-making and symbolic reasoning capabilities of generative AI systems by focusing on integrating complementary strengths from other general methods/frameworks in AI.

Related to the decision-making aspect, recent approaches demonstrate how tree search algorithms can be effectively combined with language model prompting techniques, like chain-of-thought prompting[68], to enable better exploration of solution spaces[69]. Looking forward, methods from planning and reinforcement learning research[70,71] could inspire mechanisms for enabling models to learn how to break down complex problems into manageable components. Active exploration is a crucial aspect of reinforcement learning methods because it allows uncovering new information that can lead to better overall strategies/outcomes in the longer term. Advances in UQ for deep generative models can directly support this exploration process by sampling in regions where its knowledge is lacking (i.e., high uncertainty).

Indeed, recent releases of OpenAI O1/DeepSeek R1 models have shown the potential of leveraging reinforcement learning approaches to domains such as code generation and mathematical problem solving. However, both these domains are unique in providing relatively cheaper in-silico verification for candidate solutions. In contrast, many natural science domains present significantly higher evaluation challenges due to resource-intensive requirements (e.g., wet-lab experiments or expert assessment).

Moreover, incorporating symbolic methods into the generative AI architectures (for instance, neurosymbolic constrained sampling[72] and reasoning) could enhance the logical consistency of these approaches by enforcing explicit rules and constraints. For example, Symbol-LLM[73] utilizes an LLM (GPT-3.5) to extract all relevant symbols relating to human activity labels and uses the same LLM as a reasoning proxy for effective symbol composition and construction rules. By leveraging these complementary AI paradigms, we can equip generative models with more robust and interpretable decision-making abilities and enhanced capacity for complex problem-solving.

Building the next-generation ecosystem for generative AI in science

Just as breakthroughs in computer vision and language modeling have historically driven major advances in AI—including the recent surge in generative AI—we now need next-generation ecosystems that enable the AI community to engage more deeply with scientific applications. These ecosystems should be designed to advance generative AI in ways that are informed and driven by the needs of the scientific community. This section highlights key components of such an ecosystem: (i) Access to AI-ready Benchmark Datasets in Science, (ii) Standardizing Evaluation Metrics and Protocols for Generative AI in Science, (iii) Cyber-infrastructure Support to Build and Deploy Generative AI solutions in Scientific Applications, (iv) Trans-disciplinary Training and Engagement, and (v) Trans-disciplinary Collaborations between AI Researchers and Domain Scientists. Some of the discussions in this section build upon the generic recommendations presented in the 2023 Workshop Report on AI in Science[2]. Here, we recontextualize these recommendations in light of recent developments in generative AI, with an emphasis on actionable priorities for the scientific community.

Building AI-ready benchmark datasets

One of the key enablers for making progress in generative AI for science is having access to large-scale AI-ready datasets in scientific domains that allow pre-training and fine-tuning of scientific foundation models, as possible on Internet-scale data in language and vision domains. Despite the excitement in the scientific community for generative AI, publicly accessible AI-ready datasets in science remain scarce for the following reasons. First, scientific processes often involve complex systems characterized by numerous interdependent observations from different modalities, scales, and qualities that serve as proxies for the underlying phenomena being studied. This makes it challenging to standardize the set of observations needed for modeling complex scientific systems. Second, many scientific variables need to be sourced from disparate origins, such as simulations, laboratory experiments, or sensor measurements, requiring extraction from unstructured data formats such as lab notes, figures, or tables in literature, which are sometimes also inaccessible due to proprietary restrictions. This results in large collections of small datasets, where data are fragmented and insufficiently standardized for broader use. Harmonizing such data from multiple sources introduces further complexities, including variability in copyright restrictions, public versus private access, and inconsistent

benchmarking standards across laboratories. Third, there is reluctance within the scientific community to share datasets due to the significant effort involved in their creation and the preference for publishing results before releasing data. Scientists want to safeguard their datasets so that they do not become mere data-providers to AI scientists. While this reluctance is beginning to diminish with funding agencies mandating public data access, we need more coordinated efforts to promote data sharing, establish common benchmarks, and incentivize the development of comprehensive, high-quality, and fully-attributable datasets tailored to scientific applications. We also need protocols and incentive mechanisms for data sharing and attribution in collaborative setups, as it is difficult to build AI-ready datasets in science by individual researchers.

Standardizing evaluation metrics and protocols

We need evaluation ecosystems for generative AI that motivate researchers to think out of the box while assessing the impact of AI solutions on real-world scientific and societally relevant applications. A key ingredient for creating such an ecosystem is defining standardized metrics of performance of generative AI in science that prioritize objectivity, comparability, and efficiency, while accommodating the unique needs of scientific problems. In particular, the complexity of scientific problems—often involving numerous interacting processes at different spatial and temporal scales—renders it difficult to encapsulate their essence within a single metric. For instance, evaluating the synthesizability of chemical materials requires multiple criteria, such as molecular stability and safety considerations, rather than a singular measure. Incorporating collections of disparate metrics may better capture the multidimensional nature of scientific problems, but synthesizing these into a coherent evaluation framework is non-trivial. Traditional metrics such as mean squared error (MSE), commonly used for predictive modeling tasks, often fall short in capturing the nuances of generative tasks. This is especially true in scientific domains that require higher standards of accuracy compared to mainstream AI applications, with varying severity of errors in different contexts. For example, slight spatial or temporal variations in scientific outputs can lead to disproportionately large errors under standard error measures such as MSE, further complicating evaluation. Metrics must also be tailored to the specific domain, as generic measures rarely capture the complexity of scientific progress. In particular, we need specialized metrics to assess the quality of generated samples in terms of their fidelity to observed data and their consistency or coherence with known scientific principles. Moreover, the emerging use of LLMs to evaluate other generative AI models introduces risks, including biases and the challenge of quantifying uncertainty reliably. While LLMs may serve as potential evaluators, their capacity to assess the performance of generative models objectively and consistently remains an open question.

Another key ingredient in building an evaluation ecosystem for generative AI in science is developing evaluation protocols that ensure reproducibility and robust validation. While benchmarking serves as a necessary first step to validate models on known datasets and tasks, the true value of generative models lies in their application to novel, unseen scenarios that lead to new scientific discoveries. Task-specific evaluation processes and domain-specific benchmarks will be vital for advancing generative modeling and enabling AI to contribute effectively to scientific discovery. As an example, benchmarks like ImageNet and AlphaFold have historically been critical in driving major AI advancements by providing well-defined evaluation metrics. However, in many scientific applications, we need to go beyond benchmark-driven research to analyze the potential of a generative AI model in advancing scientific knowledge. For instance, a model generating biologically plausible images of organisms may require evaluation not only through standard metrics like classification accuracy and Frechet inception distance but also via domain-specific insights, such as changes in the traits of generated images through biologically-informed changes in the conditioning vectors of generative AI models, leading to novel hypotheses about species evolution[74,75]. Scientific understanding and explainability of generative AI models are essential for advancing the field, yet these qualities are inherently difficult to quantify.

Building AI-ready cyber-infrastructure

We need to continue investing in cyber-infrastructure to scale the power of generative AI models to their fullest limit in science. Initiatives like the National AI Research Resources program and RAISE (Research in AI for Science and Engineering) provide opportunities to democratize the use of AI and integrate AI into scientific discovery and engineering pipelines. The OpenKIM model curation framework and ColabFit also offer platforms for managing, validating, and sharing models, ensuring transparency and reproducibility in AI-driven research. Additionally, developing new platforms with comprehensive resources—such as curated datasets, model repositories, and scalable computational infrastructure—enables researchers from diverse fields to collaborate effectively. These efforts collectively form the backbone of an AI-ready ecosystem, equipping scientists and engineers with the tools and infrastructure needed to advance their work and accelerate breakthroughs.

Trans-disciplinary training and engagement

We need a strong emphasis on interdisciplinary education and training for generative AI in science. We need to educate students and professionals across scientific backgrounds about the opportunities and risks of using generative AI in science to ensure safe and responsible use of AI in the future and to meet growing workforce demands. Additionally, AI experts need sufficient exposure to scientific practices to collaborate effectively with domain scientists. Progress in generative AI for science requires crossing disciplinary boundaries between experimentalists, model developers, and AI researchers, bridging scientists across these fields. We thus need educational programs that enable such bridges while rethinking computational science and AI curricula to integrate AI tools and foundational knowledge of the scientific discovery process, respectively.

Promoting, nurturing, and sustaining collaboration across disciplinary boundaries

Sustained transdisciplinary collaboration between AI and domain sciences is essential for advancing the role of generative AI in scientific discovery. These collaborations enable experts at the intersection of AI and scientific disciplines to share knowledge, learn from one another, and co-develop solutions. Partnerships between AI researchers and domain scientists are especially critical for translating innovations in generative modeling into meaningful scientific contributions with real-world impact. In addition to bridging academic disciplines, we must seek creative ways to strengthen collaboration across academia, industry, and government labs. Each sector brings unique strengths: industry offers access to proprietary data and scalable infrastructure; government labs are mission-driven and focused on long-term societal challenges; and academia plays a key role in foundational research and training the next generation of scientists.

A noteworthy program from NSF is the Ideas Labs program that provides opportunities for researchers to find collaborators and grow their network. More resources and structural support are needed to cultivate the science of team science that encourages researchers to think outside their disciplinary silos and foster trust in interdisciplinary teams with fully articulated goals and mutual milestones.

Data availability

No datasets were generated or analysed during the current study.

Received: 31 May 2025; Accepted: 21 June 2025;
Published online: 11 August 2025

References

1. NSF Sponsored Workshop on AI-Enabled Scientific Revolution 2023. https://sites.google.com/umn.edu/nsfaiworkshop2023/home. (2023).
2. Kumar, V. et al. *Report on NSF Sponsored Workshop on AI-Enabled Scientific Revolution* https://drive.google.com/file/d/1lU5ZNTRlm9gmgvSlUQGp1aEzqalQXDeQ/view (2023).

3. Bommasani, R et al. On the opportunities and risks of foundation models. *arXiv* preprint arXiv:2108.07258 (2021).
4. Zhou, C. et al. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. In *International Journal of Machine Learning and Cybernetics.* 1–65 (2024).
5. Achiam, J. et al. Gpt-4 technical report. *arXiv* preprint arXiv:2303.08774 (2023).
6. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
7. Reddy, C. K. & Shojaee, P. Towards scientific discovery with generative AI: progress, opportunities, and challenges. *Proc. AAAI Conf. Artif. Intell.* **39**, 28601–28609 (2025).
8. AI4Science, Microsoft Research, and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv* preprint arXiv:2311.07361 (2023).
9. https://sites.google.com/umn.edu/2024ai4sc Second NSF Sponsored Workshop on AI-Enabled Scientific Revolution (2024).
10. Jesson, A. et al. Estimating the hallucination rate of generative AI. *Adv. Neural Inf. Process. Syst.* **37**, 31154–31201 (2024).
11. Bang, Y. et al. HalluLens: LLM Hallucination Benchmark. In *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 24128–24156 (Association for Computational Linguistics, Vienna, Austria, 2025).
12. Gui, J. et al. A survey on self-supervised learning: algorithms, applications, and future trends. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
13. Vaswani, et al. Attention is all you need. In *Advances in neural information processing systems,* 30 (2017).
14. Yang, J. et al. Harnessing the power of llms in practice: a survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data* **18**, 1–32 (2024).
15. Jiang, J., Wang, F.,Shen, J., Kim, S. & Kim, S. A Survey on Large Language Models for Code Generation. *ACM Trans. Softw. Eng. Methodol* https://doi.org/10.1145/3747588 (2025).
16. Yang, L. et al. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**, 1–39 (2023).
17. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
18. Zheng, Y. et al. Large language models in drug discovery and development: from disease mechanisms to clinical trials. *arXiv* **2409**, 04481 (2024).
19. Zhang, X. et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *Found. Trends® Mach. Learn.* **18**, 385–912 (2025).
20. Irrgang, C. et al. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nat. Mach. Intell.* **3**, 667–674 (2021).
21. Zhang, J., Huang, J. Jin, S. & Lu, S. Vision-language models for vision tasks: a survey. In *IEEE Transactions On Pattern Analysis And Machine Intelligence* (2024).
22. Dong, Q. et al. A survey on in-context learning. *arXiv* arXiv:2301.00234 (2022).
23. Yang, K. et al. Leandojo: theorem proving with retrieval-augmented language models. *Adv. Neural Inf. Process. Syst.* **36**, 21573–21612 (2023).
24. Maruf, M. et al. Vlm4bio: a benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *Adv. Neural Inf. Process. Syst.* **37**, 131035–131071 (2024).
25. Li, T. et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *NPJ Digit. Med.* **7**, 40 (2024).
26. Yu, R. et al. Physics-guided foundation model for scientific discovery: an application to aquatic science. *Proc. AAAI Conf. Artif. Intell*. **39**, 28548–28556 (2025).
27. Kurth, T. et al. Fourcastnet: accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11 (2023).
28. Fedik, N. et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat. Rev. Chem.* **6**, 653–672 (2022).
29. Lee, K. L. et al. Towards foundation models for materials science: the open matsci ml toolkit. In *Proceedings of the SC'23 Workshops of the International Conference on High Performance Computing*, Network, Storage, and Analysis, pp. 51–59. (2023).
30. Stevens, S. et al. Bioclip: a vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19412–19424. (2024).
31. Yamada, K. & Hamada, M. Prediction of RNA–protein interactions using a nucleotide language model. *Bioinform. Adv.* **2**, vbac023 (2022).
32. Tan, C. et al. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. *arXiv* preprint arXiv:2312.17016 (2023).
33. Shi, Y. et al. Open-transmind: a new baseline and benchmark for 1st foundation model challenge of intelligent transportation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6328–6335 (2023).
34. Ravirathinam, Khandelwal, R., Ghosh, R. & Kumar, V. Towards a knowledge guided multimodal foundation model for spatio-temporal remote sensing applications. *arXiv* preprint arXiv:2407.19660 (2024).
35. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
36. Requeima, J., John Bronskill, D. C., Richard E. T. & D. Duvenaud. LLM processes: numerical predictive distributions conditioned on natural language. *NeurIPS* (2024).
37. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
38. Sullivan, T. J. "Introduction to uncertainty quantification". pp 63. (Springer, 2015).
39. Kuhn, L., Gal, Y. & Farquhar S. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation. In *ICLR* (2023).
40. Asai, A. et al. Openscholar: synthesizing scientific literature with retrieval-augmented lms. *arXiv* **2411**, 14199 (2024).
41. Kovačević, L. et al. No foundations without foundations-why semi-mechanistic models are essential for regulatory biology. *arXiv* **2501**, 19178 (2025).
42. Karpatne, A. et al. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. data Eng.* **29**, 2318–2331 (2017).
43. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **55**, 1–37 (2022).
44. Karpatne, A., Jia, X. & V. Kumar, V. Knowledge-guided machine learning: current trends and future prospects. *arXiv* preprint arXiv:2403.15989 (2024).
45. Jia, X. et al. Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM/IMS Trans. Data Sci.* **2**, 1–26 (2021).
46. Wan, Y., Wu, J., Hou, T., Hsieh, C.-Y. & Jia, X. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nat. Commun.* **16**, 413 (2025).
47. Read, J. S. et al. Process-guided deep learning predictions of lake water temperature. *Water Resour. Res.* **55**, 9173–9190 (2019).
48. Liu et al. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nat. Commun.* **15**, 357 (2024).
49. Khandelwal, A. et al. Physics guided machine learning methods for hydrology. *arXiv* preprint arXiv:2012.02854 (2020).
50. Pan, H. et al. Quantum many-body physics calculations with large language models. *Commun. Phys.* **8**, 49 (2025).

51. Prince, M. H. et al. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Comput. Mater.* **10**, 251 (2024).
52. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. "Benchmarking retrieval-augmented generation for medicine". The ACL Findings, 2024.
53. Liu, J. et al. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness *In* NeurIPS 7498–7512 (2020).
54. Van Amersfoort, J., Smith, L., Jesson, A., Key, O. & Y. Gal, Y. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv* preprint arXiv:2102.11409 (2021).
55. Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J. & Jordan, M. I. Conformal prediction under feedback covariate shift for biomolecular design. *Proc. Natl Acad. Sci.* **119**, e2204569119 (2022).
56. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
57. Lin, S., Hilton, J. & Evans, O. Teaching models to express their uncertainty in words. *arXiv* preprint arXiv:2205.14334 (2022).
58. Johnson, D. D., Tarlow, D., Duvenaud,D., & C. J. Maddison experts don't cheat: learning what you don't know by predicting pairs. In ICML (2024).
59. Kadavath, S. et al. Language models (mostly) know what they know. *arXiv* preprint arXiv:2207.05221 (2022).
60. Kapoor, S. et al. Calibration-tuning: teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP* (UncertaiNLP 2024), pp. 1–14. (2024).
61. Garnett, R. "Bayesian optimization". Cambridge University Press (2023).
62. Deshwal, A. & Doppa, J. Combining latent space and structured kernels for Bayesian optimization over combinatorial spaces. *Adv. neural Inf. Process. Syst.* **34**, 8185–8200 (2021).
63. Maus, N. et al. Local latent space bayesian optimization over structured inputs. *Adv. neural Inf. Process. Syst.* **35**, 34505–34518 (2022).
64. Deshwal, A., Cory, M. S. & Janardhan Rao Doppa. Bayesian optimization of nanoporous materials. *Mol. Syst. Des. Eng.* **6**, 1066–1086 (2021).
65. Belakaria, S., Deshwal, A. & Rao Doppa, J. Max-value entropy search for multi-objective Bayesian optimization. In NeurIPS (2019).
66. Takeno, S. et al. Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization. In ICML (2020).
67. Neiswanger, W., K. Alexander Wang, K. & Ermon, E. Bayesian algorithm execution: estimating computable properties of black-box functions using mutual information. In ICML (2021).
68. Hoffman, M. D. et al. Training chain-of-thought via latent-variable inference. In *NeurIPS* https://neurips.cc/virtual/2023/poster/71210 (2024).
69. Yao, S. et al. Tree of thoughts: deliberate problem solving with large language models. *NeurIPS*, **517**, 11809–11822 (2024).
70. Zelikman, E., Wu, Y., Mu, J. & Goodman, N.star: bootstrapping reasoning with reasoning. In *NeurIPS*, 15476–15488 (2022).
71. Kumar, A. et al. Training language models to self-correct via reinforcement learning. *arXiv* **2409**, 12917 (2024).
72. Zhang, H., Kung, P.-N., Yoshida, M., Broeck, G V den. & Peng, N. Adaptable logical control for large language models. In *NeurIPS,* 115563–115587 (2024).
73. X. Wu, Y.-L. Li, Sun, J. & Lu C. "Symbol-LLM: leverage language models for symbolic system in visual human activity reasoning". In *Advances in Neural Information Processing Systems*, 36 (2024).
74. Elhamod, M. et al. Discovering novel biological traits from images using phylogeny-guided neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3966–3978 (2023).
75. Khurana, M. et al. Hierarchical conditioning of diffusion models using tree-of-life for studying species evolution. In *European Conference on Computer Vision*, pp. 137–153 (Cham: Springer Nature Switzerland, 2024).

Acknowledgements

We would like to acknowledge all workshop participants listed in Appendix A who contributed to the workshop discussions and in preparation of the report. This work was supported in part by NSF Award #2309660.

Author contributions

A.K.,A.D., X.J., and V.K. designed the structure of the article and prepared the first draft of the report. W.D.,M.S., and A.Z. provided feedback on the draft and made edits to the manuscript. A.K., A.D., X.J., M.S., and W.D. took copious notes during the meeting to capture the discussions. The authors iterated with the workshop participants to incorporate their feedback into the manuscript. V.K. provided overall leadership in the preparation of the manuscript. A.K., A.D., and X.J. contributed equally to this work as joint first authors. V.K., and A.Z. served as the workshop co-organizers.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00018-6.

Correspondence and requests for materials should be addressed to Vipin Kumar.