Review

Not peer-reviewed version

# AI for Science: A Systematic Survey of Architectures, Autonomy, and Open Challenges

Yongqiang Zhao , Yuhan Liu , Yuqian Han , Huanyao Zhang , Yanfang Zhou , Ziliang Wang , Yu Huang , Yan Zhang , Yonghong Tian *

Posted Date: 1 July 2026

doi: 10.20944/preprints202607.0102.v1

Keywords: artificial intelligence for science; levels of autonomy; six-layer hierarchical framework; autonomous scientific discovery; scientific foundation models

 doi:10.20944/preprints202607.0102.v1



*Review*

# AI for Science: A Systematic Survey of Architectures, Autonomy, and Open Challenges

**Yongqiang Zhao [1], Yuhan Liu [2], Yuqian Han [1], Huanyao Zhang [1], Yanfang Zhou [2], Ziliang Wang [3], Yu Huang [1], Yan Zhang [1] and Yonghong Tian [1,*]**

[1] Peking University, Beijing, China
[2] Academy of Military Science of the People's Liberation Army, Changsha/Beijing, China
[3] Central South University, Changsha, China
* Correspondence: yhtian@pku.edu.cn

**Abstract**

Artificial Intelligence for Science (AI4S) is a pivotal research frontier revolutionizing scientific discovery, yet it suffers from fragmentation and lacks system-level exploration. This survey unifies its landscape: it defines AI4S against related concepts, verifies its status as the fifth scientific paradigm, and presents a six-layer framework and five-tier autonomy taxonomy. It also summarizes key techniques, core challenges and future directions, aiming to guide relevant research and infrastructure development.

**Keywords:** artificial intelligence for science; levels of autonomy; six-layer hierarchical framework; autonomous scientific discovery; scientific foundation models

## 1. Introduction

Recent years have witnessed the rapid rise of artificial intelligence (AI) as a transformative force across the natural and engineering sciences [1–3]. Its impact has already been demonstrated in a wide range of domains, including protein structure prediction in the life sciences [4], global weather and climate modeling in the geosciences [5–7], and complex system modeling in materials science and particle physics [8]. By substantially extending the capabilities of traditional scientific computing and data analysis, AI is increasingly recognized as an important enabler of scientific discovery [9,10]. In this context, Artificial Intelligence for Science (AI4S) has emerged as a frontier research area of growing interest to both academia and industry [9,11].

From the perspective of the evolution of scientific paradigms, AI4S can be viewed as part of a broader transition in how science is conducted. Jim Gray famously characterized scientific inquiry in terms of four paradigms: experimental science, theoretical science, computational science, and data-intensive science [12]. As modern scientific research increasingly relies on high-throughput experiments, large-scale observations, and complex instruments, the scale, heterogeneity, and complexity of scientific data have grown dramatically. Conventional workflows based on manual analysis, handcrafted modeling, and fragmented toolchains are therefore facing growing limitations. Meanwhile, advances in deep learning, multimodal learning, and large-scale AI models have created new opportunities for AI not only to accelerate existing scientific workflows, but also to support knowledge discovery, reasoning, and experimental automation. This trend has motivated an increasing discussion of whether AI-driven scientific discovery may represent a potential new paradigm beyond the traditional four [1,10,13,14].

The strategic importance of AI4S has also been increasingly reflected in national research agendas. In the United States, AI-enabled scientific research has been advanced through initiatives such as the National AI Research Institutes, the National Artificial Intelligence Research Resource (NAIRR) pilot, and high-performance computing programs supported by the Department of Energy [15,16]. In parallel, China has also seen rapid growth in AI-enabled scientific research, with increasing efforts devoted

 doi:10.20944/preprints202607.0102.v1



to scientific foundation models, research agents based on large language models, and automated experimental platforms [17–20]. With the rapid development of large models, multimodal learning, and laboratory automation, AI4S is moving beyond isolated task-specific optimization toward more integrated and system-level scientific capabilities.

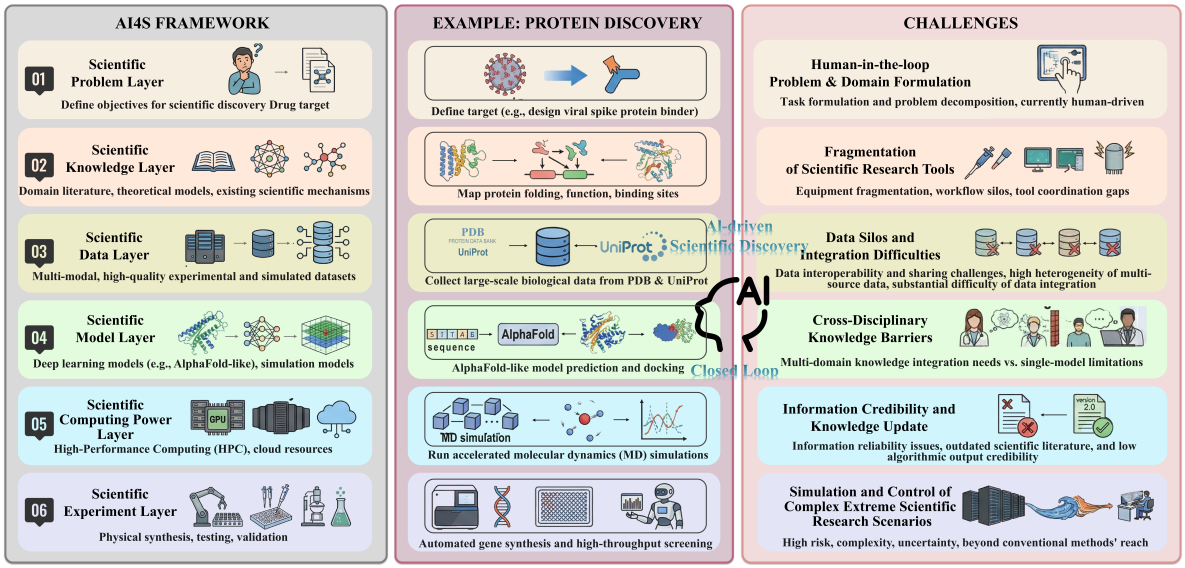


**Figure 1. Overview of the AI4S framework, its application in protein discovery, and core challenges. (Left)** The six-layer hierarchical architecture of AI4S, spanning from scientific experiments and computing power to models, data, knowledge, and problem formulation. **(Middle)** A detailed workflow of protein discovery as a case study, illustrating target definition, functional mapping, data collection (PDB/UniProt), AlphaFold-like prediction, MD simulation, and automated experimental screening. **(Right)** The AI-driven scientific discovery closed loop and its associated key challenges, such as tool fragmentation, data silos, cross-disciplinary barriers, and simulation of complex scenarios.

A fundamental question for AI4S is what kinds of scientific problems are particularly suitable for AI-driven approaches. In general, AI is most effective for problems characterized by large combinatorial search spaces, high-dimensional and nonlinear interactions, costly or risky experimentation, incomplete mechanistic understanding, and the availability of heterogeneous data or simulators. Typical examples include molecular and materials design, protein engineering, reaction optimization, climate and plasma modeling, and literature-driven hypothesis generation [21–24]. In these settings, AI can contribute not only by accelerating prediction and optimization, but also by generating candidate hypotheses, prioritizing experiments, integrating evidence across domains, and updating models through feedback. However, AI is less suitable as a standalone solution when problems lack measurable feedback, reliable data, falsifiable hypotheses, or clear validation criteria. Therefore, clarifying the types of scientific problems that AI can effectively address is essential for understanding the scope, boundaries, and system-level organization of AI4S.

Despite this rapid progress and the broad applicability of AI to these classes of scientific problems, current AI4S research remains highly fragmented. Existing studies often focus on individual scientific tasks, isolated technical directions, or particular disciplinary applications, such as scientific computing acceleration, literature-based knowledge mining, domain-specific prediction, or automated experimentation. In practice, however, scientific discovery is inherently a multi-stage process involving problem formulation, knowledge organization and reasoning, data acquisition and management, model construction and prediction, large-scale computation, and experimental validation [25]. These stages are tightly coupled, and bottlenecks in any one of them may constrain the efficiency and reliability of the overall research workflow. This motivates the need for a system-level framework for organizing AI4S research.

 doi:10.20944/preprints202607.0102.v1



Several influential perspectives and surveys have shaped the discourse on AI-enabled scientific discovery. In particular, [1] provided a landmark perspective on how recent advances such as self-supervised learning, geometric deep learning, and generative AI can support multiple stages of the scientific process, including hypothesis generation, experiment design, data interpretation, and the integration of experiments with simulation. These studies have been instrumental in establishing the promise of AI for scientific discovery. However, existing overviews have generally focused either on methodological breakthroughs or on selected scientific domains, leaving the system-level organization of AI4S across the full lifecycle of scientific discovery comparatively underexplored. In contrast, this survey explicitly positions AI4S as an integrated research ecosystem and organizes the field through a unified architectural framework, a clarified conceptual scope, an autonomy-oriented developmental trajectory, and a challenge-driven synthesis of the major bottlenecks that constrain AI-enabled scientific research.

To this end, this paper introduces the unified AI4S framework shown in Figure 1, which consists of six core layers: the scientific problem layer, scientific knowledge layer, scientific data layer, scientific model layer, scientific computing power layer, and scientific experiment layer. Specifically, the scientific problem layer concerns task formulation and decomposition; the scientific knowledge layer organizes scientific theories, literature knowledge, and domain knowledge graphs; the scientific data layer supports the acquisition, governance, and integration of heterogeneous scientific data; the scientific model layer focuses on modeling and prediction through machine learning and scientific computing methods [26,27]; the scientific computing power layer provides the infrastructure required for large-scale training, simulation, and inference; and the scientific experiment layer enables automated design, execution, and validation through intelligent platforms and robotic systems. Together, these six layers characterize how AI participates in the closed-loop process of scientific discovery, from problem formulation and model reasoning to experimental validation and knowledge updating.

At the same time, this survey argues that a clear conceptual boundary is necessary for understanding what AI4S includes and what differentiates it from adjacent fields. Although AI4S overlaps with Scientific Machine Learning, Autonomous Science, and AI for Scientific Discovery, we use the term in a broader system-level sense. In this paper, AI4S refers not merely to the use of AI for isolated prediction tasks, model acceleration, or laboratory automation, but to an intelligent research ecosystem that spans the full trajectory from problem formulation to knowledge abstraction, data integration, model construction, scientific computing, and experimental validation. Under this perspective, the role of AI is also evolving along a continuum of autonomy: from foundational tools for prediction and simulation, to research copilots, to collaborative co-scientists, and prospectively toward autonomous AI scientists and self-driving laboratories [28–31]. Clarifying this conceptual scope and developmental trajectory is essential for organizing the technical landscape reviewed in the remainder of this paper.

Beyond clarifying the conceptual scope of AI4S, this survey introduces a formal taxonomy for characterizing the degree of autonomy exhibited by AI4S systems. Specifically, we propose a Levels of Autonomy (LoA) framework that spans five tiers—from basic computational tools (L1) to fully autonomous self-driving laboratories (L5)—representing a progressive reduction in human involvement. We also define a Five-Dimensional Quantitative Evaluation Framework that decomposes AI4S autonomy into five measurable dimensions: Lifecycle Coverage Rate (LCR%), Autonomous Iteration Depth (AID%), Human Disengagement Degree (HDD%), Knowledge Novelty Percentage (NP%), and Domain Transfer Success Rate (DTS%). Together, these two components provide both a conceptual scaffold for understanding where any given AI4S system sits on the autonomy spectrum, and a practical instrument for tracking quantitative progress toward higher levels of system maturity. The final autonomy level is determined by a bottleneck rule, ensuring that overall system classification reflects its weakest dimension rather than its strongest.

It is worth noting that scientific problem formulation and the definition of research goals still largely depend on human researchers and their domain expertise. Accordingly, this survey focuses primarily on the remaining five technical dimensions through which AI most directly reshapes scientific

 doi:10.20944/preprints202607.0102.v1



discovery: scientific knowledge, scientific data, scientific models, scientific computing, and scientific experimentation. Across these dimensions, AI4S continues to face a number of fundamental and recurring bottlenecks, including the fragmentation of scientific research tools, data silos and integration difficulties, cross-disciplinary knowledge barriers, information credibility and timely knowledge updating, and the simulation and control of complex extreme-condition scientific scenarios.

Motivated by the above perspective, this survey makes four main contributions. First, we introduce a unified AI4S framework that characterizes the role of AI across the scientific discovery lifecycle, spanning scientific problems, knowledge, data, models, computing, and experimentation. Second, we clarify the conceptual scope of AI4S by distinguishing it from related paradigms such as Scientific Machine Learning, Autonomous Science, and AI for Scientific Discovery, and we introduce a formal Levels of Autonomy (LoA) taxonomy alongside a Five-Dimensional Quantitative Evaluation Framework to systematically characterize and measure the degree of AI autonomy across the scientific discovery lifecycle, outlining a developmental trajectory from AI as a foundational tool to a collaborative co-scientist and, prospectively, an autonomous scientific agent. Third, we provide a structured review of representative technical pathways across five core dimensions of AI-enabled scientific research, namely scientific knowledge, scientific data, scientific models, scientific computing, and scientific experimentation. Fourth, we identify five fundamental challenges that currently limit the development of AI4S, including tool fragmentation, data silos, cross-disciplinary knowledge barriers, information credibility and timely knowledge updating, and the simulation and control of complex extreme-condition scientific scenarios, and we discuss future opportunities for building the next generation of AI-native scientific research systems.

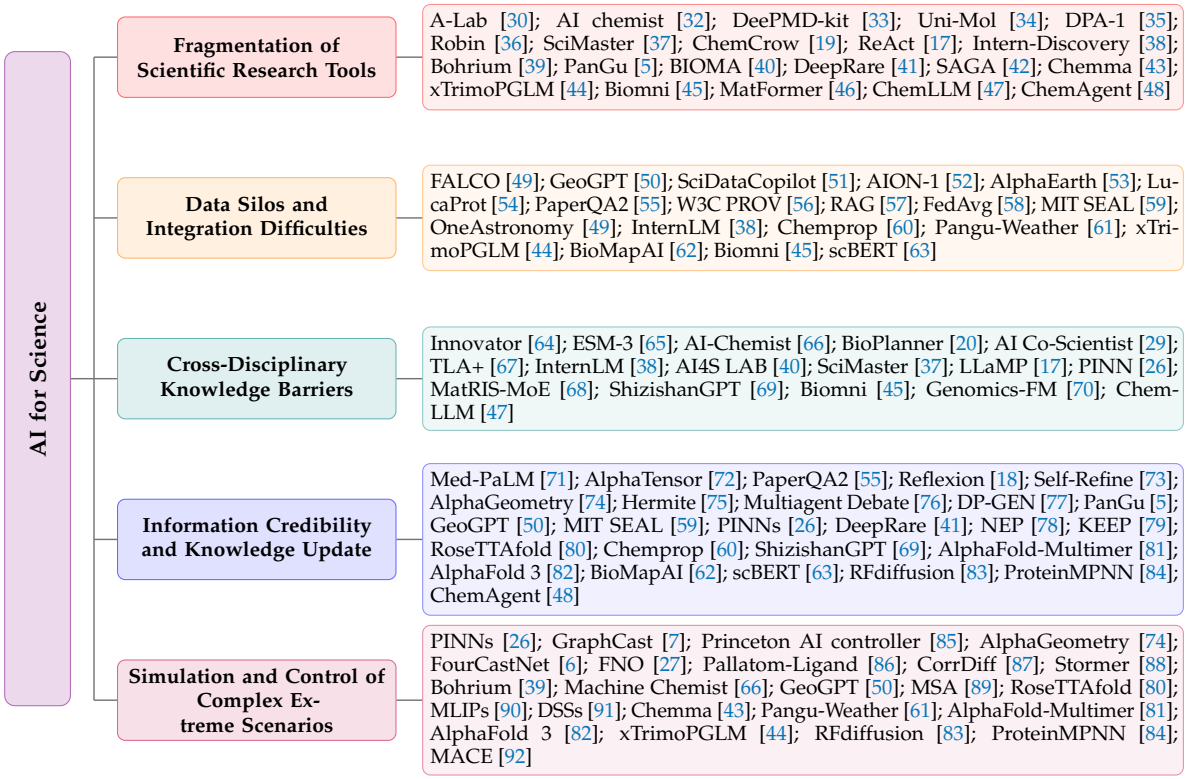


**Figure 2.** Systematic Framework of AI for Science.

## 2. AI for Science: Conceptual Framework and Autonomy Classification

### *2.1. Definition and Disciplinary Boundaries*

**AI for Science (AI4S)** represents an emerging paradigm that conceptualizes artificial intelligence as a foundational infrastructure for scientific cognition. Beyond merely accelerating computational segments within existing workflows, AI4S actively engages in the knowledge creation process—encompassing hypothesis generation, pattern discovery, inverse design, and autonomous experimenta-

 doi:10.20944/preprints202607.0102.v1



tion. By capturing the underlying structures of large-scale scientific data, it establishes a **fifth scientific paradigm** centered on machine agents, distinguishing itself from the fourth paradigm characterized by statistical law discovery [1,13,14], which relied primarily on extracting statistical correlations from large-scale data. In contrast, the fifth paradigm is characterized by machines that actively formulate falsifiable causal hypotheses and autonomously design experiments to test them—a shift from pattern recognition to hypothesis-driven inquiry.

Conceptually, AI4S must be distinguished from several related fields. While **Scientific Machine Learning (SciML)** prioritizes the integration of physical priors with neural networks to ensure model-level consistency and interpretability, **Self-driving Laboratories** represent the cyber-physical instantiation of AI4S at the highest autonomy tier (L5), where AI-directed robotic platforms close the experimental loop—rather than constituting a separate paradigm, they operationalize the full-lifecycle integration that defines AI4S at its most advanced form. Furthermore, **AI for Scientific Discovery** tends to concentrate on leveraging AI to expedite specific breakthroughs. In contrast, AI4S spans the entire trajectory from problem formulation to knowledge abstraction, emphasizing system-level integration and the holistic restructuring of scientific workflows. In a broad sense, AI4S encompasses any system that employs AI as a core driver for scientific inquiry, spanning from isolated computational tools (L1) to fully autonomous cyber-physical agents (L5). In a narrower, stronger sense, AI4S denotes systems capable of cross-task integration and inherent autonomy—systems that not only accelerate individual research steps but restructure the scientific workflow itself into a self-sustaining cycle of hypothesis generation, experimental validation, and knowledge abstraction.

The AI4S paradigm has undergone a structural transition from "static computational tools" to "dynamic autonomous agents," establishing a comprehensive **Levels of Autonomy** spectrum—ranging from isolated task execution to closed-loop physical experimentation. This transition has been made tractable by the advent of large language models and multi-agent frameworks, which provide the reasoning, planning, and tool-use capabilities necessary to bridge isolated research steps into coherent autonomous pipelines. As illustrated in Table 1, this spectrum delineates a progressive reduction in human intervention: evolving from human-centric assistance at stages L1–L2 to the multi-agent collaborative frameworks typified by the Google AI Co-scientist at L3 [29], and ultimately toward the self-driven discovery loops of L4–L5.

**Table 1.** Levels of Autonomy: Quantitative performance standards across five dimensions.

| Level | D1: LCR% | D2: AID% | D3: HDD% | D4: NP% | D5: DTS% |
|---|---|---|---|---|---|
| **L1** *Basic Tool* | 0–25% (0–1.25 stages) | 0–25% (<0.25 cycles) | 0–25% (>75% human) | 0–25% (<5% novel) | 0% |
| **L2** *Copilot* | 25–45% (1.25–2.25 stages) | 25–45% (0.25–0.45 cycles) | 25–45% (55–75% human) | 5–30% | <30% |
| **L3** *Co-Scientist* | 45–70% (2.25–3.5 stages) | 45–70% (0.45–0.70 cycles) | 45–70% (30–55% human) | 30–70% | 30–60% |
| **L4** *AI Scientist* | 70–90% (3.5–4.5 stages) | 70–90% (0.70–0.90 cycles) | 70–90% (10–30% human) | 30–70% | 60–80% |
| **L5** *Self-driving Lab* | 90–100% (4.5–5 stages) | 90–100% (>0.90 cycles) | 90–100% (<10% human) | 50–100% | 80–100% |

### *2.2. The Spectrum of Autonomy and Quantitative Evaluation Framework*

The remarkable diversity in AI4S system capabilities necessitates a principled framework for systematic classification and assessment. To address this need, we introduce the **Levels of Autonomy (LoA)** taxonomy, which provides a structured spectrum from human-centric tool assistance to fully autonomous cyber-physical scientific agents. This multi-tier framework serves as both a conceptual scaffold for understanding AI4S system architectures and a practical foundation for comparing systems across different scientific domains and implementation maturity stages. As shown in Table 1, the LoA spectrum defines five distinct tiers, each with specific quantitative thresholds across the five evaluation dimensions.

 doi:10.20944/preprints202607.0102.v1



However, while qualitative categorization alone offers intuitive guidance, it proves insufficient for precise boundary discrimination between adjacent autonomy tiers or for tracking quantitative progress toward higher levels of system maturity. To remedy this limitation, we introduce a **Five-Dimensional Quantitative Evaluation Framework** that complements the qualitative LoA taxonomy. This framework decomposes AI4S autonomy into five independent, measurable dimensions, each capturing a distinct aspect of system capability.

#### 2.2.1. The Five Evaluation Dimensions

Each AI4S system is evaluated along five orthogonal dimensions, each scored as a continuous percentage (0–100%) that subsequently maps to the corresponding autonomy level (L1–L5):

**D1: Lifecycle Coverage Rate (LCR%).** This dimension quantifies the breadth of system autonomy across the research process. It measures what fraction of five canonical scientific research stages—problem formulation, hypothesis generation, experimental design, data analysis, and knowledge dissemination—the system can autonomously execute:

$$\text{LCR\%} = \frac{\text{Number of autonomous stages}}{5} \times 100\% \tag{1}$$

Each stage is scored on a three-point scale: 1.0 if the system executes the stage fully without human intervention, 0.7 if human approval or correction is required at one or more steps within the stage, and 0 if the stage is not covered. LCR% is thus equivalently computed as the sum of the five stage scores multiplied by 20%.

**D2: Autonomous Iteration Depth (AID%).** This dimension captures the system's ability to form closed-loop feedback cycles without human intervention. It measures the proportion of research iterations in which the system completes a full hypothesis-verification-refinement cycle autonomously:

$$\text{AID\%} = \frac{\text{Uninterrupted feedback cycles}}{\text{Total cycles}} \times 100\% \tag{2}$$

A single cycle is operationally defined as one complete pass through hypothesis generation, experimental execution or simulation, result analysis, and hypothesis revision. Any human intervention requiring approval or corrective instruction terminates the cycle and marks it as interrupted; passive monitoring or log inspection does not constitute an interruption.

**D3: Human Disengagement Degree (HDD%).** This dimension quantifies the reduction in human decision-making authority across the system lifecycle. Recognizing that not all decisions carry equal weight, we apply a weighted calculation:

$$\text{HDD\%} = 100\% - \frac{\sum(\text{weighted human decisions})}{\sum(\text{total decision weights})} \times 100\% \tag{3}$$

where decision weights are assigned as: problem definition decisions (weight 1×), mid-process decisions such as direction approval and error correction (weight 2×, most critical), and post-hoc result validation decisions (weight 0.5×, least critical).

In practice, pre-process decisions typically comprise a single problem-scoping or objective-definition step, in-process decisions encompass all direction changes, error corrections, and hypothesis approvals occurring mid-workflow, and post-process decisions include result validation and dissemination approval. The denominator sums the weights of all decision points identified for the evaluated workflow, while the numerator sums the weights of only those points at which a human actually intervened.

 doi:10.20944/preprints202607.0102.v1



**D4: Knowledge Novelty Percentage (NP%).** This dimension assesses the originality of system outputs relative to existing literature. It integrates three sub-dimensions—novelty of hypotheses, novelty of methodologies, and novelty of conclusions—and computes their mean:

$$\text{NP\%} = \text{mean}(\text{hypothesis novelty}, \text{method novelty}, \text{conclusion novelty}) \tag{4}$$

Each sub-dimension ranges from 0% (pure reproduction of known work) to 100% (fully original contribution) and is independently assessed by domain experts.

To facilitate consistent scoring, evaluators are provided with four anchor points: 0% indicates pure replication of known results; approximately 30% corresponds to incremental improvement using established methods; approximately 70% denotes a novel methodology or finding absent from prior literature; and 100% reflects a paradigm-shifting contribution. At least two independent domain experts are required per evaluation; when their scores diverge by more than 20 percentage points, a third arbitrator is consulted and the median value is adopted.

**D5: Domain Transfer Success Rate (DTS%).** This dimension measures the system's generalization and cross-disciplinary applicability. It captures the fraction of test domains (with >50% dissimilarity to the training domain) in which the system achieves comparable or superior performance:

$$\text{DTS\%} = \frac{\text{Successful domain transfers}}{\text{Total test domains}} \times 100\% \tag{5}$$

A minimum of three test domains is required for a statistically meaningful DTS% estimate, with each test domain exhibiting greater than 50% dissimilarity to the training domain as assessed by task type, data modality, or disciplinary classification. A transfer is considered successful if the system achieves at least 90% of the performance attained by a domain-specific baseline on the same task in the target domain.

#### 2.2.2. Mapping Dimensions to Autonomy Levels

Each dimension's percentage score maps to an autonomy tier according to fixed thresholds. As shown in Table 2, dimensions D1–D4 follow a uniform linear mapping, while D5 (domain transfer) employs a more conservative mapping to reflect the substantially greater difficulty of cross-domain generalization:

**Table 2.** Mapping Dimension Percentage Scores to Autonomy Levels (L1–L5)

| Dimension | 0–25% | 25–45% | 45–70% | 70–90% | 90–100% |
|---|---|---|---|---|---|
| D 1 | L1 | L2 | L3 | L4 | L5 |
| D 2 | L1 | L2 | L3 | L4 | L5 |
| D 3 | L1 | L2 | L3 | L4 | L5 |
| D 4 | L1 | L2 | L3 | L4 | L5 |
| D 5 | L1 | L1–L2 | L2–L3 | L3–L4 | L4–L5 |

#### 2.2.3. Determining Final Autonomy Level: The Bottleneck Rule

The final autonomy level of an AI4S system is determined by the **bottleneck rule**, which applies the principle that overall system maturity is constrained by its weakest dimension:

$$L_{\text{final}} = \min(L_{D_i}) + \Delta, \quad \Delta = \begin{cases} 1 & \text{if } |\{i : L_{D_i} > \min(L_{D_j})\}| \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This rule identifies the dimension with the lowest mapped autonomy level (the "bottleneck") and conditionally increments it by one. The compensation factor $\Delta$ encodes a *sufficiency-of-evidence* condition: it equals 1 only when at least two dimensions exceed the bottleneck level, providing

 doi:10.20944/preprints202607.0102.v1



convergent evidence from multiple independent aspects that the system's true capability surpasses what its weakest dimension alone would suggest. A single dimension exceeding the bottleneck—however dramatically—may reflect specialized engineering of one subsystem rather than genuine overall advancement, and therefore does not warrant elevation. This design yields three desirable properties: (1) systems uniformly at L1 remain at L1 without ad hoc capping; (2) systems with a single outlier dimension (e.g., strong iteration depth but otherwise uniform) are not artificially inflated; and (3) systems with broad multi-dimensional strength above their bottleneck receive appropriate credit.

### *2.3. Illustrative Case Studies*

To demonstrate the practical application of the LoA framework, we present four representative systems spanning distinct autonomy tiers. For each case, we show how the five-dimensional scores are derived from published system capabilities and how the bottleneck rule determines the final classification.

Case 1: CLAIRify — Level L1 (Basic Tool).

CLAIRify [93] is a natural-language-to-code translation framework developed at the University of Toronto. It uses LLM-based iterative prompting with program verification to convert free-text experiment descriptions into syntactically valid XDL (Chemical Description Language) programs that can be executed by laboratory robots. Its five-dimensional assessment is:

| | D1 (LCR) | D2 (AID) | D3 (HDD) | D4 (NP) | D5 (DTS) | Level per dim. |
|---|---|---|---|---|---|---|
| Score | 14% | 0% | 15% | 0% | 15% | — |
| Mapped Level | L1 | L1 | L1 | L1 | L1 | — |

**Scoring rationale.** D1 is minimal (14%) because CLAIRify addresses only a single sub-task within experimental design—translating a human-written natural language description into structured robot-executable code. It does not formulate problems, generate hypotheses, execute experiments with feedback, analyze data, or disseminate findings. Only the "experimental design" stage receives a partial score (0.7) due to the translation being constrained to pre-defined instructions rather than autonomous planning. D2 is zero because the system operates as a one-directional pipeline: natural language input $\rightarrow$ XDL output, with no closed-loop iteration or refinement based on experimental outcomes. D3 is low (15%) because the human retains nearly all decision authority—writing the experiment description, verifying the generated plan, and interpreting any results. D4 is zero as CLAIRify is a pure translation tool that generates no scientific knowledge; it faithfully reproduces what the human describes rather than proposing novel experiments. D5 scores 15% because, while the framework is architecturally DSL-agnostic (it can theoretically target any structured language), it has only been validated with XDL for chemistry.

**Bottleneck calculation.** All dimensions map to L1, so $\min = \max = \text{L1}$, yielding $\Delta = 0$. Thus $L_{\text{final}} = \text{L1} + 0 = \textbf{L1}$. Because the system exhibits no dimension above its bottleneck level, no compensation is warranted. CLAIRify exemplifies the L1 (Basic Tool) archetype: a valuable enabling component that lowers the barrier to using robotic systems, but contributes no autonomous scientific reasoning. It serves as a building block within larger AI4S architectures—for example, it could function as the "instruction translation module" in a higher-autonomy system like ORGANA—but on its own it remains firmly in the tool-assistance tier.

Case 2: Coscientist — Level L2 (Copilot).

Coscientist [94] is an LLM-based agent that autonomously searches literature, designs multi-step chemical procedures, controls robotic instruments via APIs, and analyzes experimental results. Its five-dimensional assessment is:

 doi:10.20944/preprints202607.0102.v1



| | D1 (LCR) | D2 (AID) | D3 (HDD) | D4 (NP) | D5 (DTS) | Level per dim. |
|---|---|---|---|---|---|---|
| Score | 74% | 50% | 55% | 10% | 40% | — |
| Mapped Level | L4 | L3 | L3 | L1 | L2 | — |

**Scoring rationale.** D1 is high (74%) because the system covers four of five stages: partial problem formulation via prompt interpretation (0.7), literature-based hypothesis generation (1.0), experimental design (1.0), and data analysis (1.0), lacking only autonomous dissemination. D4 is critically low (10%) because all demonstrated experiments—including Suzuki and Sonogashira couplings—reproduce known reactions rather than generating novel scientific knowledge.
**Bottleneck calculation.** $\min(\text{L4}, \text{L3}, \text{L3}, \text{L1}, \text{L2}) = \text{L1}$; four dimensions exceed L1 ($\geq$ 2), so $\Delta = 1$. Thus $L_{\text{final}} = \text{L1} + 1 = \textbf{L2}$. Despite broad lifecycle coverage and strong architectural generalizability, the system's inability to produce novel findings constrains it to the Copilot tier—a powerful executor of human-directed science.

Case 3: A-Lab — Level L3 (Co-Scientist).

A-Lab [30] is an autonomous materials synthesis platform at UC Berkeley that uses ML-predicted synthesis recipes, active learning for condition optimization, robotic powder handling, and automated XRD characterization:

| | D1 (LCR) | D2 (AID) | D3 (HDD) | D4 (NP) | D5 (DTS) | Level per dim. |
|---|---|---|---|---|---|---|
| Score | 60% | 75% | 65% | 55% | 5% | — |
| Mapped Level | L3 | L4 | L3 | L3 | L1 | — |

**Scoring rationale.** D2 is high (75%) reflecting 17 days of fully autonomous operation with closed-loop iteration (synthesize $\rightarrow$ XRD characterize $\rightarrow$ ML-adjust recipe $\rightarrow$ re-synthesize). D4 reaches 55% because the system realized 41 novel compounds that had been computationally predicted but never physically synthesized—genuine first-time experimental validation. D5 is near-zero (5%) as the platform is purpose-built exclusively for solid-state inorganic powder synthesis.
**Bottleneck calculation.** Full: $\min(\text{L3}, \text{L4}, \text{L3}, \text{L3}, \text{L1}) = \text{L1}$; four dimensions exceed L1, so $\Delta = 1$, yielding $L_{\text{full}} = \textbf{L2}$. Core (D1–D4): $\min(\text{L3}, \text{L4}, \text{L3}, \text{L3}) = \text{L3}$; only one dimension (D2) exceeds L3, so $\Delta = 0$, yielding $L_{\text{core}} = \textbf{L3}$ (Co-Scientist). The gap between the strong D2 performance (L4) and the final Core classification (L3) illustrates the sufficiency-of-evidence principle: a single outstanding dimension does not warrant elevation when the remaining dimensions uniformly indicate the bottleneck level.

Case 4: LUMI-Lab — Level L4 (AI Scientist).

LUMI-Lab [95] represents the current frontier: a foundation model pretrained on 28 million molecules proposes lipid nanoparticle candidates, active learning designs synthesis experiments, an Opentrons robot executes them, and automated transfection assays feed results back to update the model:

| | D1 (LCR) | D2 (AID) | D3 (HDD) | D4 (NP) | D5 (DTS) | Level per dim. |
|---|---|---|---|---|---|---|
| Score | 74% | 90% | 80% | 75% | 15% | — |
| Mapped Level | L4 | L5 | L4 | L4 | L1 | — |

**Scoring rationale.** D2 reaches 90% through a fully closed self-driving laboratory loop (predict $\rightarrow$ synthesize $\rightarrow$ evaluate $\rightarrow$ update model) operating with minimal human intervention across multiple rounds. D4 scores 75% because the system made a genuinely unexpected discovery: brominated lipid tails—a class of molecules not previously associated with mRNA delivery—significantly enhance transfection efficiency. This finding was *not anticipated by the human research team*, representing authentic AI-driven scientific serendipity. D3 is high (80%) as humans define only the therapeutic objective (mRNA delivery efficiency), with all subsequent exploration autonomously directed.

 doi:10.20944/preprints202607.0102.v1



**Bottleneck calculation.** Core (D1–D4): $\min(\text{L4, L5, L4, L4}) = \text{L4}$; only one dimension (D2) exceeds L4, so $\Delta = 0$, yielding $L_{\text{core}} = \textbf{L4}$ (AI Scientist). This result naturally captures the intuition that LUMI-Lab, while exhibiting exceptional autonomous iteration depth, has not yet achieved L5-level performance across multiple dimensions simultaneously. It represents the highest autonomy tier achieved by any current system.

These four cases reveal a consistent pattern: **domain transfer (D5) constitutes the universal bottleneck** across all existing AI4S systems, regardless of their single-domain sophistication. Moreover, the contrast between CLAIRify (L1) and higher-tier systems highlights that *closed-loop feedback and autonomous iteration are the critical differentiators* separating basic tools from genuine scientific agents. Achieving the fully autonomous L5 (Self-driving Lab) tier will require not merely perfecting closed-loop experimentation within one domain, but developing transferable scientific reasoning capabilities that generalize across disciplines.

## 3. Core Challenges and Solutions in AI-Enabled Scientific Research

### *3.1. Fragmentation of Scientific Research Tools*

The fragmentation of scientific research tools represents one of the core challenges confronting AI-driven scientific discovery (AI4S). This challenge manifests primarily in the dispersion of experimental equipment, the disconnection of workflows, and the difficulty of cross-tool coordination. In traditional scientific research paradigms, the journey from hypothesis formulation to experimental validation typically requires invoking a wide variety of disparate tools and instruments, resulting in diminished research efficiency, redundant labor, and significant barriers to interdisciplinary collaboration. Although the intelligence level of individual research tools has improved with the deep integration of AI technologies into scientific practice, the lack of unified interface standards across tools, inconsistencies in data formats, and incoherent workflow pipelines have become increasingly prominent, severely impeding the large-scale deployment and broader adoption of AI4S. Addressing this challenge is of critical importance for achieving the deep integration of "AI + scientific research" and for driving the paradigm shift toward a "full-pipeline intelligent" mode of scientific inquiry. To effectively respond to this challenge, the research community is actively exploring a range of innovative pathways, progressively realizing the intelligent transformation of scientific workflows through the construction of unified platforms, the adoption of advanced model paradigms, and the development of agent-based architectures.

#### 3.1.1. Building Unified Automated Experimental Platforms

Building unified automated experimental platforms constitutes a key pathway to resolving the fragmentation of scientific research tools. The core mechanism of this approach lies in integrating experimental equipment, data streams, and AI agents to connect the entire pipeline from hypothesis generation to experimental validation, thereby forming an intelligent closed loop of "design–execute–analyze." This effectively addresses the problems of dispersed experimental equipment, disconnected workflows, and low efficiency of manual operations.

The Shanghai Artificial Intelligence Laboratory's "InternLM" Scientific Discovery Platform [38] has successfully connected over a hundred categories of experimental devices, enabling remote invocation and automated scheduling. The AI chemist developed at the University of Liverpool [32] integrates synthesis and analytical equipment through a mobile robotic system, achieving fully automated decision-making across the entire experimental pipeline. The A-Lab system has established a hardware integration framework comprising robotic arms, high-temperature furnaces, and X-ray diffraction (XRD) instruments, enabling the complete automation of the "design–synthesize–characterize" workflow for inorganic materials, with a daily sample throughput 50 to 100 times greater than that achievable by human researchers. DP Technology's "Bohrium Research Space Station" [39] integrates AI force field models such as DeePMD-kit [33] and Uni-Mol [34] with robotic experimental interfaces, constructing an end-to-end workflow spanning molecular design, property prediction, and experimental valida-

 doi:10.20944/preprints202607.0102.v1



tion, thereby significantly reducing the cost associated with researchers switching between different software tools. These implementations collectively demonstrate that unified automated experimental platforms are driving a transformation in the scientific research paradigm from "manual operation" to "unmanned, high-throughput" execution, substantially enhancing experimental efficiency. In addition to the above, Biomni [45] and Chemma [43] focus on mitigating tool fragmentation by integrating agents into closed-loop experimental pipelines or unified action spaces, thereby facilitating seamless synergy between computational reasoning and physical or database-driven experimentation.

#### 3.1.2. Adopting the Pre-training and Fine-tuning Paradigm

Adopting the pre-training and fine-tuning paradigm represents an effective strategy for mitigating the fragmentation of scientific research tools. The core mechanism involves leveraging large-scale general-purpose pre-trained models to cover common scientific research scenarios, requiring only lightweight fine-tuning to adapt to diverse disciplines and experimental requirements. This approach avoids the redundant development of specialized tools for individual scenarios or devices, thereby enabling a single general-purpose model to support multi-device, multi-domain workflows.

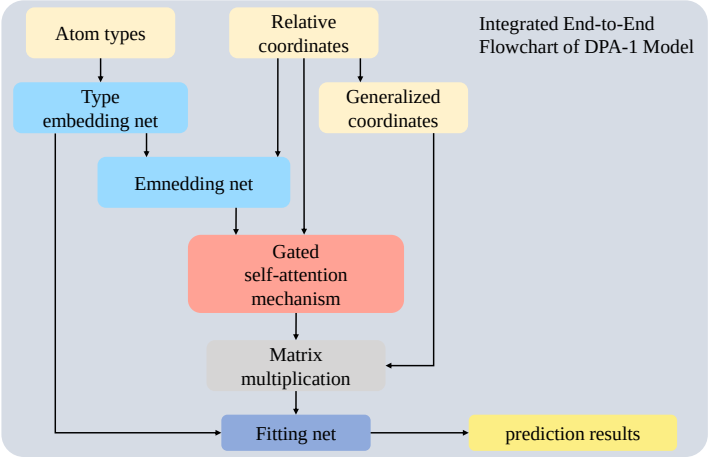


**Figure 3. DPA-1 Model: End-to-End Pipeline from Atomic Inputs to Prediction Results.** This flowchart illustrates the DPA-1 model pipeline, which takes atom types and relative coordinates as inputs, processes them through type embedding, generalized coordinate transformation, local embedding, gated self-attention, and matrix multiplication, and finally outputs prediction results via a fitting network.

Taking the DPA-1 platform [35] as an example, with its architecture illustrated in Figure 3, this platform integrates a large-scale dataset encompassing 56 elements to complete pre-training of interatomic potential functions. Through fine-tuning on small amounts of domain-specific data, the model can be readily adapted to molecular simulations across diverse material systems—including elemental solids, alloys, and semiconductors—effectively resolving the problems of tool dispersion and repetitive labor inherent in conventional molecular simulation frameworks, and significantly improving model transferability and tool reusability. The core loss function is formulated as follows:

$$\mathcal{L}_w(E^w, \mathcal{F}^w) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \left( p_\epsilon |E_t - E_t^w|^2 + p_f |\mathcal{F}_t - \mathcal{F}_t^w|^2 \right). \tag{7}$$

It adopts the Adamstochastic gradient descent method to minimize the discrepancies between the predicted values $E^w$ and $\mathcal{F}^w$ and their corresponding ground truth values by optimizing the model parameters $w$. Here $\mathcal{B}$ represents a minibatch, $|\mathcal{B}|$ is the batch size, $t$ denotes the index of the training sample. It also adopts a scheduler to tune the pre-factors $p_\epsilon$ and $p_f$ during the training process to make a better balance [35].

The "PanGu" foundational scientific large model released by the Chinese Academy of Sciences [96] adopts a symbolic–neural hybrid architecture [97] for general pre-training, followed by domain-specific fine-tuning for condensed matter physics, high-energy physics, and related fields. It has been

 doi:10.20944/preprints202607.0102.v1



successfully applied to synchrotron radiation data analysis and particle physics event reconstruction, avoiding the redundant investment associated with developing dedicated analytical tools for each large-scale scientific facility. Robin [36] forms an iterative closed loop by orchestrating multiple specialized agents responsible for literature review, candidate molecule evaluation, and data analysis. The "InternLM" Scientific Discovery Platform employs a general–domain fused large model as the "cognitive brain" of research tasks, deeply integrating general AI capabilities with domain-specific expertise in materials science and biomedical research. In addition, xTrimoPGLM [44], ChemLLM [47], and MatFormer [46] leverage the large-scale pre-training and fine-tuning paradigm to establish domain-specific foundation models, seeking to consolidate fragmented, task-specific utilities into a unified and versatile generative framework.

#### 3.1.3. Developing Agent Hub Architectures

Developing agent hub architectures represents the core pathway to realizing the intelligent transformation of scientific research workflows. The fundamental mechanism of this approach is to treat containers as the minimum deliverable execution units, ensuring environmental consistency and execution safety, and effectively resolving the canonical problem of "it works on my machine." By positioning AI agents as the "neural hub" of the research pipeline, this approach establishes specialized scientific agents and multi-agent collaboration frameworks, enabling dynamic scheduling of experimental equipment, interpretation of multidisciplinary knowledge, and coordination of full-pipeline research tasks, driving the transformation of the scientific research paradigm from "tool stacking" to "intelligent collaboration."

The AI4S LAB platform co-developed by Peking University Shenzhen Graduate School and Baidu Intelligent Cloud constructs four functional agents—PredAgent, ProAgent, OperAgent, and ComAgent—to simulate the cognitive and operational workflows of scientists, which is referred to as the BIOMA Multi-Agent System, realizing a research paradigm characterized by "wet–dry closed loops" and "full-pipeline digital intelligence" [40], as shown in Figure 4. The "InternLM" Scientific Discovery Platform achieves coordinated scheduling of experimental equipment and data resources through a multi-agent framework, effectively overcoming the challenge of invoking multiple heterogeneous tools across the pipeline from hypothesis generation to experimental validation. The SciMaster general-purpose scientific research agent [37] enables collaborative operation among internal agents dedicated to literature review, experimental design, and code generation, dynamically invoking external tool chains to provide users with an integrated scientific research experience. Finch employs containerized sandboxes (Aviary/Docker [98]) pre-loaded with mainstream bioinformatics libraries such as Scanpy and Seurat, ensuring environmental consistency for bioinformatics workflows. DeepRare [41], a multi-agent system integrating more than 40 specialized tools, adopts agent hub architectures as a concrete solution to the fragmentation of scientific research tools. GROMACS Copilot adopts a multi-layer abstraction architecture to enable natural language-driven molecular dynamics simulation. To fundamentally resolve the problem of tool chain heterogeneity, ChemCrow [19] proposes a plug-and-play tool ecosystem design that encapsulates the capabilities required for specific scientific tasks as standardized "tools," articulates a "capability matrix" comprising input–output specifications, preconditions, and failure modes, and defines a unified invocation contract. Furthermore, by introducing a suite of safety-oriented tools, it effectively mitigates the risk of error propagation arising from external tool invocations. Building upon this foundation, the ReAct [17] paradigm supplements existing command-line interfaces (CLIs) or scripts with wrappers, enabling the model to interleave reasoning and action (tool invocation) in an alternating fashion so that these interfaces can be scheduled by agents in a structured manner, thereby effectively resolving the problems of task disconnection and hallucination propagation that result from purely text-based generation. Aside from the above discussion, SAGA [42]and ChemAgent [48] are centered on developing sophisticated agent hub architectures characterized by autonomous goal evolution or self-updating memory systems, which orchestrate multifaceted research workflows through high-level strategic planning and iterative experience accumulation.

 doi:10.20944/preprints202607.0102.v1



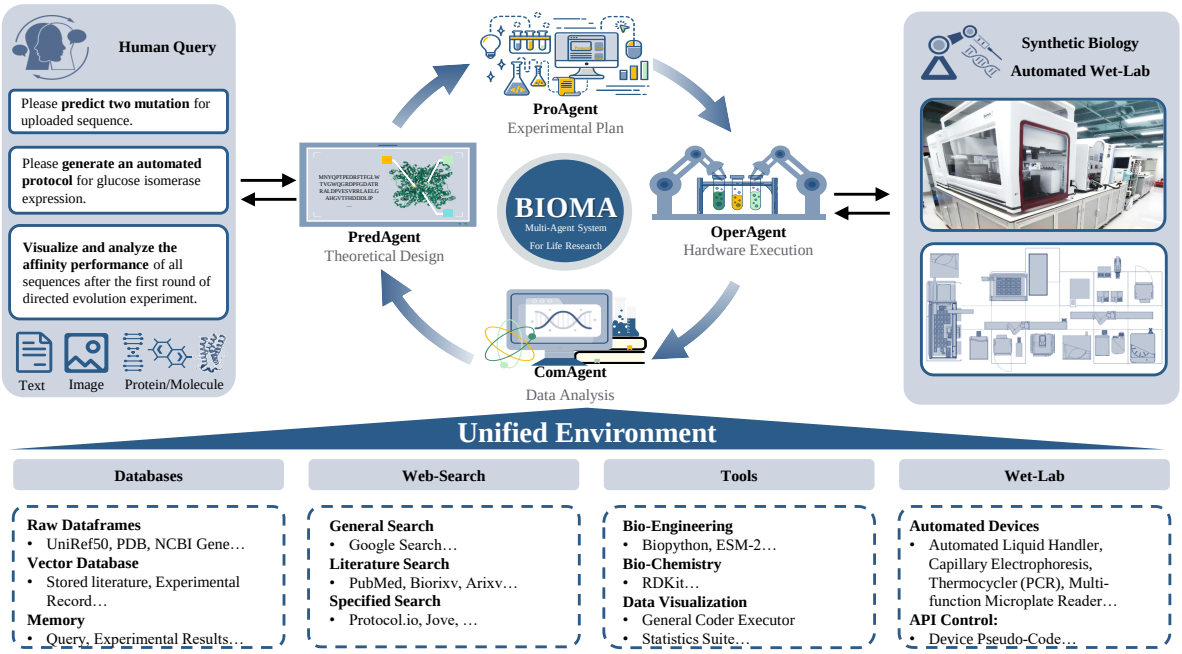


**Figure 4. BIOMA Multi-Agent System.** Coordinated by four types of agents—PredAgent, ProAgent, OperAgent, and ComAgent—this system can accomplish the entire workflow of theoretical prediction, experimental planning, automated unmanned experimentation, and data analysis iteration. This enables an autonomous research closed-loop of "conceive – design – execute – analyze – optimize," significantly enhancing the efficiency of life science research.

### *3.2. Data Silos and Integration Difficulties*

Data silos and integration difficulties constitute a critical bottleneck constraining AI4S, manifesting primarily in the challenges of data interoperability and sharing, the high heterogeneity of multi-source data, and the substantial difficulty of data integration. Against the backdrop of explosive growth in scientific research data, data are dispersed across systems belonging to different institutions, employing different formats, and adhering to different standards, giving rise to the phenomenon of "data silos." This leads to limitations on AI model training, impediments to interdisciplinary research, and reduced efficiency in scientific discovery. In interdisciplinary research in particular, the differences in data standards and semantic systems across disciplines multiply the difficulty of data integration, severely constraining the application potential of AI in scientific discovery. Consequently, resolving the problems of data silos and integration difficulties is of foundational significance for constructing a unified scientific research data infrastructure and enabling data-driven scientific discovery. In response to the data silo problem, the academic and industrial communities are actively building data fusion platforms, improving data governance frameworks, and advancing open sharing mechanisms to achieve the efficient utilization of scientific research data.

#### 3.2.1. Building Cross-Source Heterogeneous Data Fusion Platforms

Building cross-source heterogeneous data fusion platforms is a key strategy for resolving data silo problems. The core mechanism involves mapping heterogeneous scientific research data from multiple domains, formats, and dimensions into a unified semantic space, achieving standardized processing and integrated analysis of data through techniques such as unified representation, feature extraction, and knowledge graph construction, thereby enabling AI systems to uniformly understand and utilize data from diverse sources. This approach also entails establishing rigorous data quality control standards and full lifecycle management systems to cleanse, annotate, and validate scientific research data, ensuring data accuracy, completeness, and consistency.

The OneAstronomy platform of Zhejiang Laboratory, as shown in Figure 5, performs standardized processing, denoising, and feature extraction on multi-band astronomical spectral data, constructing

 doi:10.20944/preprints202607.0102.v1



a unified dataset adapted for AI training [49]. The model adopts the Huber loss function, which combines the advantages of mean squared error (MSE) and mean absolute error (MAE), making it less sensitive to outliers. Its definition is as follows:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \leqslant \delta, \\ \delta\left(|y - f(x)| - \frac{1}{2}\delta\right) & \text{otherwise,} \end{cases} \tag{8}$$

where $y$ is the true value, $f(x)$ is the predicted value, and $\delta$ is a hyperparameter that determines the point at which the loss transitions from quadratic to linear. It uses the squared loss (similar to MSE) for small errors ($|y - f(x)| \leq \delta$, promoting faster convergence with smooth gradients, while transitioning to a linear loss (similar to MAE) for larger errors, providing robustness against outliers [49].

Other platforms are also committed to addressing the problem of heterogeneous data fusion. The GeoGPT geoscience large model led by the same institution [50], designed specifically for the earth science domain, constructs a "digital twin foundation of the Earth" and successfully integrates multi-source heterogeneous spatiotemporal data from high-resolution satellites, ground-based meteorological stations, and ocean buoys [99], providing a high-quality unified data view for flood early warning and carbon sink estimation. SciDataCopilot, the "scientific data navigator," adopts a Data Fabric architecture [51] and employs a virtualization layer to achieve logically unified access to data dispersed across systems such as LIMS, ELN, and HPC, enabling cross-source querying and integration without physical data migration, thereby substantially lowering the threshold for data governance. Polymathic AI's AION-1 [52] converts 200 million heterogeneous astronomical data records into unified representations through a "unified tokenization" scheme. Google AlphaEarth [53] generates unified feature vectors from multi-source Earth observation data, resolving the integration challenges posed by multi-source heterogeneous data in astronomy and earth sciences. BioMapAI [62] and Pangu-Weather [61] construct fusion platforms for cross-domain heterogeneous data sources, by realizing multi-omics clinical mapping and three-dimensional meteorological dimension integration respectively.

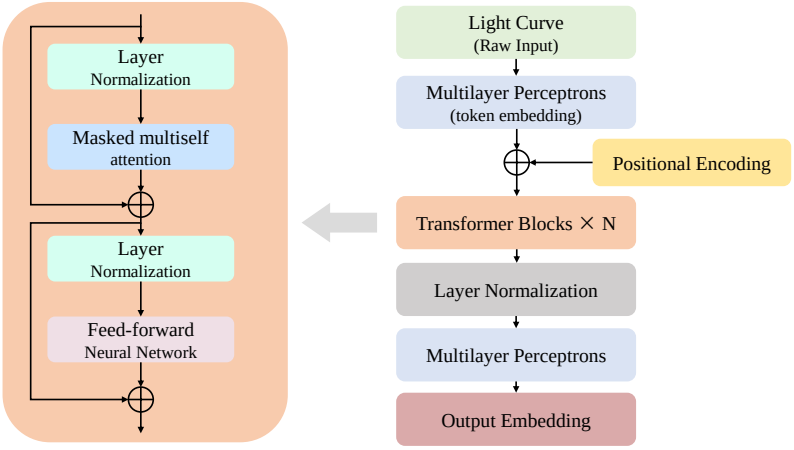


**Figure 5. Architecture of the FALCO foundation model.** This figure illustrates the end-to-end workflow and core modules of FALCO, which replaces discrete token embedding with MLP encoding for continuous time-series data to accommodate the continuous flux properties of astronomical light curves.

### 3.2.2. Establishing Scientific Data Governance Frameworks

Establishing scientific data governance frameworks is a critical mechanism for ensuring data quality. The core of this approach lies in building rigorous data quality control standards and full lifecycle management systems to cleanse, annotate, and validate scientific research data, ensuring data accuracy, completeness, and consistency. This provides high-quality data sources for AI model training,

 doi:10.20944/preprints202607.0102.v1



addressing the problems of uneven data quality and difficulty in extracting effective information from scientific research data.

When the "InternLM" Scientific Discovery Platform integrated over 200 high-value datasets from 50 leading global research institutions, it implemented data validation and cross-validation through a multi-agent framework. SciDataCopilot incorporates a built-in AI metadata engine capable of automatically identifying data types, extracting key parameters, and associating relevant experimental records, while also providing FAIR principles compliance checks, offering automated tools for the full lifecycle governance of scientific research data. The LucaProt model completed standardized processing when integrating metatranscriptomic data, significantly improving the accuracy of RNA virus identification [54]. PaperQA2 [55] simultaneously captures metadata such as DOIs and citation information upon literature ingestion, and employs a "citation graph traversal" tool that uses highly relevant evidence as seeds to expand candidate sets along citation relationships, thereby improving evidence recall and coverage. Furthermore, for the pain points of untraceable evidence and mixed use of multi-version data, establishing data provenance and lineage record abstractions is of paramount importance. For instance, adopting the W3C PROV data model [56] as a cross-system exchangeable provenance abstraction enables detailed recording of source entities, processing activities (such as cleansing and training), and responsible agents for scientific research data, forming chain-level provenance that fundamentally satisfies the audit requirements of high-compliance scientific research scenarios. These practices fully demonstrate the core value of scientific data governance in improving data quality and AI model performance. Biomni [45] and Chemprop [60] focus on establishing standardized governance and specification frameworks for scientific data, relying on automated literature mining pipelines and rigorous data quality control mechanisms.

#### 3.2.3. Open Sharing Mechanisms for Scientific Data

Advancing open sharing mechanisms for scientific data is a key pathway to breaking through data silos. The core mechanism involves achieving joint modeling without data leaving its domain through technologies such as federated learning, thereby resolving issues of data privacy and compliance restrictions. This approach also entails constructing retrieval-augmented generation pipelines that ingest documents from both internal and external organizational sources into a unified repository, enabling the accumulation and reuse of evidence through evidence summarization and scoring.

In retrieval-augmented generation (RAG [57]) pipelines, PaperQA2 [55] requires the model to reason on the basis of generated "evidence summaries and relevance scores" rather than merely concatenating raw text fragments. In the LitQA2 question-answering benchmark, this approach effectively improved the model's precision and accuracy. For addressing the systemic blockages in cross-organizational sharing caused by compliance restrictions, federated learning minimum viable validation has become mainstream. For example, participating parties train locally and upload only gradients, which are then securely aggregated on the central side. The FedAvg [58] algorithm performs distributed training under non-independent and identically distributed (non-IID) data conditions, reducing communication rounds by a substantial factor of 10 to 100 compared to conventional synchronous SGD methods, providing a highly valuable solution for privacy-preserving joint modeling.Its optimization objective can be written as a weighted sum of local client losses:

$$\min_{w\in\mathbb{R}^d} f(w), \qquad f(w) \overset{\text{def}}{=} \sum_{k=1}^{K} \frac{n_k}{n} F_k(w), \qquad F_k(w) = \frac{1}{n_k} \sum_{i\in P_k} f_i(w). \tag{9}$$

Here $w \in \mathbb{R}^d$ denotes the shared global model parameters, $K$ is the number of participating clients, $n_k$ is the number of data samples held by client $k$, $n = \sum_k n_k$ is the total sample count, $P_k$ is the set of data indices local to client $k$, $F_k(w)$ is the empirical loss of client $k$, and $f_i(w)$ is the per-sample loss. This objective clarifies why privacy-preserving scientific data sharing can still support a single global model without forcing raw data to leave each institution. The "InternLM" Scientific Discovery Platform constructs a global scientific research data network, integrating thousands of terabytes of

 doi:10.20944/preprints202607.0102.v1



cross-domain scientific data. The MIT SEAL framework [59] integrates scientific literature from global databases including arXiv and PubMed, building a structured scientific knowledge graph. All of these practices leverage data sharing to achieve cross-domain data fusion, laying a solid data foundation for AI4S. xTrimoPGLM [44] and scBERT [63] are pre-trained on large-scale public scientific datasets. By openly releasing model weights and source codes, they jointly facilitate the open sharing and global circulation of knowledge of scientific data.

### *3.3. Cross-Disciplinary Knowledge Barriers*

Cross-disciplinary knowledge barriers represent another critical challenge confronting AI4S, manifesting primarily in the need for interdisciplinary research to integrate knowledge from multiple domains and the inability of single models to handle complex tasks. As scientific problems become increasingly complex, the knowledge systems of individual disciplines are insufficient to address intricate scientific questions, making interdisciplinary research an inevitable trend. However, the substantial differences in knowledge systems, terminological frameworks, and research methodologies across disciplines give rise to challenges of "knowledge asymmetry," "communication barriers," and "poor model adaptability" in interdisciplinary research. This not only limits the application of AI models to complex scientific problems but also impedes the generation of innovative scientific discoveries [100]. Therefore, breaking down cross-disciplinary knowledge barriers and constructing a unified interdisciplinary knowledge framework are essential for achieving breakthroughs in AI4S. To this end, researchers are committed to developing general–domain fused large models, establishing interdisciplinary collaboration mechanisms, promoting interactive collaboration and slow thinking, and leveraging visualization and artifactization, thereby facilitating the seamless integration and application of knowledge across different disciplines.

#### 3.3.1. Developing General–Domain Fused Large Models

Developing general–domain fused large models is the core strategy for breaking through cross-disciplinary knowledge barriers. The core mechanism involves constructing large models that integrate general artificial intelligence capabilities with domain-specific knowledge [101], using general-purpose foundation models as a base and incorporating professional theories, experimental regularities, and knowledge systems from various disciplines. This enables such models to serve as "translators" of interdisciplinary knowledge, capable of understanding multi-disciplinary language and processing complex cross-domain scientific research tasks, thereby resolving the problem of single models being poorly suited to interdisciplinary research.

The "InternLM" Scientific Discovery Platform employs a general–domain fused large model as the "cognitive brain" of research tasks, integrating general AI capabilities with domain-specific expertise across multiple fields to lower the threshold for interdisciplinary research. The Innovator general scientific foundation model [64], developed by the Beijing Academy of Scientific Intelligence, integrates three scientific modalities—text, structure, and numerical data—and embeds a "scientific fact graph" containing over 50 million entity relationships. This endows the model with powerful cross-disciplinary knowledge reasoning capabilities, enabling it to answer complex cross-domain questions such as "How should radiation-resistant aerospace alloys be designed?" The ESM-3 multimodal generative platform [65] integrates multidimensional professional information on protein sequences, structures, and functions, achieving cross-modal joint reasoning in the field of molecular biology and transcending the disciplinary knowledge limitations of traditional protein design. MatRIS-MoE [68], Genomics-FM [70] and ChemLLM [47] are all committed to the joint pre-training on ultra-large-scale heterogeneous data, so as to realize the internalized integration of interdisciplinary knowledge covering physics, genomics, and chemistry at the fundamental model level.

#### 3.3.2. Establishing Interdisciplinary Collaboration Mechanisms

Establishing interdisciplinary collaboration mechanisms is a key pathway for promoting the deep integration of AI and science. The core mechanism involves building cross-disciplinary research

 doi:10.20944/preprints202607.0102.v1



platforms and collaborative systems that integrate the research methodologies, technical approaches, and knowledge systems of different disciplines, driving the deep fusion of computer science, artificial intelligence, and foundational scientific disciplines to realize collaborative innovation in "AI + science."

The "Machine Chemist" platform [66] developed at the University of Science and Technology of China integrates multidisciplinary knowledge from chemistry, computer science, and automation and control, constructing a tripartite methodology of "chemical brain + robotic experimentation + intelligent optimization," as illustrated in Figure 6 The AI4S LAB platform integrates the four core elements of AI4S—computing power, models, data, and experiments—building a collaborative bridge between life sciences and AI technologies, and realizing cross-disciplinary linkage between theoretical prediction and experimental validation [40]. The SciMaster general-purpose scientific research agent [37], by incorporating multiple built-in domain expert roles (such as a "condensed matter physics assistant"), enables non-specialist users to conveniently access and apply knowledge in specific domains, effectively constructing a new paradigm of human–machine collaborative interdisciplinary cooperation. These practices have effectively broken down the disciplinary barriers of traditional scientific research and have driven the cross-disciplinary integration of fields. Biomni [45] and ShizishanGPT [69] focus on constructing agent collaboration platforms capable of autonomously scheduling multi-disciplinary professional tools and knowledge resources. Through modular collaborative mechanisms, such platforms break down cognitive and operational barriers across diverse research domains.

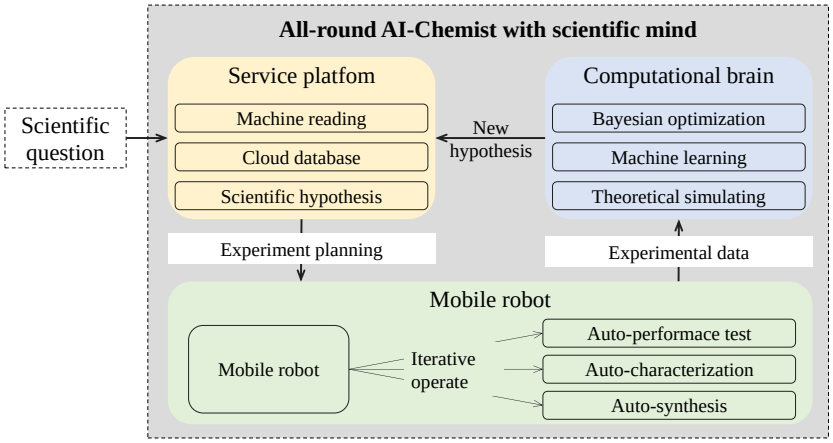


**Figure 6. All-round AI-Chemist with scientific mind.** This figure fully illustrates the top-level architecture, core module composition, and full closed-loop working mechanism of the AI-Chemist system platform, providing a visual representation of the trinitarian methodology of "chemical brain + robotic experimentation + intelligent optimization".

### 3.3.3. Interactive Collaboration and Slow Thinking

Interactive collaboration and slow thinking represent a key pathway for promoting the deep integration of AI and science. The core mechanism involves designing "requirement clarification" as a fixed process, using multi-turn dialogue to delineate research boundaries and parameters, thereby resolving the problem of repeated rework caused by unclear requirements. Complex operations are compressed into a finite set of pseudo-functions, and static checking is employed to ensure executability, resolving the problem that natural language cannot be directly executed.

BioPlanner [20] transforms natural language descriptions into standardized experimental protocols through a "constrained domain-specific language (DSL) / pseudo-function library." LLaMP employs a hierarchical ReAct agent to enable natural language-driven materials screening. AI Co-Scientist [29] adopts a "goal–hypothesis–parameter locking" process and achieves collaborative reasoning through dedicated agents for generation, reflection, ranking, and evolution. Kosmos implements "collaborative artifact automation," generating structured artifacts at each iteration to ensure auditability.

 doi:10.20944/preprints202607.0102.v1



To make the protocol-oriented evaluation landscape around BioPlanner more concrete, Table 3 summarizes representative biology-focused benchmarks highlighted in the project source slides.

**Table 3.** Representative biology protocol benchmarks.

| Benchmark | Domain | Scale | Primary Task Focus |
|---|---|---|---|
| BioLP-bench | Math / physics / biology | 200 | ErrorsQA for cross-domain scientific reasoning. |
| BioPlanner | Molecular biology | 100 | Retrieval-based protocol planning evaluation. |
| ProBio | Biology | 3,724 | Protocol question answering. |
| LAB-Bench | Biology | 135 | Laboratory protocol question answering. |
| BioProBench | 20+ biological domains | 27k protocols / 556k instances | Protocol QA, step ordering, error correction, protocol generation, and protocol reasoning. |

#### 3.3.4. Visualization and Artifactization

Visualization and artifactization are effective mechanisms for improving the efficiency of cross-disciplinary collaboration. The core mechanism involves mandating that each iteration produce structured artifacts that are bound to data and code versions, thereby resolving the problem that verbally established consensus is difficult to consolidate. Kosmos stipulates that each iteration must produce structured artifacts including problem statements, hypotheses, evidence lists, tool invocation summaries, conclusions, and open items. BioPlanner employs "self-consistency verification" to ensure that generated experimental protocols can directly drive liquid-handling robots.

#### 3.3.5. Formal Verification of Critical Processes

Formal verification of critical processes is an important mechanism for eliminating boundary ambiguities in cross-disciplinary collaboration. When collaborating across disciplines, boundary conditions established through verbal agreement are highly susceptible to being overlooked. For high-risk processes involving multi-agent coordination or cross-system concurrent scheduling, formal methods can be applied for verification. For instance, TLA+ [67] can be used to extract concise specifications, and the TLC model checker can be employed for exhaustive verification to ensure the absolute correctness of critical invariants—such as "no write without authorization" and "eventual consistency conditions"—thereby providing rigorous logical proofs under conditions of concurrency and complex collaboration.

### *3.4. Information Credibility and Knowledge Update*

Information credibility and knowledge update represent significant challenges in AI4S, manifesting primarily in issues of information reliability (particularly in high-stakes domains such as healthcare), untimely updates of scientific literature, and insufficient credibility of algorithmic outputs. In scientific discovery, the accuracy and timeliness of information directly bear on the reliability of research outcomes. However, AI models are typically trained on historical data and struggle to incorporate the latest scientific knowledge in a timely manner; furthermore, the credibility of model outputs is difficult to verify when handling complex scientific problems. In high-stakes domains such as healthcare in particular, unreliable information can lead to serious consequences. Moreover, the rapid advancement of scientific knowledge imposes requirements of continuous learning and knowledge updating on AI models [102]. Consequently, addressing the challenges of information credibility and knowledge update is of great importance for ensuring the reliability of AI-driven scientific discovery and for facilitating the timely dissemination and application of scientific knowledge. Researchers are tackling this core challenge through approaches such as constructing dynamic trusted closed loops, improving algorithmic interpretability, establishing scientific fact anchors, and enabling continuous learning and knowledge internalization.

 doi:10.20944/preprints202607.0102.v1



### 3.4.1. Constructing Dynamic Trusted Closed Loops

Constructing dynamic trusted closed loops is the core mechanism for improving the credibility of AI models. The fundamental approach involves decomposing answer and conclusion generation into two stages: first generating an "evidence package," and then deriving conclusions on the basis of this evidence package, thereby ensuring that all conclusions are traceable to documented evidence. This entails establishing a dynamic error correction and credibility verification system following the cycle of "data input–model output–validation feedback–model optimization," employing multi-dimensional validation, fact-checking, and autonomous falsification to verify the outputs of AI models.

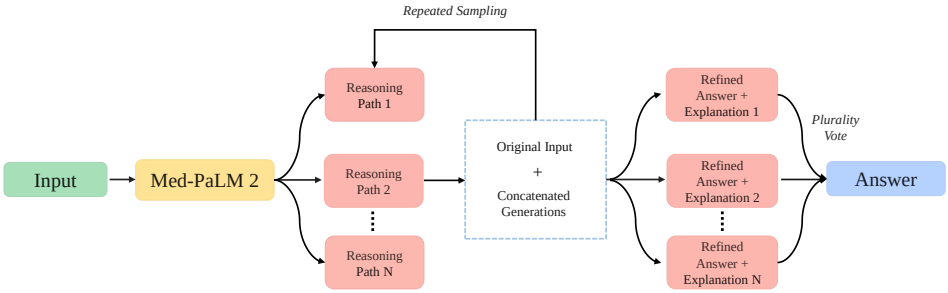


**Figure 7. Ensemble Refinement (ER) with Med-PaLM 2.** The Med-PaLM model proposes Ensemble Refinement (ER) to enhance the reasoning capability of large medical models. It takes multiple possible reasoning paths generated by itself as conditions to refine and improve the final answer.

The Med-PaLM model [71] introduces a chained retrieval and verification mechanism to fact-check medical knowledge outputs, as illustrated in Figure 7, thereby improving the credibility of clinical decision-making recommendations. AlphaTensor ensures the correctness of matrix multiplication algorithms through rigorous mathematical verification [72], endowing AI-discovered research findings with scientific rigor and credibility. Pipelines based on agentic RAG, as exemplified by PaperQA2, further demonstrate the critical role of "evidence summarization and scoring" in constructing trusted closed loops. After retrieval, this system performs "contextual summarization and relevance scoring" on each piece of evidence, storing high-scoring evidence summaries in the session state for repeated invocation during subsequent reasoning and writing. Simultaneously, the system mandatorily requires that the generation stage reason on the basis of these evidence summaries rather than merely concatenating raw text fragments, thereby achieving evidence consolidation and reuse, and significantly improving the traceability and accuracy of generated conclusions. DeepRare [41] and ChemAgent [48] adopt self-reflective loops or self-updating memory banks, and guarantee the real-time credibility of reasoning chains via dynamic verification mechanisms. During the execution of long-chain scientific tasks, addressing error propagation is equally critical to constructing trusted closed loops. By introducing runtime error capture and self-correction mechanisms, one can draw on the distributed tracing concepts of OpenTelemetry to structure agent tool invocations and execution logs; when a task fails, errors are precisely categorized, and for recoverable errors, iterative "reflect–revise–retry" cycles based on frameworks such as Reflexion and Self-Refine [18,73] are automatically triggered. Self-Refine formalizes this iterative update process as follows:

$$y_{t+1} = \mathcal{M}(p_{\text{refine}} \| x \| y_0 \| fb_0 \| \cdots \| y_t \| fb_t). \tag{10}$$

Here $\mathcal{M}$ is the frozen language model, $x$ is the original task input, $y_t$ is the model output at iteration $t$, $fb_t$ is the textual feedback generated by the same model on $y_t$, $p_{\text{refine}}$ is the refinement prompt, and $\|$ denotes string concatenation. This recurrence makes the reflect–revise–retry loop explicit and explains how trusted closed loops can improve outputs without updating model weights. This self-improvement pathway, which requires no updates to model weights, effectively avoids the pain point of requiring expert manual debugging for long-chain tasks and ensures the dynamic trustworthiness of the system.

 doi:10.20944/preprints202607.0102.v1



#### 3.4.2. Improving Algorithmic Interpretability

Improving algorithmic interpretability is a key mechanism for resolving the "black box" problem of AI. The core of this approach lies in developing interpretable AI algorithms—such as rule-based algorithms, decision tree algorithms, and knowledge graph associations—to make the decision-making process and reasoning logic of AI models transparent, thereby enhancing scientists' acceptance of and trust in AI results.

The Med-PaLM model associates diverse medical concepts through a medical knowledge graph, achieving "clinical interpretability" for the medical large model and ensuring that diagnostic recommendations are explicitly grounded in a knowledge base. The "Machine Chemist" platform's predictive model combines machine learning with physical models, allowing its catalytic activity prediction results to be explained through chemical principles, thereby avoiding the uninterpretability of purely data-driven models. Introducing contradiction detection and negative evidence retrieval processes transforms the black-box reasoning of models into a structured "dialectical process." For instance, PaperQA2 [55] explicitly incorporates negative queries in the retrieval stage, performs contradiction detection on candidate evidence, and treats this as a primary output. This requires the model, when generating conclusions, to clearly explain "why evidence A is adopted rather than evidence B." This practice of explicitly presenting conflicting evidence alongside the rationale for the adopted evidence makes the decision logic of the AI transparent to scientists, substantially improving algorithmic interpretability and result credibility in complex knowledge integration tasks. DeepRare [41] provides traceable and verifiable transparent reasoning chains linked to authoritative medical evidence, ensuring the credibility of information by improving algorithmic interpretability and establishing scientific fact anchors. In addition to the aforementioned content, RoseTTAFold [80], Chemprop [60], BioMapAI [62], scBERT [63] and the AlphaFold series [81,82] transform black-box models into traceable scientific insights through architectural transparency and feature attribution.

#### 3.4.3. Establishing Scientific Fact Anchors

Establishing scientific fact anchors is the core mechanism for ensuring consistency between AI models and scientific rigor. The fundamental approach involves using rigorously validated classical theories, experimental conclusions, and authoritative data from various disciplines as scientific fact anchors, treating them as boundary constraints for AI model training and generation. This prevents models from producing outputs that violate scientific laws, ensuring information credibility from the source and resolving the tension between AI models and scientific rigor.

The AlphaGeometry series uses formal logic and geometric theorems as fact anchors, verifying model-generated proof processes through a symbolic reasoning engine to ensure the scientific validity of geometric theorem proofs [74]. Both the "PanGu" large model and the GeoGPT geoscience large model adopt physics-informed neural network (PINN) technology, embedding fundamental equations from geoscience and physics—such as the Navier–Stokes equations and the laws of thermodynamics—as soft constraints into the models, ensuring that they do not violate basic scientific facts when simulating extreme scenarios. The Hermite drug design platform uses foundational theories from chemistry and biology as anchors, integrating AI with physical modeling to improve the scientific validity and accuracy of predictions regarding drug molecule–target binding [75]. The Multiagent Debate [76] framework introduces a "multi-agent debate review loop." This framework generates multiple candidate answers and their corresponding evidence chains in parallel for the same question, engages multiple agents in multi-round debate over the sufficiency of evidence and the soundness of reasoning chains, and finally has an adjudicator synthesize the results according to established scientific rules (such as evidence-level priority). This process has been demonstrated to effectively improve model reasoning capabilities and reduce the risk of hallucinations. ShizishanGPT [69], ProteinMPNN [84] and RFdiffusion [83] build a rigorous factual foundation for scientific content generation by leveraging knowledge graph anchoring and physical structural constraints.

 doi:10.20944/preprints202607.0102.v1



#### 3.4.4. Enabling Continuous Learning and Knowledge Internalization

Enabling continuous learning and knowledge internalization is the key mechanism for addressing the problem of lagging knowledge updates in AI models. The core of this approach involves building a continuous learning system for models that incorporates the latest research findings, literature data, and experimental conclusions into model training in real time, consolidating new information into model weights through methods such as self-editing and meta-learning, thereby achieving proactive knowledge updating and internalization, and resolving the problem of AI models relying on historical data with outdated knowledge [103].

The MIT SEAL framework achieves continuous learning and knowledge internalization through a dual-loop optimization system, consolidating information from new scientific literature into model weights. The Hermite drug design platform undergoes continuous upgrades on a weekly basis, integrating the latest research findings to maintain the timeliness of drug development knowledge. Med-PaLM incorporates the latest medical research through continual pre-training, preventing outdated information from misleading clinical decision-making. The Darwin–Gödel Discovery Machine (DGDM) proposes a "dual-loop" architectural concept: its inner loop performs constrained evolution on candidate solutions to execute specific tasks, while its outer loop conducts risk-controlled self-adaptation of the pipeline itself. This mechanism has been demonstrated to simultaneously improve binding energy and maintain 100% chemical validity in concrete examples such as molecular docking. DP-GEN [77] explicitly organizes the active learning closed loop into three fixed phases: exploration, labeling, and training. This workflow uses the uncertainty or divergence in model outputs as a trigger condition to prioritize the computation and labeling of regions with the greatest model extrapolation risk, thereby obtaining reliable potential energy surface models with minimal reference data and effectively resisting performance degradation caused by distributional shifts in data. The NEP series [78] (e.g., MACE [92], MatRIS [68]) focuses on the underlying internalization of physical laws and the continuous iteration of knowledge systems through large-scale data pre-training.

### *3.5. Simulation and Control of Complex Extreme Scientific Research Scenarios*

The simulation and control of complex extreme scientific research scenarios represent a formidable challenge in AI4S, manifesting in highly difficult problems in extreme settings such as Mars environment simulation, nuclear fusion reactor control, and geometric theorem proving. These scenarios are typically characterized by high risk, high complexity, and high uncertainty, making them difficult to address effectively with conventional methods. For instance, Mars environment simulation requires reproducing Martian conditions in terrestrial laboratories, yet the extreme and non-reproducible nature of the Martian environment renders experimentation exceptionally difficult; plasma instability in nuclear fusion reactors requires precise real-time control, otherwise experimental interruption or even equipment damage may result; and geometric theorem proving requires handling highly abstract mathematical structures that conventional methods struggle to address effectively. Resolving this challenge is of great significance for advancing the application of AI to high-risk, high-complexity scientific problems and for achieving breakthrough scientific discoveries. Confronting the challenges posed by complex extreme scientific research scenarios, researchers are developing high-fidelity AI simulation–control closed loops, adopting synthetic data-driven strategies, and integrating multi-scale unified modeling frameworks to achieve effective simulation and control of extreme scenarios.

#### 3.5.1. Developing High-Fidelity AI Simulation–Control Closed Loops

Developing high-fidelity AI simulation–control closed loops is the core mechanism for resolving control problems in complex extreme scientific research scenarios. The fundamental approach involves constructing a high-fidelity real-time closed-loop system following the cycle of "simulate–predict–adjust–verify," combining deep learning, reinforcement learning, and related technologies to perform precise simulation and real-time perception of complex extreme scenarios, anticipate risks

 doi:10.20944/preprints202607.0102.v1



and changes within such scenarios in advance, and dynamically adjust control parameters, thereby achieving a transition from "post-hoc verification" to "real-time control."

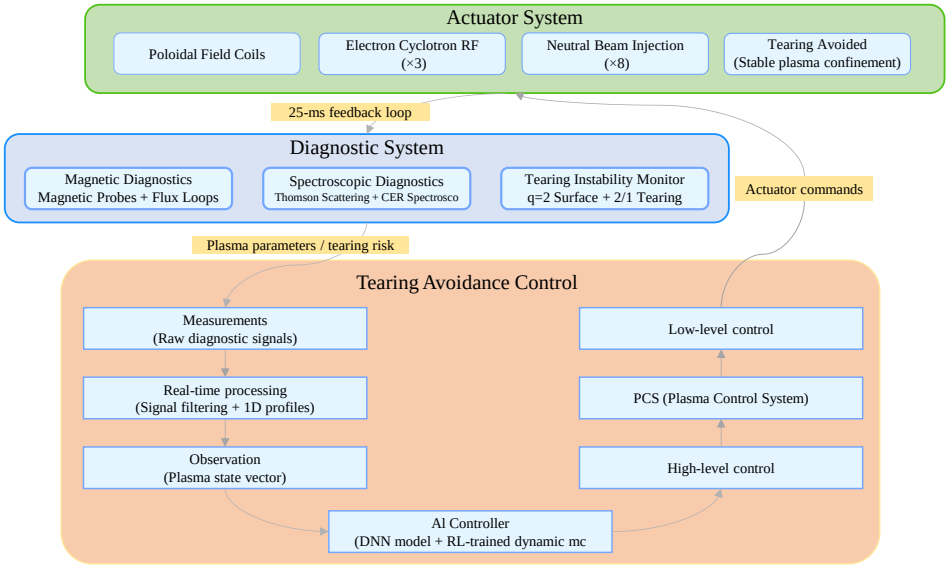


**Figure 8.** This figure fully presents the overall architecture and full-link working mechanism of the tearing instability avoidance system in the DIII-D tokamak.

The Princeton AI controller integrates real-time sensor data from a tokamak device and uses deep reinforcement learning to predict plasma tearing instability 300 milliseconds in advance, dynamically adjusting beam power and plasma shape to upgrade the control paradigm of nuclear fusion experiments from "reactive firefighting" to "proactive hazard prevention," resolving the extreme scenario control challenges of nuclear fusion reactors [85]. Its workflow is illustrated in Figure 8. The design of the reward function is crucial in guiding the AI controller to learn optimal strategies. Its form is as follows:

$$R(\beta_{\mathrm{N}}, T; k) = \begin{cases} \beta_{\mathrm{N}}, & \text{if } T < k \\ k - T, & \text{otherwise} \end{cases} \tag{11}$$

Here, $R$ represents the reward obtained by the AI controller. This function aims to incentivize the AI controller to maximize the normalized plasma pressure $\beta_{\mathrm{N}}$ while ensuring that the tearing instability $T$ remains below an acceptable threshold $k$. Should $T$ reach or exceed $k$, the controller will be penalized, thereby prompting it to prioritize measures for reducing instability [85].

The Bohr Scientific Research Space Station developed by DP Technology leverages its high-precision AI force field model to successfully simulate the extreme chemical reaction environment involved in oxygen production from Martian meteorites, providing critical materials design solutions for future deep-space exploration. Physics-informed neural networks (PINNs) [26] directly incorporate the residuals of partial differential equations (PDEs) or physical laws into the loss function of the neural network and compute the required derivatives through automatic differentiation.A standard PINN training objective combines data fidelity and physics residual penalties:

$$\begin{aligned} \mathrm{MSE} &= \mathrm{MSE}_u + \mathrm{MSE}_f, \\ \mathrm{MSE}_u &= \frac{1}{N_u} \sum_{i=1}^{N_u} \left| u(t_u^i, x_u^i) - u^i \right|^2, \\ \mathrm{MSE}_f &= \frac{1}{N_f} \sum_{i=1}^{N_f} \left| f(t_f^i, x_f^i) \right|^2. \end{aligned} \tag{12}$$

 doi:10.20944/preprints202607.0102.v1



Here $\mathrm{MSE}_u$ measures data fidelity against $N_u$ labeled observation points $\{(t_u^i, x_u^i, u^i)\}$, where $u(t,x)$ is the neural-network solution surrogate; $\mathrm{MSE}_f$ penalizes the PDE residual $f(t,x)$ evaluated at $N_f$ collocation points, enforcing the governing physics. Requiring only a small number of observation points combined with a large number of collocation points for training, PINNs can solve forward problems and perform high-fidelity parameter inversion with minimal data. GraphCast [7] models global weather prediction as a neural operator surrogate model, learning a "function-to-function" mapping from initial and boundary conditions to solution fields, thereby bypassing the computationally expensive iterative physical equation-solving process of traditional numerical solvers at inference time. The system adopts an autoregressive recursive architecture with a 6-hour time step, taking the current atmospheric state field as conditional input and generating continuous spatiotemporal sequences through rolling prediction, completing 10-day global forecasts in under one minute of inference time, thereby providing an extremely fast engineering foundation for large-scale ensemble forecasting. Chemma [43] and DSSs [91] embed intelligent agents into the real-time loops of physical experiments and crop simulators, achieving autonomous optimization and closed-loop regulation for complex scientific research scenarios.

#### 3.5.2. Adopting Synthetic Data-Driven Strategies

Adopting synthetic data-driven strategies is an effective approach for addressing data scarcity in extreme scientific research scenarios. The core mechanism targets the problem of scarce and difficult-to-obtain real experimental data in extreme scientific research scenarios by employing AI technologies to generate high-fidelity synthetic data that fills training data gaps. This enables models to learn the regularities and characteristics of extreme scenarios from synthetic data, resolving the bottlenecks of simulation and control caused by data scarcity.

The AlphaGeometry series pioneered the use of synthetic data for geometric theorem proving, generating 100 million theorem-proving steps as training data, thereby resolving the problems of human geometric proofs being difficult to convert into machine-verifiable formats and the scarcity of training data. FourCastNet adopts the Fourier Neural Operator (FNO) [6,27], parameterizing kernel functions in the frequency domain to achieve resolution-invariant inference capabilities.The core Fourier layer can be written as:

$$(K(\phi)v_t)(x) = \mathcal{F}^{-1}\big(R_\phi \cdot (\mathcal{F}v_t)\big)(x), \qquad x \in D. \tag{13}$$

Here $v_t$ is the function representation at layer $t$, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, $R_\phi$ is a learnable weight tensor applied in the frequency domain (parameterized by $\phi$), and $D$ is the spatial domain. By performing the convolution as a pointwise multiplication in the frequency domain, the operator achieves resolution-invariant inference. Its week-scale forecasts can be generated in seconds—orders of magnitude faster than traditional solvers—thereby making large-scale ensemble forecasting feasible.

Furthermore, xTrimoPGLM [44], RFdiffusion [83] and ProteinMPNN [84] focus on constructing synthetic data-driven strategies for the robust generation of extreme protein structures via large-scale knowledge internalization and physical noise augmentation.

#### 3.5.3. Integrating Multi-Scale Unified Modeling Frameworks

Integrating multi-scale unified modeling frameworks is a key mechanism for resolving simulation problems in complex extreme scientific research scenarios. The core of this approach lies in incorporating the microscopic and macroscopic laws of extreme scientific research scenarios, along with research objects at different scales, into a unified modeling framework. By fusing physical models, machine learning models, and other model types, it captures latent correlations in multi-dimensional data, ensures the robustness and accuracy of complex system simulations, and realizes full-scale simulation from the molecular level to macroscopic systems.

 doi:10.20944/preprints202607.0102.v1



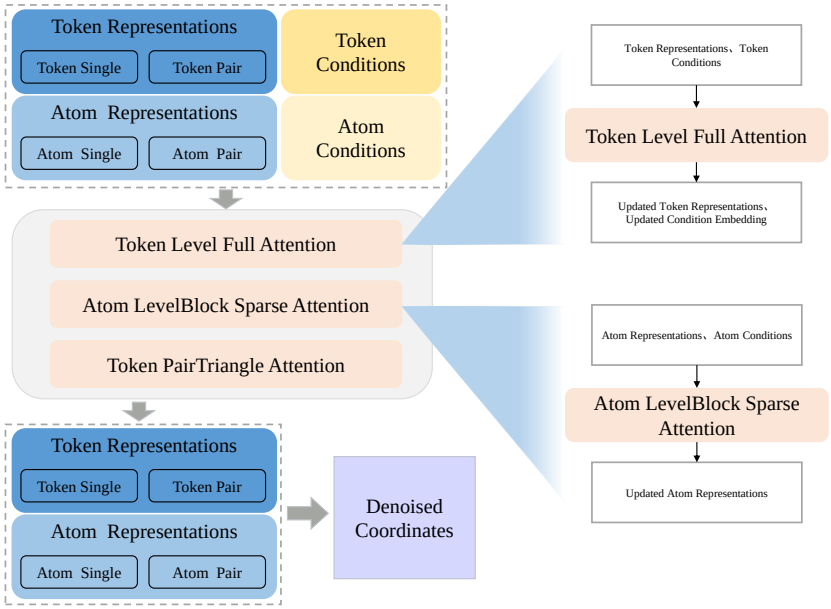


**Figure 9. Architecture of the Pallatom-Ligand Transformer.** The Pallatom-Ligand Transformer architecture employs a triple-attention module pipeline to jointly update coarse-grained token representations and fine-grained atom representations, enabling end-to-end denoising generation of all-atom 3D coordinates for protein–ligand complexes under conditional constraints.

The "Machine Chemist" platform fuses machine learning (random forests and deep learning) with physical models (density functional theory) to construct a multi-scale "chemical brain" predictive model, successfully simulating the extreme chemical scenario of oxygen production from Martian meteorites and reducing the development cycle for Martian oxygen-producing catalysts from 2,000 years to 6 weeks. The Pallatom-Ligand all-atom diffusion model [86] performs information fusion at the amino acid and atomic levels, achieving precise all-atom simulation and design of proteins, and resolving the multi-scale modeling challenges of complex molecular scenarios. The method employs a tripe-attention module, as depicted in Figure 9. Its core idea is illustrated by the following equation:

$$a = a + \text{Layernorm}(\text{SegmentMean}(q)). \tag{14}$$

Here, $a$ represents the token-level coarse-grained representation of the protein-ligand system. $\text{Layernorm}(\text{SegmentMean}(q))$ denotes the feature obtained by averaging the atomic-level fine-grained features $q$ across token segments and then applying layer normalization. The residual fusion of these two features enables efficient transfer of fine-grained atomic information to the coarse-grained token representation.

The GeoGPT geoscience large model fuses macroscopic surface information acquired from satellite remote sensing with microscopic meteorological data collected from ground stations, constructing a unified climate simulation framework spanning from the global to the local scale, and has been successfully applied to high-accuracy extreme weather prediction. CorrDiff [87] adopts a two-step strategy in which a mean model first outputs a baseline prediction, after which the low-resolution prediction is used as conditional input for a diffusion model to learn residuals and perform bias correction.This decomposition is expressed directly in the original paper as:

$$\mathbf{x} = \underbrace{\mathbb{E}[\mathbf{x} \mid \mathbf{y}]}_{:=\boldsymbol{\mu}\ \text{(regression)}} + \underbrace{(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}])}_{:=\mathbf{r}\ \text{(generation)}}. \tag{15}$$

Here $\mathbf{x}$ is the high-resolution target field, $\mathbf{y}$ is the coarse-resolution input, $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x} \mid \mathbf{y}]$ is the deterministic regression component learned by a mean model, and $\mathbf{r}$ is the stochastic residual captured by a conditional diffusion model. This formulation clarifies why CorrDiff separates baseline regression

 doi:10.20944/preprints202607.0102.v1



from residual generation before bias correction. This approach successfully downscales coarse-grained predictions at the 25 km scale to generate high-resolution detail fields at 2 km, effectively recovering the tail risks of extreme events and local spatial detail. Stormer [88] proposes a training strategy of randomized prediction targets (training across different time intervals) and combines multiple prediction results at inference time. This approach, which upgrades single deterministic outputs to distributional outputs, significantly improves the accuracy at the target forecast lead time and enhances system reliability. MSA Transformer [89], RoseTTAFold [80], MLIPs [90] (e.g., MACE), Pangu-Weather [61] and the AlphaFold series [81,82] reconstruct multi-scale correlations of scientific systems within a unified modeling framework, through multi-dimensional integration and multi-track architectural design.

## 4. Future Trends of AI for Science

With the rapid advancement of artificial intelligence, AI for Science is evolving from task-specific optimization toward the construction of systematic scientific research capabilities. In the future, AI is expected not only to enhance scientific data analysis and complex system modeling, but also to participate more deeply in scientific problem formulation, experimental design, and research workflow orchestration. This trend will drive scientific research toward greater automation, intelligence, and collaboration. From this perspective, the future development of AI4S can be discussed along seven major directions.

**Intelligent research systems for workflow-level collaboration.** Traditional scientific research often requires researchers to switch frequently among multiple software tools, computational platforms, and experimental environments, with the overall workflow heavily dependent on manual coordination. In the future, with the development of large language models and research agents, AI is likely to play a more central role in problem definition, task planning, and workflow organization. On the one hand, AI scientist systems may assist researchers by analyzing large-scale scientific literature and research data to identify promising questions and generate research hypotheses. On the other hand, intelligent research platforms may automatically decompose complex scientific tasks and coordinate data analysis, model training, and experimental validation, thereby enabling end-to-end research workflows. As tool interfaces become increasingly standardized and research platform ecosystems mature, AI-centered collaborative research systems may gradually emerge to support the full process from problem formulation to experimental verification.

**Interdisciplinary knowledge modeling and scientific knowledge graphs.** Modern scientific research is increasingly interdisciplinary, and many important problems require the integration of theories and methods from multiple domains. In areas such as materials design, climate science, and the life sciences, research questions often involve knowledge from physics, chemistry, biology, and computational science simultaneously. Consequently, interdisciplinary knowledge integration and unified representation will remain a major direction for AI4S. In the future, scientific knowledge modeling based on knowledge graphs and large models is expected to become more mature, enabling the structured organization of scientific literature, experimental results, and theoretical models within unified knowledge networks. At the same time, AI systems may develop stronger scientific reasoning capabilities, allowing them to generate hypotheses and perform theory-informed inference based on existing knowledge. Moreover, with the continued progress of multimodal models, scientific knowledge representation will increasingly extend beyond text to include images, molecular structures, and experimental data, further improving AI's ability to reason over complex scientific problems.

**Open and shared scientific data infrastructures.** Scientific data constitute a fundamental basis for AI-driven discovery. However, in the current research environment, large volumes of scientific data remain scattered across institutions and repositories, with substantial differences in data standards and semantic schemas. These issues continue to hinder data-driven scientific research. As a result, building open and shared scientific data infrastructures will be a key trend in AI4S. Future scientific data platforms are likely to move toward greater standardization and interoperability through unified data formats, metadata specifications, and data governance mechanisms. At the same time, advances

 doi:10.20944/preprints202607.0102.v1



in privacy-preserving computation and federated learning may enable broader data sharing while maintaining security and privacy. In addition, multimodal scientific data platforms that integrate experimental, simulation, and observational data are expected to provide richer resources for AI models and further promote data-centric scientific discovery.

**Trustworthy scientific foundation models.** Large-scale foundation models have achieved remarkable success in natural language processing and computer vision, and similar trends are emerging in scientific research. Scientific foundation models trained on large-scale scientific data are becoming an important direction in domains such as protein structure prediction, climate modeling, and materials discovery. Looking forward, the development of such models will place increasing emphasis on trustworthiness and interpretability. On the one hand, researchers are likely to further explore approaches that combine scientific mechanisms with data-driven learning, such as physics-informed neural networks and mechanism-constrained learning, to improve scientific consistency. On the other hand, model interpretability and uncertainty quantification will become increasingly important for ensuring that AI-generated scientific predictions and inferences are reliable and verifiable. With continued progress in cross-domain foundation models, future AI systems may eventually support more general scientific problem solving across multiple disciplines.

**Mechanism-aware AI for scientific discovery.** Beyond prediction and optimization, an important future direction of AI4S is to support the discovery of scientific mechanisms. Rather than treating AI-generated results as isolated predictions, future systems should transform data-driven patterns into interpretable, testable, and reusable scientific knowledge. This process may involve learning scientific representations from heterogeneous data, extracting stable patterns from predictive models, formulating mechanism hypotheses with the support of domain knowledge and literature evidence, and validating these hypotheses through simulation, high-throughput experimentation, or autonomous laboratories. In this way, AI4S can move from accelerating scientific workflows toward generating mechanistic understanding that guides future prediction, intervention, and discovery.

**Intelligent computing systems for extreme-scale scientific computation.** Many frontier scientific problems, including climate system simulation, plasma dynamics, and cosmic evolution, require large-scale numerical simulation and complex system modeling. These applications place extremely high demands on computational resources, making efficient scientific computing infrastructure a critical foundation for AI4S. In the future, scientific computing is likely to evolve toward a deeply integrated computing paradigm that combines high-performance computing (HPC) with AI computing. On the one hand, GPUs, AI accelerators, and specialized architectures can significantly improve the efficiency of scientific model training and simulation. On the other hand, AI itself can be used to optimize computational processes, for example by accelerating numerical solvers or serving as surrogate models for traditional simulations, thereby reducing overall computational cost. In addition, continued progress in cloud and distributed computing may provide researchers with more flexible access to large-scale computing resources and broaden the accessibility of advanced scientific computation.

**Self-driving laboratories and closed-loop scientific discovery.** Experimental validation is a core component of scientific research, and advances in robotics and laboratory automation are enabling AI to play an increasingly important role in this stage. In the future, self-driving laboratories are expected to become a major paradigm in AI4S. Under this paradigm, AI systems can automatically design experimental protocols based on model predictions, execute experiments through robotic platforms, and then use the resulting observations to update the underlying models, thereby forming a closed-loop workflow of "model prediction–experimental validation–model updating." Early progress has already been demonstrated in areas such as materials discovery and chemical synthesis. As experimental robotics and real-time control systems continue to improve, self-driving experimental platforms are likely to be adopted across a wider range of scientific domains, substantially increasing the efficiency of scientific discovery.

 doi:10.20944/preprints202607.0102.v1



## References


1. Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60. https://doi.org/10.1038/s41586-023-06221-2.
2. Zhao, T.; Wang, S.; Ouyang, C.; Chen, M.; Liu, C.; Zhang, J.; Yu, L.; Wang, F.; Xie, Y.; Li, J.; et al. Artificial intelligence for geoscience: Progress, challenges, and perspectives. *The Innovation* **2024**, *5*.
3. Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J.C.; Yang, S. Artificial intelligence in drug development. *Nature medicine* **2025**, *31*, 45–59.
4. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *nature* **2021**, *596*, 583–589.
5. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556* **2022**, [2211.02556]. https://doi.org/10.48550/arXiv.2211.02556.
6. Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214* **2022**, [arXiv:physics.ao-ph/2202.11214]. https://doi.org/10.48550/arXiv.2202.11214.
7. Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421.
8. Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Reviews of Modern Physics* **2019**, *91*, 045002.
9. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.W.; et al. Artificial intelligence: A powerful paradigm for scientific research. *The innovation* **2021**, *2*.
10. Berens, P.; Cranmer, K.; Lawrence, N.D.; von Luxburg, U.; Montgomery, J. AI for science: an emerging agenda. *arXiv preprint arXiv:2303.04217* **2023**, [2303.04217]. https://doi.org/10.48550/arXiv.2303.04217.
11. Davenport, T.H.; Bean, R. Five trends in AI and data science for 2025. *MIT Sloan Management Review (Online)* **2025**, pp. 1–4.
12. Hey, A.J.; Tansley, S.; Tolle, K.M.; et al. *The fourth paradigm: data-intensive scientific discovery*; Vol. 1, Microsoft research Redmond, WA, 2009.
13. Ioannidis, Y. The 5th Paradigm: AI-Driven Scientific Discovery. *Communications of the ACM* **2024**, *67*, 5–5. https://doi.org/10.1145/3702970.
14. Miolane, N. The fifth era of science: Artificial scientific intelligence. *PLOS Biology* **2025**, *23*, e3003230. https://doi.org/10.1371/journal.pbio.3003230.
15. Carter, J.; Feddema, J.; Kothe, D.; Neely, R.; Pruet, J.; Stevens, R.; Balaprakash, P.; Beckman, P.; Foster, I.; Iskra, K.; et al. Advanced research directions on ai for science, energy, and security: Report on summer 2022 workshops **2023**.
16. Stevens, R.; Taylor, V.; Nichols, J.; Maccabe, A.B.; Yelick, K.; Brown, D. Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science. Technical report, Argonne National Laboratory (ANL), Argonne, IL (United States), 2020.
17. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the Proceedings of the 11th International Conference on Learning Representations (ICLR). OpenReview.net, 2023.
18. Shinn, N.; Cassano, F.; Labash, B.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
19. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* **2023**, [2304.05376]. https://doi.org/10.48550/arXiv.2304.05376.
20. O'Donoghue, O.; Shtedritski, A.; Ginger, J.; Abboud, R.; Ghareeb, A.; Rodriques, S. BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2023, pp. 2676–2694.

 doi:10.20944/preprints202607.0102.v1




21. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8.
22. Wang, D.; Tan, Z.; Gao, J.; Zhang, S.; Shen, J.; Lu, Y. AI4Protein: transforming the future of protein design. *Science China Life Sciences* **2025**, *68*, 2880–2890. https://doi.org/10.1007/s11427-024-2906-3.
23. Jiang, X.; Zhang, M.; Luo, X.; Yu, Z.; Meng, Y.; Wang, D. FiberKAN: Kolmogorov–Arnold Networks for Nonlinear Fiber Optics. *Journal of Lightwave Technology* **2025**, *43*, 7083–7100. https://doi.org/10.1109/jlt.2025.3571017.
24. Jacobsson, T.J. AI-Generated Hypotheses and the Emergence of Autonomous Scientific Discovery. *ACS Materials Letters* **2026**. https://doi.org/10.1021/acsmaterialslett.6c00224.
25. Davies, A.; Veličković, P.; Buesing, L.; Blackwell, S.; Zheng, D.; Tomašev, N.; Tanburn, R.; Battaglia, P.; Blundell, C.; Juhász, A.; et al. Advancing mathematics by guiding human intuition with AI. *Nature* **2021**, *600*, 70–74.
26. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* **2019**, *378*, 686–707.
27. Li, Z.; Kovachki, N.B.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2021.
28. Lu, C.; Lu, C.; Lange, R.T.; Foerster, J.; Clune, J.; Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* **2024**, [2408.06292]. https://doi.org/10.48550/arXiv.2408.06292.
29. Gottweis, J.; Weng, W.H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* **2025**, [2502.18864]. https://doi.org/10.48550/arXiv.2502.18864.
30. Szymanski, N.J.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.; Cubuk, E.D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **2023**, *624*, 86–91. https://doi.org/10.1038/s41586-023-06734-w.
31. Zheng, T.; Deng, Z.; Tsang, H.T.; Wang, W.; Bai, J.; Wang, Z.; Song, Y. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2025, p. 17744–17761. https://doi.org/10.18653/v1/2025.emnlp-main.895.
32. Dai, T.; Vijayakrishnan, S.; Szczypiński, F.T.; Ayme, J.F.; Simaei, E.; Fellowes, T.; Clowes, R.; Kotopanov, L.; Shields, C.E.; Zhou, Z.; et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **2024**, *635*, 890–897.
33. Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178–184. https://doi.org/10.1016/j.cpc.2018.03.016.
34. Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In Proceedings of the International Conference on Learning Representations, 2023.
35. Zhang, D.; Bi, H.; et al. Pretraining of attention-based deep learning potential model for molecular simulation. *npj Computational Materials* **2024**. https://doi.org/10.1038/s41524-024-01278-7.
36. Ghareeb, A.E.; et al. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400* **2025**, [2505.13400]. https://doi.org/10.48550/arXiv.2505.13400.
37. Chai, J.; Tang, S.; Ye, R.; Du, Y.; Zhu, X.; Zhou, M.; Wang, Y.; E, W.; Zhang, Y.; Zhang, L.; et al. SciMaster: Towards General-Purpose Scientific AI Agents, Part I. X-Master as Foundation: Can We Lead on Humanity's Last Exam? *arXiv preprint arXiv:2507.05241* **2025**, [2507.05241]. https://doi.org/10.48550/arXiv.2507.05241.
38. Shanghai Artificial Intelligence Laboratory. Intern-Discovery Scientific Discovery Platform. Available online at: https://discovery-home.intern-ai.org.cn/, n.d. Accessed: 17 March 2026.
39. Zhang, L.; Chen, S.; Cai, Y.; Chai, J.; Chang, J.; Chen, K.; Chen, Z.X.; Ding, Z.; Du, Y.; Gao, Y.; et al. Bohrium + SciMaster: Building the Infrastructure and Ecosystem for Agentic Science at Scale **2025**. https://doi.org/10.48550/ARXIV.2512.20469.
40. AI4S LAB, Peking University Shenzhen Graduate School. AI4S LAB: Intelligent Automation Laboratory for Life Sciences, 2026. Accessed: 17 March 2026.

 doi:10.20944/preprints202607.0102.v1




41. Zhao, W.; Wu, C.; Fan, Y.; Qiu, P.; Zhang, X.; Sun, Y.; Zhou, X.; Zhang, S.; Peng, Y.; Wang, Y.; et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature* **2026**, *651*, 775–784. https://doi.org/10.1038/s41586-025-10097-9.
42. Du, Y.; Yu, B.; Liu, T.; Shen, T.; Chen, J.; Rittig, J.G.; Sun, K.; Zhang, Y.; Krishnan, A.; Zhang, Y.; et al. Accelerating Scientific Discovery with Autonomous Goal-evolving Agents **2026**. [arXiv:cs.AI/2512.21782].
43. Zhang, Y.; Han, Y.; Chen, S.; Yu, R.; Zhao, X.; Liu, X.; Zeng, K.; Yu, M.; Tian, J.; Zhu, F.; et al. Large language models to accelerate organic chemistry synthesis. *Nature Machine Intelligence* **2025**, *7*, 1010–1022. https://doi.org/10.1038/s42256-025-01066-y.
44. Chen, B.; Cheng, X.; Li, P.; Geng, Y.a.; Gong, J.; Li, S.; Bei, Z.; Tan, X.; Wang, B.; Zeng, X.; et al. xTrimoPGLM: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nature Methods* **2025**, *22*, 1028–1039. https://doi.org/10.1038/s41592-025-02636-z.
45. Huang, K.; Zhang, S.; Wang, H.; Qu, Y.; Lu, Y.; Roohani, Y.; Li, R.; Qiu, L.; Li, G.; Zhang, J.; et al. Biomni: A General-Purpose Biomedical AI Agent **2025**. https://doi.org/10.1101/2025.05.30.656746.
46. Guerrero, P.; Hašan, M.; Sunkavalli, K.; Měch, R.; Boubekeur, T.; Mitra, N.J. MatFormer: a generative model for procedural materials. *ACM Transactions on Graphics* **2022**, *41*, 1–12. https://doi.org/10.1145/3528223.3530173.
47. Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.; et al. ChemLLM: A Chemical Large Language Model **2024**. https://doi.org/10.48550/ARXIV.2402.06852.
48. Tang, X.; Hu, T.; Ye, M.; Shao, Y.; Yin, X.; Ouyang, S.; Zhou, W.; Lu, P.; Zhang, Z.; Zhao, Y.; et al. ChemAgent: Self-updating Library in Large Language Models Improves Chemical Reasoning. In Proceedings of the 13th International Conference on Learning Representations, (ICLR), 2025, pp. 52724 – 52760. https://doi.org/10.48550/ARXIV.2501.06590.
49. Zuo, X.; Tao, Y.; Huang, Y.; Kang, Z.; Chen, H.; Cui, C.; Pan, J.; Kong, X.; Ting, Y.S.; Tang, X.; et al. FALCO: Foundation Model of Astronomical Light Curves for Time Domain Astronomy. *The Astronomical Journal* **2026**, *171*, 10. https://doi.org/10.3847/1538-3881/ae1467.
50. Huang, F.; Wu, F.; Zhang, Z.; Wang, Q.; Zhang, L.; Boquet, G.M.; Chen, H. GeoGPT-RAG Technical Report. Technical report, Zhejiang Lab, 2025, [arXiv:cs.IR/2509.09686].
51. Rao, J.; Qiu, Y.; Zhang, J.; Deng, J.; Sun, S.; Ling, F.; Chen, H.; Dong, N.; Gao, Z.; Sun, S.; et al. SciDataCopilot: An Agentic Data Preparation Framework for AGI-driven Scientific Discovery. *arXiv preprint arXiv:2602.09132* **2026**, [arXiv:cs.DB/2602.09132].
52. Parker, L.; Lanusse, F.; Shen, J.; Liu, O.; Hehir, T.; et al. AION-1: Omnimodal Foundation Model for Astronomical Sciences. *arXiv preprint arXiv:2510.17960* **2025**, [arXiv:astro-ph.IM/2510.17960].
53. Brown, C.F.; Kazmierski, M.R.; Pasquarella, V.J.; et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291* **2025**, [2507.22291]. Version 2, cs.CV, Google DeepMind, https://doi.org/10.48550/arXiv.2507.22291.
54. Shi, M.; Li, Z.; et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* **2024**, *187*, 5696–5710.e12. https://doi.org/10.1016/j.cell.2024.09.009.
55. Skarlinski, M.D.; Cox, S.; Laurent, J.M.; Braza, J.D.; Hinks, M.; Hammerling, M.J.; Ponnapati, M.; Rodriques, S.G.; White, A.D. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740* **2024**.
56. Moreau, L.; Missier, P.; et al. PROV-DM: The PROV Data Model. W3c recommendation, World Wide Web Consortium (W3C), 2013.
57. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, Inc., 2020, Vol. 33, pp. 9459–9474.
58. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, 2017, Vol. 54, pp. 1273–1282.
59. Zweiger, A.; Pari, J.; Guo, H.; Akyürek, E.; Kim, Y.; Agrawal, P. Self-Adapting Language Models. In Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025), Massachusetts Institute of Technology, 2025; [arXiv:cs.LG/2506.10943].
60. Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackermann, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702. https://doi.org/10.1016/j.cell.2020.01.021.

 doi:10.20944/preprints202607.0102.v1




61. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. https://doi.org/10.1038/s41586-023-06185-3.
62. Xiong, R.; Aiken, E.; Caldwell, R.; Vernon, S.D.; Kozhaya, L.; Gunter, C.; Bateman, L.; Unutmaz, D.; Oh, J. BioMapAI: Artificial Intelligence Multi-Omics Modeling of Myalgic Encephalomyelitis / Chronic Fatigue Syndrome **2024**. https://doi.org/10.1101/2024.06.24.600378.
63. Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; Yao, J. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **2022**, *4*, 852–866. https://doi.org/10.1038/s42256-022-00534-z.
64. Liao, N.; Wang, X.; Lin, Z.; et al. Innovator: Scientific Continued Pretraining with Fine-grained MoE Upcycling. *arXiv preprint arXiv:2507.18671* **2025**, [arXiv:cs.LG/2507.18671]. https://doi.org/10.48550/arXiv.2507.18671.
65. Hayes, T.; Rao, R.; Akin, H.; et al. Simulating 500 million years of evolution with a language model. *Science* **2025**, *387*, 850–858. https://doi.org/10.1126/science.ads0018.
66. Zhu, Q.; Zhang, F.; Huang, Y.; et al. An all-round AI-Chemist with a scientific mind. *National Science Review* **2022**, *9*, nwac190. https://doi.org/10.1093/nsr/nwac190.
67. Lamport, L. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*; Addison-Wesley, 2002; pp. 1–83.
68. Zhou, Y.; Wang, H.; Du, Y.; Wang, Y.; Li, M.; Hu, S.; Zhang, X.; Liu, W.; Wang, C.; Guo, Z.; et al. Breaking the Training Barrier of Billion-Parameter Universal Machine Learning Interatomic Potentials **2026**. https://doi.org/10.48550/ARXIV.2604.15821.
69. Yang, S.; Liu, Z.; Mayer, W.; Ding, N.; Wang, Y.; Huang, Y.; Wu, P.; Li, W.; Li, L.; Zhang, H.Y.; et al. ShizishanGPT: An Agricultural Large Language Model Integrating Tools and Resources. In Proceedings of the Web Information Systems Engineering – WISE 2024. Springer Nature Singapore, 2025, pp. 284–298.
70. Ye, P.; Bai, W.; Ren, Y.; Li, W.; Qiao, L.; Liang, C.; Wang, L.; Cai, Y.; Sun, J.; Yang, Z.; et al. Genomics-FM: Universal Foundation Model for Versatile and Data-Efficient Functional Genomic Analysis **2024**. https://doi.org/10.1101/2024.07.16.603653.
71. Singhal, K.; Azizi, S.; Tu, T.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.
72. Fawzi, A.; Balog, M.; Huang, A.; et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **2022**, *610*, 47–53. https://doi.org/10.1038/s41586-022-05172-4.
73. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, Inc., 2023, Vol. 36, pp. 46534–46594.
74. Trinh, T.H.; Wu, Y.; Le, Q.V.; He, H.; Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **2024**, *625*, 476–482. https://doi.org/10.1038/s41586-023-06747-5.
75. DP Technology. Hermite®: New-Generation AI-Powered Drug Design Platform, 2024. Accessed: 17 March 2026; AI4S-enabled CADD tool for rational drug discovery.
76. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
77. Zhang, L.; Lin, D.Y.; Wang, H.; Car, R.; E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials* **2019**, *3*, 023804.
78. Lai, S.; Yang, X.; Shi, J.; Liu, S.; Guo, Y.; Yan, L.; Zang, J.; Zhang, Z.; Jia, Q.; Sun, J.; et al. Bulk hexagonal diamond. *Nature* **2026**, *651*, 621–625. https://doi.org/10.1038/s41586-026-10212-4.
79. Zhou, X.; Sun, L.; He, D.; Guan, W.; Wang, G.; Wang, R.; Wang, L.; Yuan, X.; Sun, X.; Zhang, Y.; et al. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell* **2026**, *44*, 777–791. https://doi.org/10.1016/j.ccell.2026.01.019.
80. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. https://doi.org/10.1126/science.abj8754.
81. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**. https://doi.org/10.1101/2021.10.04.463034.

 doi:10.20944/preprints202607.0102.v1




82. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.
83. Watson, J.L.; Juergens, D.; Bennett, N.R.; Trippe, B.L.; Yim, J.; Eisenach, H.E.; Ahern, W.; Borst, A.J.; Ragotte, R.J.; Milles, L.F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.
84. Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R.J.; Milles, L.F.; Wicky, B.I.M.; Courbet, A.; de Haas, R.J.; Bethel, N.; et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49–56. https://doi.org/10.1126/science.add2187.
85. Seo, J.; Kim, S.; Jalalvand, A.; et al. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature* **2024**, *626*, 746–751. https://doi.org/10.1038/s41586-024-07024-9.
86. Wang, H.; Wang, Q.; Ma, R.; Guan, J.; Wu, W.; Wang, H.; Dou, J. PALLATOM-LIGAND: An All-Atom Diffusion Model for Designing Ligand-Binding Proteins. In Proceedings of the Proceedings of the 14th International Conference on Learning Representations (ICLR 2026), 2026.
87. Mardani, M.; Brenowitz, N.; Cohen, Y.; Pathak, J.; Chen, C.Y.; Cheng, C.K.; Kashinath, K.; Pritchard, M.; et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *COMMUNICATIONS EARTH & ENVIRONMENT* **2025**, *6*, 124. https://doi.org/10.1038/s43247-025-01424-5.
88. Nguyen, T.; Shah, R.; Bansal, H.; Arcomano, T.; Maulik, R.; Kotamarthi, V.; Foster, I.; Madireddy, S.; Grover, A. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2024, Vol. 37, pp. 68740–68771.
89. Wen, J.; Yang, S.; Jiang, L.; Shi, Y.; Huang, Z.; Li, P.; Xiong, H.; Yu, Z.; Zhao, X.; Xu, B.; et al. Accelerated discovery of high-performance small-molecule hole transport materials via molecular splicing, high-throughput screening, and machine learning. *Journal of Materials Informatics* **2025**, *5*. https://doi.org/10.20517/jmi.2024.102.
90. Wang, L.; Zhou, X.; Luo, Z.; Liu, S.; Yue, S.; Chen, Y.; Liu, Y. Review of External Field Effects on Electrocatalysis: Machine Learning Guided Design. *Advanced Functional Materials* **2024**, *34*. https://doi.org/10.1002/adfm.202408870.
91. Chen, D.; Huang, Y. Integrating Reinforcement Learning and Large Language Models for Crop Production Process Management Optimization and Control through A New Knowledge-Based Deep Learning Paradigm. *Computers and Electronics in Agriculture* **2025**, *232*, 110028. https://doi.org/https://doi.org/10.1016/j.compag.2025.110028.
92. Batatia, I.; Kovács, D.P.; Simm, G.N.C.; Ortner, C.; Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2022; NIPS '22.
93. Yoshikawa, N.; Skreta, M.; Darvish, K.; Arellano-Rubach, S.; Ji, Z.; Bjørn Kristensen, L.; Li, A.Z.; Zhao, Y.; Xu, H.; Kuramshin, A.; et al. Large language models for chemistry robotics. *Autonomous Robots* **2023**, *47*, 1057–1086.
94. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.
95. Cui, H.; Xu, Y.; Pang, K.; Li, G.; Gong, F.; Wang, B.; Li, B. LUMI-lab: a foundation model-driven autonomous platform enabling discovery of new ionizable lipid designs for mRNA delivery. *BioRxiv* **2025**, pp. 2025–02.
96. ScienceOne. ScienceOne Scientific Foundation Model Platform, 2026. Accessed: 17 March 2026.
97. Zhang, J.; Jiang, J.; et al. A closed-loop architecture with knowledge-of-results feedback for neural-symbolic planning. *Knowledge-Based Systems* **2025**. https://doi.org/10.1016/j.knosys.2025.114041.
98. Docker Inc.. *Docker Engine Security and Resource Constraints*, 2024. Official Documentation.
99. Zhejiang Lab. GeoGPT: A Non-Profit Geoscience Exploration and Research Platform, 2026. Accessed: 17 March 2026.
100. Li, G.; Deng, L.; Tang, H.; Pan, G.; Tian, Y.; Roy, K.; Maass, W. Brain-Inspired Computing: A Systematic Survey and Future Trends. *Proceedings of the IEEE* **2024**, *112*, 544–584. https://doi.org/10.1109/JPROC.2024.3429360.
101. He, Y.H. AI-driven research in pure mathematics and theoretical physics. *Nature Reviews Physics* **2024**, *6*, 546–553. https://doi.org/10.1038/s42254-024-00740-1.

 doi:10.20944/preprints202607.0102.v1



102. Messeri, L.; Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **2024**, *627*, 49–58. https://doi.org/10.1038/s41586-024-07146-0.
103. Yang, R. Unlearning as Ablation: Toward a Falsifiable Benchmark for Generative Scientific Discovery **2025**. https://doi.org/10.48550/ARXIV.2508.17681.