

Foundation Models for Scientific Discovery: From Paradigm Enhancement to Paradigm Transition

**Fan Liu¹, Jindong Han², Tengfei Lyu¹, Weijia Zhang¹, Zhe-Rui Yang¹,
Lu Dai^{2,1}, Cancheng Liu¹, Hao Liu^{1,2}**

¹The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

²The Hong Kong University of Science and Technology, Hong Kong SAR, China

fliu236@connect.hkust-gz.edu.cn; hanjindong01@gmail.com;
tlyu077@connect.hkust-gz.edu.cn; vegazhang3@gmail.com;
lidaiae@connect.ust.hk; liuh@ust.hk

Abstract

Foundation models (FMs), such as GPT-4 and AlphaFold, are reshaping the landscape of scientific research. Beyond accelerating tasks such as hypothesis generation, experimental design, and result interpretation, they prompt a more fundamental question: Are FMs merely enhancing existing scientific methodologies, or are they redefining the way science is conducted? In this paper, we argue that FMs are catalyzing a transition toward a new scientific paradigm. We introduce a three-stage framework to describe this evolution: (1) Meta-Scientific Integration, where FMs enhance workflows within traditional paradigms; (2) Hybrid Human-AI Co-Creation, where FMs become active collaborators in problem formulation, reasoning, and discovery; and (3) Autonomous Scientific Discovery, where FMs operate as independent agents capable of generating new scientific knowledge with minimal human intervention. Through this lens, we review current applications and emerging capabilities of FMs across existing scientific paradigms. We further identify risks and future directions for FM-enabled scientific discovery. This position paper aims to support the scientific community in understanding the transformative role of FMs and to foster reflection on the future of scientific discovery. Our project is available at <https://github.com/usail-hkust/Awesome-Foundation-Models-for-Scientific-Discovery>.

1 Introduction

Scientific discovery has historically progressed through a series of methodological paradigms, each redefining how researchers observe, explain, and model the natural world. From the empirical investigations of Galileo and Boyle to the formal theories of Newton and Einstein, science has advanced by transforming observations into abstract and systematic knowledge. Later, computational simulation enabled the exploration of systems too complex for direct experimentation, while the rise of data-driven science in the 21st century reframed discovery as the extraction of statistical patterns from massive datasets. Together, these four paradigms, *i.e.*, experimental, theoretical, computational, and data-driven, constitute the foundations of modern scientific practice [1, 2, 3].

However, as science increasingly engages with phenomena characterized by emergent behavior, open-endedness, and irreducible complexity, the limitations of existing paradigms have become more evident [4, 5, 6]. Challenges such as understanding consciousness, modeling protein folding pathways, and predicting social polarization defy reductionist modeling and remain intractable, even in the face of recent advances in machine learning [7, 8]. In fields like drug discovery and material design, the combinatorial explosion of candidate spaces makes exhaustive search infeasible [9]. Meanwhile, the

Correspondence to Hao Liu.

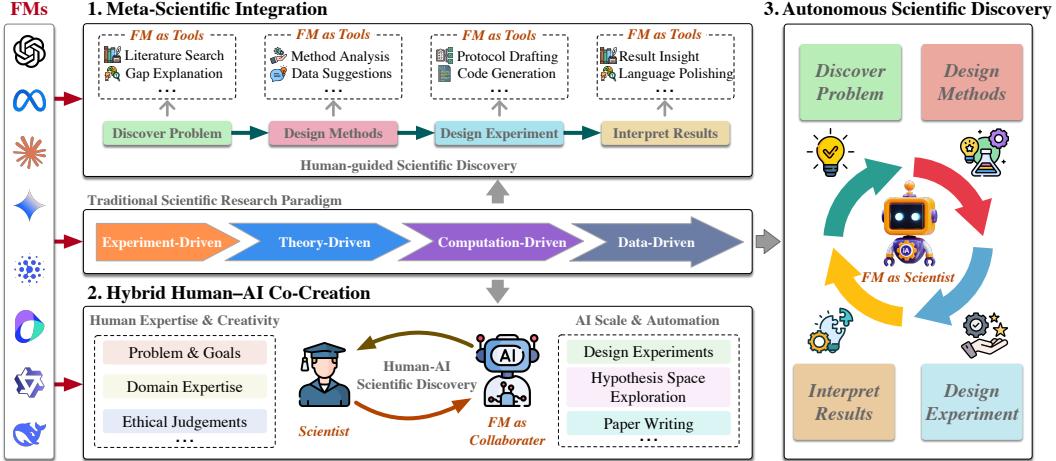


Figure 1: Evolving scientific paradigms empowered by FMs. FMs progressively transition from tool-like infrastructure (meta-scientific integration), to interactive co-creators (hybrid human–AI collaboration), and ultimately to autonomous agents capable of end-to-end scientific discovery.

rapid accumulation of experimental and observational data has outpaced our capacity to synthesize unifying theories or explanatory frameworks, widening the gap between empirical richness and conceptual understanding. Even state-of-the-art computational models often rely on simplifying assumptions such as linearity, stationarity, or equilibrium, which are fundamentally misaligned with the dynamic, non-linear, and adaptive nature of many real-world systems [10, 11]. These tensions underscore a growing mismatch between the increasing complexity of scientific problems and the methodological frameworks currently available to address them.

Foundation Models (FMs) [12] offer a promising response to these challenges. As large-scale neural networks trained on diverse and extensive datasets, FMs exhibit remarkable adaptability, performing a wide range of tasks via prompting or fine-tuning. Models such as GPT-4 [13], AlphaFold [4], and DeepSeek [14] have demonstrated unprecedented capabilities in language understanding, code generation, and scientific reasoning. For instance, AlphaFold [4] resolved the long-standing protein folding challenge by navigating an intractable configuration space using learned structural priors. FunSearch [15], developed by DeepMind, goes even further, showing that FMs can autonomously propose and validate new mathematical conjectures, rivaling expert-designed algorithms on NP-hard problems. These advances reflect a broader trend: FMs not only accelerate existing scientific workflows, but also begin to change how knowledge is generated, organized, and applied. Unlike previous AI systems that were built for specific tasks, FMs offer a unified architecture capable of handling text, code, and even multi-modal inputs. More importantly, they support new ways of thinking, enabling reasoning, abstraction, and exploration at scale. In this sense, FMs blur the boundary between tool and collaborator, between algorithmic processing and cognitive engagement. This brings us to a critical question: Are foundation models simply enhancing the current scientific paradigm, or are they catalyzing the emergence of a new one?

Throughout history, paradigm shifts in science have not only introduced new tools but also transformed the way science is understood and practiced. Transitions from observation to explanation, or from simulation to data-driven inference, have introduced new epistemologies for formulating problems, generating evidence, and establishing scientific validity. Today, FMs may represent a similar inflection point. By unifying language, code, and multimodal inputs within a single framework, FMs can retrieve literature, formulate hypotheses, simulate complex phenomena, interpret results, and even coordinate end-to-end research workflows. Supporters argue that FMs are reshaping the structure of scientific discovery by lowering entry barriers, facilitating exploratory iteration, and redistributing agency between humans and machines [16, 17]. Skeptics, however, view FMs as powerful yet conventional tools that amplify existing methodologies [18, 19]. From this perspective, FMs serve to accelerate scientific progress without fundamentally transforming its underlying paradigm.

This paper enters that debate with a clear position: **FMs are not only improving parts of the scientific process, they are beginning to reshape the paradigm through which science operates.** To support this argument, we propose a three-stage framework that describes the progressive integration

of FMs into scientific discovery, as illustrated in Figure 1. (1) *Meta-Scientific Integration*. FMs operate as flexible infrastructure across traditional scientific paradigms. They integrate cross-domain data, automate reasoning steps, and support end-to-end workflows, while remaining embedded within established epistemic structures. (2) *Hybrid Human-AI Co-Creation*. In this transitional phase, FMs shift from passive tools to active collaborators. They participate in problem formulation, hypothesis generation, and experimental design, enabling more dynamic, iterative, and co-creative modes of discovery. (3) *Autonomous Scientific Discovery*. Looking ahead, we envision FMs acting as autonomous agents of science. These systems will be capable of initiating questions, executing simulations, interpreting results, and generating new knowledge across both virtual and physical domains. At this stage, the scientific process becomes partially self-directed, presenting a shift toward a fundamentally new epistemic regime.

Our Contributions: (1) *A new conceptual framework*. We introduce a three-stage framework to position FMs as catalysts of scientific paradigm evolution, spanning infrastructure support, collaborative reasoning, and autonomous discovery. (2) *A systematic review and taxonomy*. We present a systematic analysis of FM-enabled scientific discovery, organized by their integration into experimental, theoretical, computational, and data-driven workflows. (3) *A research agenda*. We identify key risks that should be addressed to realize the full scientific potential of FMs, and we also outline directions for future research on aligning epistemic goals with emerging AI capabilities.

2 Background and Preliminary

A Brief History of Scientific Discovery. Scientific discovery has advanced through a series of methodological paradigms, each reshaping the ways we observe, explain, and intervene in the natural world, as illustrated in Figure 2. These paradigms emerged in response to the limitations of their predecessors and were enabled by conceptual breakthroughs or technological innovations. Specifically, today’s scientific practice is shaped by four foundational paradigms: experiment-driven, theory-driven, computation-driven, and data-driven science. Each has introduced new standards for reasoning, validation, and knowledge generation.

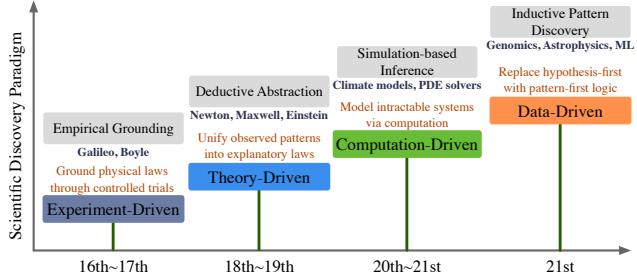


Figure 2: A roadmap of scientific discovery paradigms and their epistemic capabilities.

- (1) The *experiment-driven paradigm* arose during the scientific revolution of the 16th and 17th century, emphasizing systematic observation and controlled experimentation [1]. Pioneers such as Galileo and Boyle designed repeatable experiments to validate natural laws [20, 21], establishing measurability, verifiability, and reproducibility as empirical norms. However, this approach struggled with large-scale, highly complex, or inaccessible systems, where direct manipulation was impractical or impossible.
- (2) The *theory-driven paradigm* emerged in the 18th and 19th centuries, driven by advances in mathematics and formal logic [22]. Newton, Maxwell, and Einstein proposed abstract, unified theories that explained a wide range of phenomena under compact formulations [23, 24]. While these models offered significantly expanded explanatory power, they also introduced a growing gap between theoretical complexity and empirical testability.
- (3) The *computation-driven paradigm* gained traction in the mid-20th century with the advent of digital computing and numerical simulation [2, 25]. It enabled scientists to model systems that were either analytically intractable or experimentally inaccessible, such as global climate or molecular interactions. This gave rise to new forms of scientific reasoning, including scenario testing, model-based inference, and virtual experimentation. More recently, hybrid methods like physics-informed machine learning have further blurred the boundaries between theory, simulation, and data [26].
- (4) The *data-driven paradigm* became prevalent in the 21st century, fueled by exponential growth in sensing technologies, digitization, and computational power. This paradigm focuses on discovering patterns in high-dimensional data using statistical and machine learning

techniques [3, 27]. Applications span diverse fields from genomics to astrophysics [28, 29], enabling data-driven insights in domains where prior models were lacking or underdeveloped. Despite its success, data-centric science often struggles with causal inference, interpretability, and robustness under distributional shifts [30, 11, 31].

The above four paradigms have progressively expanded the scope and scale of scientific discovery. However, as scientific problems become increasingly complex and interdisciplinary, the limitations of each paradigm, particularly when applied in isolation, become more apparent. This calls for rethinking how scientific discovery might evolve beyond the current frameworks.

Foundation Models. FMs are large-scale neural networks trained on massive and diverse datasets, designed to serve as general-purpose systems adaptable to a wide range of downstream tasks [32, 14, 33, 34, 35, 13, 36]. They are typically developed through unsupervised or self-supervised pertaining, for example, by predicting masked tokens in text or aligning image–captions pairs, followed by task-specific fine-tuning or prompting. This pretrain and fine-tune paradigm allows knowledge acquired during large-scale training to transfer across domains and tasks. The emergence of FMs marks a shift from narrow, task-specific models to flexible systems capable of generalizing across modalities and problem types. This is particularly valuable in scientific domains, where labeled data is scarce, tasks are often open-ended, and disciplinary boundaries are increasingly fluid. FMs provide a unified modeling framework that integrates language, vision, code, and structured data, enabling diverse tasks in reasoning, generation, and retrieval. Prominent examples include GPT-4 [13], a language model that performs a wide range of tasks, including question answering, summarization, and code generation, through zero-shot or few-shot prompting. Another is CLIP [34], a vision–language model trained on 400 million image–text pairs using a contrastive learning objective. Without additional fine-tuning, CLIP can classify images based on natural language prompts, demonstrating strong zero-shot capabilities. More recently, FMs have expanded into scientific domains such as protein folding (e.g., AlphaFold [4]), mathematical discovery (e.g., FunSearch [15]), and robotics [37]. These applications highlight the growing role of FMs not merely as tools for automation but as general-purpose engines for scientific reasoning, interpretation, and discovery.

3 Rethinking Scientific Paradigms in the Era of Foundation Models

In this section, we introduce a three-stage framework to describe the evolving role of FMs in scientific discovery and to explore their potential for paradigm-level change based on the degree of autonomy, task scope, as illustrated in Table 1. We argue that FMs are not only enhancing individual components of the scientific process but are also beginning to reshape the broader structure of scientific discovery. Although this transformation is still in its early stages, the shift is already underway.

Table 1: Comparison of FM Paradigms Across Five Dimensions

Dimension	Meta-Scientific Integration	Hybrid Human–AI Co-Creation	Autonomous Scientific Discovery
Paradigm Definition	Tool	Human–AI collaborator	Independent agent
FM Role	Backend tool	Co-creator	Autonomous actor
Task Scope	Task enhancer	Full-cycle tasks	End-to-end, self-directed
Autonomy	Low	Moderate	High
Impact on Science	Efficiency boost	Labor shift	Scientific re-foundation

Meta-Scientific Integration. In the meta-scientific integration mechanism, FMs function as an intelligent infrastructure that augments, but does not transform, scientific practice. Their core role lies in streamlining fragmented processes, enhancing interoperability, and increasing the reproducibility and efficiency of workflows across disciplinary boundaries. FMs in this paradigm serve as backend coordinators: they automate procedural tasks such as data preprocessing, literature retrieval, and methodology matching. By integrating components that were once isolated (e.g., linking sensor data with simulation models or connecting experimental planning with prior knowledge), FMs facilitate smoother, modular research pipelines. However, these systems remain fully bound by human-defined objectives and lack the capacity to initiate or reframe scientific inquiry. Crucially, the role of FMs here is instrumental, not epistemic. They execute tasks within established scientific paradigms without altering their logic or structure. Despite their technical sophistication, they exhibit low autonomy and require continuous human supervision. This paradigm is therefore augmentative rather than transformative: it improves how science is conducted but does not redefine what science

fundamentally is. FMs increase scientific throughput and integration, but the locus of reasoning and knowledge production remains firmly human.

Hybrid Human-AI Co-Creation. FMs are shifting from passive infrastructure to active collaborators within scientific workflows. Rather than remaining in the background, they now work alongside human researchers in shared reasoning and decision-making processes, enabling a hybrid intelligence model that pairs human intuition and expertise with the generalization, memory, and automation capabilities of FMs. In this new role, FMs assist across the scientific pipeline, contributing to research question generation, hypothesis structuring, experiment planning, and, in some cases, end-to-end task execution. Their involvement extends beyond operational support to ideation and interpretation, though they continue to operate within boundaries set by human oversight and scientific norms. The scope of FM contributions has expanded significantly. They engage in hypothesis generation, problem scoping, experiment design, execution support through automation, interpretation of findings, and participation in scientific discourse. While their functions span the full research cycle, their actions are still initiated, constrained, and validated by humans. FMs exhibit moderate autonomy: they can generate ideas, select methods, and adapt workflows based on feedback within scoped research environments, but they rely on human prompts for problem framing and ethical guidance. Their outputs can influence the trajectory of discovery but do not fully determine it. This evolving paradigm begins to reshape the division of cognitive labor in science. By offloading tasks such as literature synthesis, multi-step reasoning, and combinatorial experiment planning, FMs allow human researchers to focus more on judgment, creativity, and strategic framing. Although the human-AI co-creation model introduces a new form of epistemic collaboration, it does not yet constitute standalone scientific intelligence. FMs reconfigure how science is conducted without redefining who conducts it.

Autonomous Scientific Discovery. In this emerging paradigm, FMs take a decisive step beyond collaboration, evolving into autonomous agents capable of conducting scientific discovery with minimal or no human oversight. Unlike earlier stages, where FMs assist under predefined goals or structured prompts, autonomous FMs can initiate and carry out the entire scientific cycle on their own, i.e., posing research questions, generating hypotheses, selecting methods, executing experiments or simulations, interpreting results, and updating internal models based on outcomes. What distinguishes this paradigm is the high degree of autonomy and continuity in the reasoning process. FMs are no longer reactive tools triggered by human input. They operate in a self-directed manner, guided by internal objectives and feedback mechanisms. Their behaviors resemble those of scientific investigators: identifying promising research directions, exploring solution spaces, evaluating novelty and coherence, and refining strategies based on intermediate findings. Such agentic behavior enables FMs to function not just as tools, but as epistemic actors that can contribute original insights, challenge existing theories, and shape the direction of scientific discourse. These models possess the capacity to synthesize diverse knowledge sources, bridge conceptual gaps across disciplines, and adapt dynamically to new evidence or goals. They are capable of making decisions about what to explore, how to explore it, and when to revise their understanding, all without explicit instruction. The broader implications are profound. If fully realized, this paradigm would mark a fundamental shift in the structure of scientific discovery. Rather than simply accelerating human-led research, FMs would become independent engines of scientific discovery, redefining who or what can produce scientific knowledge. This shift introduces what we call the fifth scientific paradigm, where discovery is no longer exclusively human-driven but emerges from the autonomous reasoning of machine intelligence. While still in its formative stages, this trajectory is already taking shape in systems like AI scientists [38], which have conducted the whole research pipeline to solve scientific problems. As FMs continue to evolve, their role may expand beyond assistance or collaboration toward initiating, directing, and validating new lines of scientific investigation. This transformation challenges long-held assumptions about the nature of scientific agency and opens new questions about responsibility, trust, and validation in machine-led discovery.

4 Foundation Model Integration Across Scientific Paradigms

Building on our three-stage framework of FM-driven scientific evolution, this section spans the first two stages, Meta-Scientific Integration and the early signs of Hybrid Human-AI Co-Creation, by examining how FMs are increasingly embedded within and across classical scientific paradigms. We systematically review their roles in experimental, theoretical, computational, and data-driven scientific discovery, as well as their emerging capacity to mediate cross-paradigm workflows.

4.1 Experiment-Driven Paradigm

The experiment-driven paradigm emphasizes empirical observation, controlled intervention, and iterative refinement. However, traditional workflows are constrained by limited planning capacity, costly trial spaces, and brittle automation pipelines. FMs offer new opportunities to enhance this paradigm by improving design efficiency and enabling more flexible, adaptive execution. Current integrations focus on two key stages: (1) experiment design and (2) physical experiment execution.

Experimental Design. Designing informative experiments under resource constraints remains a core scientific challenge. Classical methods such as Bayesian optimization (BO) and active learning (AL) often suffer from sparse priors and poor generalization. FMs help overcome these issues by encoding domain knowledge and guiding the search for optimal configurations [39, 40, 41, 42]. For instance, FMs serve as priors or feature extractors in BO pipelines, accelerating convergence in molecular and materials discovery [43, 44]. Building on this, FMs further improve data efficiency by directly maximizing mutual information, bypassing the need for surrogate modeling [45]. These approaches point toward a future in which FMs co-adapt with experimental design processes, forming the backbone of closed-loop, context-aware optimization agents.

Physical Experiment Execution. In laboratory settings, executing experiments demands coordination across planning, perception, and control domains that are traditionally fragmented and manually programmed. FMs increasingly act as unifying interfaces and planners [46, 47, 37, 48, 49]. For example, FMs have been employed to generate Python control scripts for scientific instruments, translating user-specified objectives into directly executable lab protocols [48], while LLM-RDF orchestrates modular agents for structured reaction planning [47]. More dynamic systems, such as CLAIRify, embed FMs into robotic control, enabling physical manipulation through language-guided planning [37]. Pushing further, multimodal agents like VISION and AP-VLM incorporate vision and speech to support real-time interaction and error correction in lab environments [49, 50].

4.2 Theory-Driven Scientific Paradigm

The theory-driven paradigm seeks to construct formal, generalizable frameworks that explain observed phenomena and yield testable predictions. Traditionally dependent on human intuition and symbolic logic, this paradigm has been constrained by limited idea generation, steep formalization requirements, and the brittleness of proof systems. FMs are increasingly being integrated to augment this process by expanding the space of plausible hypotheses and supporting formal reasoning pipelines that validate theoretical claims.

Scientific Hypothesis Generation. FMs facilitate systematic hypothesis generation by synthesizing knowledge across large-scale corpora and structured priors [51, 52, 53, 54]. Rather than relying solely on intuition or fragmented evidence, recent methods guide FMs using knowledge graphs and domain constraints. For example, KG-CoI steers hypothesis formulation through ontological concept paths to enhance novelty and verifiability [52]. The HypoGen dataset further improves model outputs by grounding them in historical patterns of scientific idea evolution, improving both creativity and feasibility [53]. In scientific domains like physics and climate modeling, physics-guided foundation models embed physical laws directly into the generation process to ensure consistency with known dynamics [55, 56].

Theory Validation and Formal Reasoning. To validate hypotheses, FMs are increasingly linked with symbolic logic systems to support deductive inference, consistency checking, and falsifiability analysis [57, 58, 59, 60, 61, 62]. Logic-LM exemplifies this by coupling LLMs with symbolic solvers in a feedback loop, improving formal rigor in logical tasks [57]. General-purpose neuro-symbolic systems like SymbolicAI and Vieira extend this framework to domain-specific reasoning tasks [58, 63]. For automated theory testing, the Popper system uses LLMs to generate counterexamples and identify falsifiable conditions [60]. In formal mathematics, LeanCopilot and DeepSeekProver demonstrate the capacity of pretrained models to assist in proof construction and verification at scale [61, 62].

4.3 Computation-Driven Scientific Paradigm

The computation-driven paradigm advances scientific discovery through the formulation and execution of mathematical models that simulate, predict, or control complex systems. While traditional

workflows depend on hand-crafted equations and high-cost numerical solvers, they often face limitations in flexibility, scalability, and automation. FMs offer new capabilities by enabling automated model construction and accelerating scientific computation. We review recent progress along two key fronts: constructing executable scientific models and efficiently solving or inverting them.

Formulating Executable Scientific Models. Conventional scientific modeling relies on expert-designed equations or symbolic regressors, which struggle with multi-scale dynamics, noisy observations, and sparse priors. FMs enhance this process by supporting symbolic, latent, and differentiable formulations. For instance, in symbolic discovery, systems like LLM-SR translate diverse inputs such as plots or text into equation skeletons for subsequent refinement, while others like FUNSEARCH discover new algorithms by framing program synthesis as a language-guided search task [64, 65, 66]. When explicit equations are elusive, FMs excel at learning latent operators. PROSE-PDE, for example, simultaneously predicts system dynamics and infers underlying governing laws within a learned representation, and DIFFUSIONPDE trains generative priors over coefficient-solution pairs to sample posteriors from sparse data, effectively bypassing direct equation formulation [67, 68]. Such modeling methods unify forward and inverse modeling under shared representations.

Solving and Inverting Scientific Equations. Once scientific models are formulated, whether as partial differential equations or latent operator representations, solving and inverting them remains computationally demanding. Classical methods typically require spatial discretization, expert-crafted solvers, and often fragile optimization routines, especially in ill-posed or high-dimensional settings. FMs operate directly over function spaces, guided by learned priors and generative inference mechanisms, to accelerate solutions and enable efficient inversion. A pivotal development on Neural Operator learns continuous maps from forcing terms to partial differential equation (PDE) solutions, generalizing across mesh resolutions and becoming a cornerstone for physics surrogates [69, 70]. Building on this, models like GRAPHCAST now outperform traditional numerical weather prediction models at reduced computational cost, and specialized architectures like FACTFORMER handle massive computational grids [71, 72]. Furthermore, FMs can enhance legacy solvers; for example, PDE-Refiner architectures iteratively correct coarse solver output, trimming error without rerunning the full simulation [73]. These innovations significantly reduce the computational burden.

4.4 Data-Driven Scientific Paradigm

The data-driven paradigm begins with large-scale observations collected across instruments, populations, and modalities. It aims to discover latent scientific structures and generate predictive outputs directly from data, often without recourse to explicit physical models. Traditional workflows rely on handcrafted features, narrow supervision, and unimodal pipelines, limiting their ability to scale, integrate, or generalize. FMs offer a unified upgrade by learning statistical regularities across domains and enabling flexible reasoning over heterogeneous signals. Current applications cluster into two major directions: (1) scientific knowledge discovery from multimodal data and (2) predictive scientific inference through generative modeling.

Scientific Knowledge Discovery. Classical methods for extracting scientific knowledge, such as enrichment analysis or rule mining, struggle with noisy, multimodal, or unstructured data. FMs address these limitations by compressing vast corpora into structured representations and supporting inference across modalities. For instance, token-based FMs like DNABERT identify functional DNA elements from sequences [74]. In chemistry, MOLFORMER learns SMILES embeddings that correlate linearly with key molecular properties, enabling zero-shot retrieval of candidate molecules [75]. Beyond single modalities, multimodal FMs like CHEMVLM integrates molecular structure images and textual descriptions to answer complex multimodal chemistry questions [76, 77]. In the spatio-temporal domain, CLIMAX[5] fuses diverse climate inputs, spanning reanalysis data, climate model simulations, and satellite observations, learning unified spatio-temporal representations through masked autoencoding. These rich embeddings capture underlying climate patterns and their complex interdependencies, thereby facilitating the discovery of novel insights into Earth system dynamics and the characterization of various climate phenomena. Furthermore, large language models pretrained on vast corpuses of scientific literature, such as GALACTICA, act as powerful tools to organize, synthesize, and query scientific knowledge, effectively transforming millions of papers into an accessible and computationally tractable knowledge base [78, 79].

Predictive Scientific Inference. In many domains, predictive accuracy is now more critical than explicit mechanistic modeling. Classical surrogate models, however, often struggle with high-

dimensionality and uncertainty. FMs redefine this task as generative modeling, trained directly on observational or simulation-derived data. For instance, in spatiotemporal forecasting, GRAPHCAST and PANGU-WEATHER learns latent dynamics from re-analysis datasets to produce global weather predictions rivaling numerical models at lower computational cost [80, 81]. Diffusion-based models like DIFFUSIONSAT can generate high-resolution satellite imagery from coarser inputs, bridging observational gaps [82]. In structural prediction, FMs such as ALPHAFOLD 2 and ESMFOLD predict protein structures from sequences with near-experimental accuracy [83, 84]. Furthermore, generative models like RFDIFFUSION can design novel protein folds and functional interfaces, while MATTERGEN extends the same paradigm to inorganic crystal design, producing stable materials that satisfy user-specified property constraints [85, 86], demonstrating the capacity of FMs to turn data into actionable foresight.

4.5 Cross-Paradigm Foundation Model Integration

Classical scientific paradigms, experimental, theoretical, computational, and data-driven, have historically represented distinct methodological lenses, though they are often employed in combination in scientific practice. As modern scientific challenges grow in complexity, discovery increasingly relies on workflows that integrate these paradigms into unified, cross-cutting pipelines. FMs, with their general-purpose reasoning abilities, multimodal interfaces, and growing autonomy, are uniquely positioned to mediate such integrative, hybrid workflows.

Recent advances show that FMs can serve as integrative engines across classical scientific paradigms, experimental, theoretical, computational, and data-driven, by enabling workflows that traverse and connect traditionally siloed approaches. Crucially, these models maintain interpretability and cross-domain transferability, supporting scientific reasoning that is both coherent and generalizable across diverse methodologies [87, 88, 89, 90, 91]. For example, PROSE-FD [92] co-trains symbolic equation templates and spatial field data within a multimodal Transformer, enabling cross-regime generalization in fluid dynamics and jointly discovering both structure and solution behavior. Similarly, Latent Neural Operators (LNOs) [93] encode physical operators into latent spaces that are geometry-agnostic and resolution-invariant, allowing both forward and inverse problems to be solved within a shared learned representation. Beyond individual modeling components, FMs increasingly orchestrate end-to-end scientific workflows that couple theory, simulation, data, and experimentation. In chemistry, for instance, systems like Coscientist [46] translate high-level research goals into machine-executable protocols, control robotic synthesis, and adapt future actions based on experimental results.

5 Risks of Emerging FM-Centered Scientific Paradigms and Future Direction

In this section, we introduce the key risks posed by emerging FM-centered scientific paradigms, including challenges related to bias, misinformation, reproducibility, and scientific accountability. We then outline future directions toward autonomous scientific discovery, highlighting embodied agents, closed-loop workflows, and continual learning.

5.1 Risks of Emerging FM-driven Scientific Paradigms

While FMs promise transformative benefits across the scientific enterprise, their growing autonomy introduces critical epistemic, technical, and ethical risks. These risks evolve and intensify as FMs transition from backend tools (Meta-Scientific Integration) to collaborative partners (Hybrid Human–AI Co-Creation), and ultimately toward independent research agents (Autonomous Scientific Discovery). We identify four key risk dimensions that require anticipatory mitigation to ensure the responsible evolution of FM-driven scientific paradigms.

Bias and Epistemic Fairness. Even in early-stage applications, such as literature review and task assistance, FMs inherit biases from their training data, which often overrepresent dominant paradigms, Western institutions, and widely cited authors [94, 95]. As FMs transition into co-creators and autonomous agents, these biases shift from being passive reflections to active forces shaping scientific agendas. For instance, in global health modeling, a FM trained predominantly on English-language publications and high-impact journals may systematically prioritize research on diseases like Type 2 diabetes or cardiovascular conditions, topics well-studied in Western contexts, while overlooking pressing but underrepresented issues such as schistosomiasis or child stunting in sub-

Saharan Africa [96, 96]. Without intervention, this can lead to epistemic homogenization and the exclusion of underrepresented perspectives. Mitigating these risks calls for more diverse and inclusive training datasets, targeted fine-tuning on marginalized knowledge domains, and fairness-aware evaluation protocols embedded throughout the FM pipeline.

Hallucination and Scientific Misinformation. Across all paradigms, FMs remain fundamentally data-driven pattern recognizers rather than truth-preserving reasoners. As their role shifts from task augmentation to autonomous hypothesis generation, the risk of generating plausible but unverified claims grows substantially [97, 98]. In biomedical domains, for instance, an FM might propose a novel mechanism that appears convincing but lacks experimental grounding, potentially misguiding research efforts. In physics, it may generate elegant but physically invalid formulations. These failures can propagate if outputs are prematurely trusted or cited. To mitigate this, FMs should incorporate verification mechanisms such as symbolic logic checks, simulation-based validation, human-in-the-loop review, and provenance tracking to ensure traceability and scientific credibility.

Reproducibility and Scientific Transparency. As FMs take on more end-to-end responsibilities, such as designing experiments, running simulations, and interpreting results, their decision-making processes often remain opaque [99, 100]. This threatens scientific reproducibility: without visibility into intermediate reasoning steps, model assumptions, or version states, it becomes difficult to replicate or validate outcomes. For example, a model-generated chemical synthesis pathway may lack interpretable derivations. Addressing this requires transparent logging of reasoning steps, version-controlled model checkpoints, and open-science practices that preserve the traceability of FM-driven scientific workflows.

Authorship, Accountability, and Scientific Ethics. As FMs shift from tools to collaborators and, ultimately, autonomous agents, questions around intellectual credit, accountability, and ethical conduct become increasingly urgent [101]. If an FM generates a core hypothesis or experimental design, should it be acknowledged as a co-author? Who is accountable if its output causes harm or leads to flawed science? While such issues were peripheral in earlier paradigms, they become central in autonomous discovery. Risks include ghost authorship, diminished human contribution, and misuse of FM-generated content. Addressing these concerns requires governance frameworks that distinguish mechanical from creative contributions, mandate transparent disclosures, and track downstream impacts of AI-generated outputs.

5.2 Future Directions: Toward Autonomous Scientific Discovery

Despite recent advances, most FM deployments remain confined to static prompts, predefined tasks, and fixed schema. They typically lack persistent memory, adaptive feedback, and physical embodiment. As such, their contributions, though impressive, are largely reactive and limited to isolated stages of the scientific process. Looking ahead, we identify three concrete research directions that define the transition toward autonomous scientific discovery:

Embodied Scientific Agents. A pivotal step toward scientific autonomy is grounding FMs in the physical world. Future FMs will increasingly be deployed within laboratory robotics, automated instruments, and digital twin environments. By coupling language-based reasoning with real-world perception and control, these agents will plan experiments, interact with physical systems, and iteratively refine procedures. This integration of abstract reasoning with physical execution is essential for closing the loop between scientific modeling and empirical verification. However, challenges remain in integrating high-level task planning with low-level control, ensuring robustness under real-world uncertainty, and maintaining safety and interpretability in dynamic lab environments.

Closed-Loop Scientific Autonomy. Current scientific workflows are typically open-loop: FMs assist with parts of the pipeline, but humans still decide the next steps. Moving toward truly autonomous science requires closed-loop systems, where FMs continuously formulate hypotheses, design and perform experiments, analyze results, and update internal models based on feedback. Current progress includes reinforcement learning-based planning [102], planning-as-inference [103, 104], and neuro-symbolic agents [105]. For example, recent neuro-symbolic agents have shown how structured memory and logic-based reasoning can guide molecule design or theorem proving [64]. Similarly, planning-as-inference approaches [106] and reinforcement learning-based agents [107] have been applied to automate scientific workflows such as hypothesis selection and experimental sequencing. A key challenge is ensuring that the loop remains robust to noisy observations, adaptive to shifting objectives, and aligned with scientific validity, not just reward maximization.

Continual Learning and Generalization. To operate effectively across scientific domains, FMs must transition from static systems to continual learners capable of accumulating and refining knowledge over time. This entails addressing key challenges such as catastrophic forgetting [108] and domain drift [109]. Promising approaches include parameter-efficient online adaptation [110], memory-augmented architectures [111], and modular lifelong learning frameworks [112] that allow selective knowledge retention and update. However, existing methods still fall short in enabling robust transfer across heterogeneous tasks and modalities. Advancing continual learning mechanisms would allow FMs to incrementally build domain-bridging representations, facilitate analogical reasoning across scientific contexts, and sustain coherent research trajectories over extended periods [113].

6 Conclusions

FMs are reshaping the landscape of scientific discovery. From enhancing existing workflows to enabling autonomous inquiry, they signal a potential shift toward a fifth scientific paradigm. In this paper, we proposed a three-stage framework, *i.e.*, meta-scientific integration, hybrid human-AI co-creation, and autonomous scientific discovery, to characterize this evolving trajectory. By analyzing FM integration across classical paradigms, we showed how FMs increasingly act not only as tools but as epistemic agents. While this transformation is still emerging, it raises profound questions about agency, authorship, and the nature of knowledge itself. Looking forward, we call for rigorous exploration of FM capabilities, responsible governance mechanisms, and deeper theoretical understanding to guide their role in science. Embracing this shift may redefine not just how we do science, but who or what can do science.

References

- [1] Steven Shapin. *The scientific revolution*. University of Chicago press, 2018.
- [2] Eric Winsberg. *Science in the age of computer simulation*. University of Chicago Press, 2019.
- [3] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- [4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
- [5] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pages 25904–25938. PMLR, 2023.
- [6] Bernardino Romera-Paredes, Andrew Brock, Gonzalo Trevino, et al. Program discovery in an open-ended domain with large language models. *Nature*, 627:39–46, 2024.
- [7] Nature Reviews Physics. More and different at the interface. *Nature Reviews Physics*, 4:497, August 2022. Published 03 August 2022.
- [8] Peter Yichen Chen, Jinxu Xiang, Dong Heon Cho, Yue Chang, GA Pershing, Henrique Teles Maia, Maurizio M Chiaramonte, Kevin Carlberg, and Eitan Grinspun. Crom: Continuous reduced-order modeling of pdes using implicit neural representations. *arXiv preprint arXiv:2206.02607*, 2022.
- [9] Jing Lyu, Shuo Wang, Trent E. Balias, Inderjit Singh, Alexey Levit, Yulia S. Moroz, Matthew J. O’Meara, Tiani Che, Erdenebaatar Algaas, Ksenia Tolmachova, Yaroslav Tolmachev, Brian K. Shoichet, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019.
- [10] Roman Frigg and Stephan Hartmann. Models in science. 2006.
- [11] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

- [13] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guaracci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaressi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.
- [14] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuteng Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024.
- [15] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.
- [16] Christopher Bishop. Ai4science to empower the fifth paradigm of scientific discovery. *Microsoft Research Blog*, 2022.
- [17] Elsevier Connect. Ai for science: A paradigm shift for scientific discovery and translation, 2024. Accessed: 2025-04-30.
- [18] Stephen Wolfram. Can ai solve science?, 2024. Accessed: 2025-04-30.
- [19] Dan McQuillan. *Resisting AI: An anti-fascist approach to artificial intelligence*. Policy Press, 2022.
- [20] Galileo Galilei. *Two new sciences*. Dover New York, 1914.
- [21] Robert Boyle. *The sceptical chymist*. Courier Corporation, 2013.
- [22] Hermann Weyl. *Symmetry*. Princeton University Press, 2015.
- [23] Eugene P Wigner. The unreasonable effectiveness of mathematics in the natural sciences. In *Mathematics and science*, pages 291–306. World Scientific, 1990.
- [24] Albert Einstein. The general theory of relativity. In *The meaning of relativity*, pages 54–75. Springer, 1922.
- [25] Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646, 1994.
- [26] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
- [27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

- [28] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe’er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.
- [29] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.
- [30] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- [31] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- [32] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- [33] xAI. Grok 3 beta — the age of reasoning agents. <https://x.ai/news/grok-3>, 2025.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [35] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [36] Fan Liu, Wenshuo Chao, Naiqiang Tan, and Hao Liu. Bag of tricks for inference-time computation of llm reasoning. *arXiv preprint arXiv:2502.07191*, 2025.
- [37] Naruki Yoshikawa, Marta Skreta, Kourosh Darvish, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Andrew Zou Li, Yuchi Zhao, Haoping Xu, Artur Kuramshin, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Large language models for chemistry robotics. *Autonomous Robots*, 47(8):1057–1086, 2023.
- [38] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292, 2024.
- [39] Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments, 2025.
- [40] Tung Nguyen, Sudhanshu Agrawal, and Aditya Grover. Expt: Synthetic pretraining for few-shot experimental design, 2023.
- [41] Taigao Ma, Haozhu Wang, and L. Jay Guo. Optogpt: A foundation model for inverse design in optical multilayer thin film structures, 2024.
- [42] Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models, 2024.
- [43] Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. A sober look at llms for material discovery: Are they actually good for bayesian optimization over molecules?, 2024.
- [44] Abdoulatif Cissé, Xenophon Evangelopoulos, Vladimir V. Gusev, and Andrew I. Cooper. Language-based bayesian optimization research assistant (bora), 2025.
- [45] Jacopo Iollo, Christophe Heinkelé, Pierre Alliez, and Florence Forbes. Bayesian experimental design via contrastive diffusions, 2025.
- [46] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

- [47] Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, Qiang Zhang, and Yiming Mo. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications*, 15(1):10160, 2024.
- [48] Davi Fébba, Kingsley Egbo, William A. Callahan, and Andriy Zakutayev. From text to test: Ai-generated control software for materials science instruments. *Digital Discovery*, 4(1):35–45, 2025.
- [49] Shray Mathur, Noah van der Vleuten, Kevin Yager, and Esther Tsai. Vision: A modular ai assistant for natural human-instrument interaction at scientific user facilities, 2024.
- [50] Venkatesh Sripada, Samuel Carter, Frank Guerin, and Amir Ghalamzan. Ap-vlm: Active perception enabled by vision-language models, 2024.
- [51] Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. A survey on hypothesis generation for scientific discovery in the era of large language models, 2025.
- [52] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, and Aidong Zhang. Embracing foundation models for advancing scientific discovery. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1746–1755. IEEE, 2024.
- [53] Charles O’Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. Sparks of science: Hypothesis generation using structured paper data, 2025.
- [54] Miaosen Chai, Emily Herron, Erick Cervantes, and Tirthankar Ghosal. Exploring scientific hypothesis generation with mamba. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart, editors, *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 197–207, Miami, FL, USA, November 2024. Association for Computational Linguistics.
- [55] Majid Farhadloo, Arun Sharma, Mingzhou Yang, Bharat Jayaprakash, William Northrop, and Shashi Shekhar. Towards physics-guided foundation models, 2025.
- [56] Runlong Yu, Chonghao Qiu, Robert Ladwig, Paul Hanson, Yiqun Xie, and Xiaowei Jia. Physics-guided foundation model for scientific discovery: An application to aquatic science, 2025.
- [57] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning, 2023.
- [58] Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers, 2024.
- [59] Daniel Cunningham, Mark Law, Jorge Lobo, and Alessandra Russo. The role of foundation models in neuro-symbolic learning and reasoning, 2024.
- [60] Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications, 2025.
- [61] Peiyang Song, Kaiyu Yang, and Anima Anandkumar. Lean copilot: Large language models as copilots for theorem proving in lean, 2025.
- [62] Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data, 2024.
- [63] Ziyang Li, Jiani Huang, Jason Liu, Felix Zhu, Eric Zhao, William Dodds, Neelay Velingker, Rajeev Alur, and Mayur Naik. Relational programming with foundational models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10635–10644, March 2024.
- [64] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *ICLR*, 2025.
- [65] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

- [66] Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *arXiv preprint arXiv:2409.09359*, 2024. NeurIPS 2024 version.
- [67] Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equations: Multi-operator learning and extrapolation. *arXiv preprint arXiv:2404.12355*, 2024.
- [68] Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. In *NeurIPS*, 2024.
- [69] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.
- [70] Mohammad Abdullah Rahman, Federica Boi, Siyuan Li, Tapio Schneider, Pietro Liò, and Yisong Yue. Pretraining codomain attention neural operators for solving multiphysics PDEs. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [71] Raymond Lam, Jonathan Weyn, Enguerrand Driss, Andrew Kealy, et al. Graph neural networks for global weather forecasting. *Science*, 381(6654):351–356, 2023.
- [72] Hao Li, Xun Huan, Qiang Du, and Rose Yu. Factformer: Scalable transformer architectures for PDE surrogate modeling. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [73] Philipp Lippe, Philipp Holl, Priya Jaini, Max Welling, and Philipp Vogt. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [74] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- [75] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkitt Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.
- [76] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wenhui Wang, Wei Li, Shufei Zhang, Mao Su, Wanli Ouyang, Yuqiang Li, and Dongzhan Zhou. Chemvilm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024.
- [77] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.
- [78] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- [79] Karan* Singhal, Shekoofeh* Azizi, Tongshuang Tu, Dai Tran, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.
- [80] Alvaro Sanchez-Gonzalez. Graphcast: Learning skillful medium-range global weather forecasting. In *EGU General Assembly Conference Abstracts*, page 11381, 2024.
- [81] Keisong Bi, Binbing Chen, Lingxi Xie, Xinlong Wang, et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 620(7971):224–229, 2023.
- [82] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. DiffusionSat: A generative foundation model for satellite imagery. In *International Conference on Learning Representations (ICLR)*, 2024.
- [83] John Jumper, Richard Evans, Alexander Pritzel, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.
- [84] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

- [85] Jacob L. Watson, David Juergens, Nicholas R. Bennett, Brian L. Trippe, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1113–1122, 2023.
- [86] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, pages 1–3, 2025.
- [87] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.
- [88] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. *CoRR*, abs/2501.04227, 2025.
- [89] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as AI research agents. *CoRR*, abs/2310.03302, 2023.
- [90] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292, 2024.
- [91] Fan Liu, Zherui Yang, Cancheng Liu, Tianrui Song, Xiaofeng Gao, and Hao Liu. Mm-agent: Llm as agents for real-world mathematical modeling problem. *arXiv preprint arXiv:2505.14148*, 2025.
- [92] Yuxuan Liu, Jingmin Sun, Xinjie He, Griffin Pinney, Zecheng Zhang, and Hayden Schaeffer. PROSE-FD: A multimodal PDE foundation model for learning multiple operators for forecasting fluid dynamics. *arXiv preprint arXiv:2409.09811*, 2024.
- [93] Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse PDE problems. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [94] Hao Wang, Luxi He, Rui Gao, and Flavio Calmon. Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. *Advances in Neural Information Processing Systems*, 36:27040–27062, 2023.
- [95] Santiago Cortes-Gomez, Mateo Dulce, Carlos Patino, and Bryan Wilder. Statistical inference under constrained selection bias. *arXiv preprint arXiv:2306.03302*, 2023.
- [96] Kai Gan and Tong Wei. Erasing the bias: Fine-tuning foundation models for semi-supervised learning. *arXiv preprint arXiv:2405.11756*, 2024.
- [97] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.
- [98] Huiwen Wu, Xiaohan Li, Xiaogang Xu, Jiafei Wu, Deyi Zhang, and Zhe Liu. Iter-ahmcl: Alleviate hallucination for large language model via iterative model-level contrastive learning. *arXiv preprint arXiv:2410.12130*, 2024.
- [99] Sachin Mehta, Mohammad Hosseini Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. Openelm: An efficient language model family with open training and inference framework. *arXiv preprint arXiv:2404.14619*, 2024.
- [100] Moritz Wolter and Lokesh Veeramacheneni. Position: More rigorous software engineering would improve reproducibility in machine learning research. *arXiv preprint arXiv:2502.00902*, 2025.
- [101] Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, 2025.
- [102] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.
- [103] Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.
- [104] Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in cognitive sciences*, 16(10):485–488, 2012.

- [105] Zenan Li, Zhaoyu Li, Wen Tang, Xian Zhang, Yuan Yao, Xujie Si, Fan Yang, Kaiyu Yang, and Xiaoxing Ma. Proving olympiad inequalities by synergizing llms and symbolic reasoning. *arXiv preprint arXiv:2502.13834*, 2025.
- [106] Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Agent planning with world knowledge model. *Advances in Neural Information Processing Systems*, 37:114843–114871, 2024.
- [107] Jonathan Bader, Nicolas Zunker, Soeren Becker, and Odej Kao. Leveraging reinforcement learning for task resource allocation in scientific workflows. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3714–3719. IEEE, 2022.
- [108] Oleksiy Ostapenko, Timothee Lesort, Pau Rodriguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In *Conference on lifelong learning agents*, pages 60–91. PMLR, 2022.
- [109] Huancheng Chen, Jingtao Li, Nidham Gazagnadou, Weiming Zhuang, Chen Chen, and Lingjuan Lyu. Dual low-rank adaptation for continual learning with pre-trained models. *arXiv preprint arXiv:2411.00623*, 2024.
- [110] Eric Nuertey Coleman, Luigi Quarantiello, Ziyue Liu, Qinwen Yang, Samrat Mukherjee, Julio Hurtado, and Vincenzo Lomonaco. Parameter-efficient continual fine-tuning: A survey. *arXiv e-prints*, pages arXiv–2504, 2025.
- [111] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.
- [112] Jiazu Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.
- [113] Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5):1–38, 2025.