Beyond Tool Use: Multimodal Distillation and the Evolution of Neural Networks toward AI Scientists

Daozheng Qu [*1] **Yanfei Ma** [*2]

Abstract

AI scientists are frequently regarded as a transition from passive instruments to autonomous research entities. This research contends that the fundamental transformation is mechanistic: neural networks have progressed from task-specific predictors to multimodal distillation systems adept at integrating scientific information, compressing expert knowledge, and facilitating discovery workflows. Scientific AI is transitioning along the spectrum from AlphaFold-style predictions to agentic systems that design experiments, utilize tools, and compose research results. We offer a conceptual framework that connects three stages: predictive neural networks, foundation models, and multimodal distillation-driven scientific agents. We contend that the autonomy of AI scientists should be assessed not solely by the quality of their outputs, but also by their capabilities in hypothesis formulation, experimental design, verification, attribution, and human oversight.

1. Introduction

Artificial intelligence in science is evolving from a passive role to an active participation in research methodologies. Historically, neural networks mostly operated as task-specific predictors or optimization elements. Modern systems now enable structure prediction, materials discovery, experimental design, tool application, and research communication. AlphaFold and AlphaFold 3 have revolutionized biological structure prediction, GNoME has accelerated neural materials discovery, and Coscientist, in conjunction with autonomous laboratories, illustrates that AI can systematically design and execute components of experimental science (Jumper et al., 2021; Abramson et al., 2024; Merchant et al., 2023; Boiko et al., 2023; Szymanski et al., 2023). These improvements suggest that the central question has transitioned from whether AI can assist in research to the degree to which it can go from a simple tool to a collaborator and finally to an autonomous scientific contributor.

This research asserts that the transition signifies not only a product-level advancement from chatbots to agents but also a significant fundamental transformation in neural representation learning. Neural networks have evolved from specialized models to foundational models that encompass broad representations spanning language, vision, molecules, code, mathematics, and scientific data (Bommasani et al., 2021; Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Yan et al., 2024a). Multimodal learning enables the integration of varied information, unifying textual hypotheses, visual evidence, symbolic formulas, experimental methods, molecular structures, and external resources into coherent thinking processes.

The principal mechanism facilitating this transformation is multimodal distillation. Distillation has progressed beyond simple model compression; it now serves as a mechanism for imparting expert knowledge, cross-modal evidence, procedural demonstrations, tool-use behavior, and reasoning traces into reusable neural representations (Wei et al., 2022; Kojima et al., 2022; Hsieh et al., 2023; Yao et al., 2023). Through this process, neural networks transition from isolated predictors to sophisticated scientific interfaces proficient in integrating knowledge, reasoning, and action.

This evolution is apparent in the advancement of emerging AI scientific systems. Coscientist demonstrates chemical reasoning and experimental implementation through language models using external tools (Boiko et al., 2023); autonomous laboratories amalgamate machine learning, robotics, and closed-loop optimization for material synthesis (Szymanski et al., 2023); and AlphaGeometry and AlphaProof-type systems showcase progress in formal reasoning and systematic problem-solving (Trinh et al., 2024). Regardless of age differences, these systems demonstrate a shared progression: scientific AI is transitioning from just answering questions to actively participating in research processes.

[1]Department of Computer Science, University of Liverpool, Liverpool, UK [2] Department of Computer Science, Fairleigh Dickinson University, Vancouver, Canada . Correspondence to: Daozheng Qu <d.qu@student.fdu.edu>, Yanfei Ma <y.ma@student.fdu.edu>.

Preprint. June 5, 2026.

This transition also raises governance and evaluation challenges. AI systems that generate hypotheses, design experiments, use tools, analyze results, and write reports go beyond traditional software assistance. Therefore, scientific AI should be evaluated not only by output quality, but also by hypothesis generation, experimental planning, verification, reproducibility, attribution, accountability, and human oversight.

This paper makes three contributions. First, it presents a neural-mechanistic perspective on the evolution from predictive neural networks to multimodal scientific agents. Second, it proposes multimodal distillation as a key mechanism connecting tool-oriented AI and collaborative scientific systems. Third, it argues that AI scientist autonomy should be evaluated according to workflow participation, including whether a system acts as a tool, collaborator, or autonomous contributor.

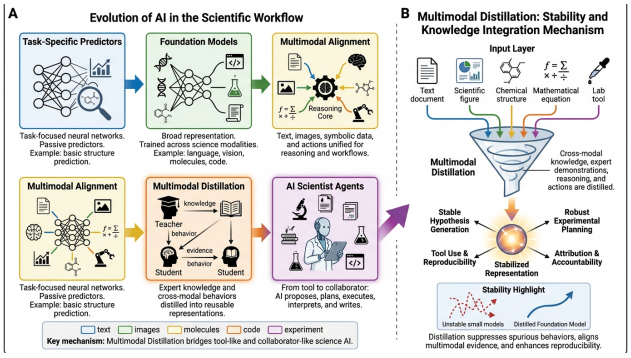

Figure 1. Conceptual overview of the mechanistic evolution from neural prediction to AI scientist agents.

Figure 1 illustrates our proposed view. Scientific AI is evolving from task-specific predictors into foundation models, multimodal alignment systems, distillation-driven representations, and AI scientist agents.

2. Related Work

2.1. AI for Science and Autonomous Discovery

AI for Science has evolved from task-specific prediction toward large-scale scientific modeling and autonomous discovery. AlphaFold and AlphaFold 3 advanced biological structure prediction, while GNoME accelerated neural materials discovery (Jumper et al., 2021; Abramson et al., 2024; Merchant et al., 2023). Beyond prediction, Coscientist and autonomous laboratories demonstrate increasing AI participation in experimental planning, tool use, and closed-loop scientific workflows (Boiko et al., 2023; Szymanski et al., 2023). However, most existing work focuses on performance and discovery speed rather than the multimodal reasoning mechanisms underlying the transition from predictive tools to collaborative scientific systems.

2.2. Foundation Models and Multimodal Representation Learning

Foundation models provide transferable representations across language, vision, code, and structured scientific data (Bommasani et al., 2021). CLIP, Flamingo, BLIP-2, and LLaVA further demonstrated cross-modal alignment and vision-language reasoning capabilities (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Yan et al., 2025b). Such multimodal learning is particularly important for scientific AI because research workflows integrate papers, figures, equations, protocols, molecular structures, code, and experimental results.

2.3. Reasoning, Tool Use, and Agentic Scientific Systems

Recent advances in reasoning and tool use enable neural systems to perform multi-step reasoning and interact with external resources. Chain-of-thought prompting, zero-shot reasoning, and ReAct demonstrate that language models can expose intermediate reasoning traces and coordinate actions with tools (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023). These capabilities are increasingly visible in AI-for-science systems such as Coscientist and autonomous laboratories (Boiko et al., 2023; Szymanski et al., 2023), while AlphaGeometry and AlphaProof-style systems highlight progress in structured mathematical reasoning (Trinh et al., 2024).

2.4. Knowledge Distillation and Multimodal Distillation

Knowledge distillation originally focused on transferring outputs from large teacher models to smaller student models, but recent approaches increasingly transfer reasoning traces, demonstrations, and tool-use behaviors (Hsieh et al., 2023). Combined with chain-of-thought reasoning and ReAct-style interaction (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023; Yan et al., 2024b), distillation can compress reasoning and action trajectories into reusable neural representations. Unlike standard knowledge distillation or multimodal representation learning, we define *multimodal distillation* as a workflow-oriented mechanism that integrates cross-modal scientific evidence, reasoning traces, expert demonstrations, and tool-use behavior into stable representations for scientific workflow participation.

2.5. Autonomy, Attribution, and Governance of AI Scientists

As AI systems increasingly participate in scientific workflows, questions of autonomy, attribution, and governance become more important. Systems such as Coscientist, autonomous laboratories, and formal reasoning frameworks

already blur the boundary between tools and collaborators (Boiko et al., 2023; Szymanski et al., 2023; Trinh et al., 2024). We argue that AI scientist systems should be evaluated not only by output quality, but also by workflow-level abilities including hypothesis generation, experimental design, verification, reproducibility, attribution clarity, accountability, and human oversight.

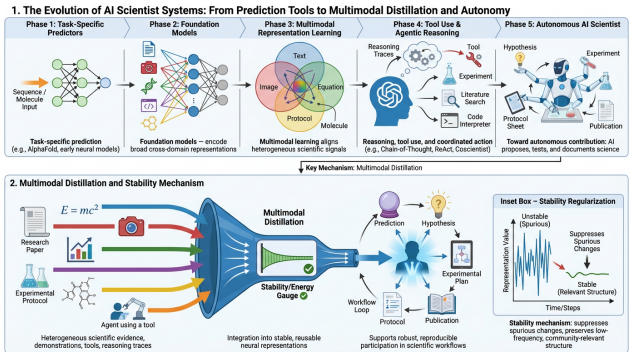

Figure 2. Conceptual framework of the evolution from prediction-oriented neural systems to autonomous AI scientist agents.

Figure 2 illustrates a conceptual trajectory through which neural systems evolve from prediction-oriented models toward autonomous AI scientist agents. Early neural networks were primarily designed for supervised prediction tasks, focusing on pattern recognition and statistical inference. The emergence of foundation models significantly expanded representation capacity and generalization ability, enabling models to learn transferable knowledge across diverse domains.

From this perspective, the development of AI scientist systems is not a discontinuous technological leap but rather the result of cumulative advances in representation learning, multimodal reasoning, tool utilization, and autonomous decision-making.

3. Methodology: A Neural-Mechanistic Framework for AI Scientist Systems

This paper adopts a conceptual and mechanistic methodology. Rather than proposing a single new model architecture, we analyze the evolution of AI scientist systems as a transition in neural learning mechanisms: from task-specific prediction, to foundation model representation, to multimodal alignment, to multimodal distillation, and finally to agentic scientific autonomy. The goal is to provide a structured framework for explaining how neural systems move from tool-like assistance toward collaborator-like participation in scientific workflows.

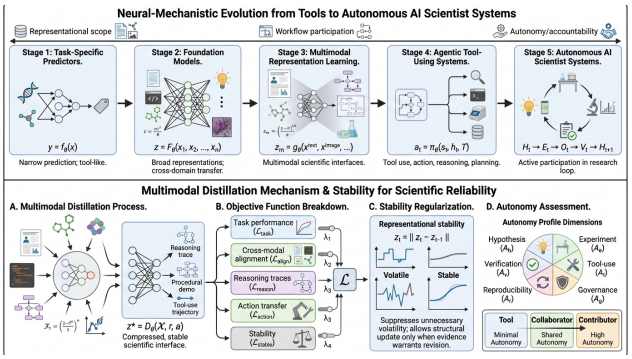

Figure 3. Neural-mechanistic framework for the evolution from tool-oriented AI systems to autonomous AI scientist systems.

3.1. Analytical Scope

We define an *AI scientist system* as a neural or neuro-symbolic system that participates in one or more stages of the scientific workflow, including literature understanding, hypothesis generation, experimental planning, tool use, data interpretation, verification, protocol generation, and manuscript drafting. Under this definition, systems differ not only by their output quality, but also by their functional role in the research process.

We analyze AI scientist systems along three dimensions:

1. **Representational scope**: whether the system operates on a single modality or integrates text, images, equations, molecular structures, protocols, code, and laboratory actions.

2. **Workflow participation**: whether the system only predicts outputs, supports human decisions, coordinates tools, or participates in closed-loop scientific discovery.

3. **Autonomy and accountability**: whether the system requires direct human instruction at every step, collaborates with human researchers, or independently proposes, tests, and documents scientific claims.

This scope allows us to compare AI systems not only as models, but as participants in scientific workflows.

3.2. Five-Stage Evolutionary Framework

We model the evolution of AI scientist systems as five stages:

$$\mathcal{S}_{AI} : P_\theta \rightarrow F_\theta \rightarrow M_\theta \rightarrow A_\theta \rightarrow R_\theta, \qquad (1)$$

where P_θ denotes task-specific predictors, F_θ denotes foundation models, M_θ denotes multimodal representation sys-

tems, A_θ denotes agentic tool-using systems, and R_θ denotes autonomous research-oriented AI scientist systems.

Stage 1: Task-specific predictors. Early neural scientific systems mainly learn narrow mappings:

$$y = f_\theta(x), \tag{2}$$

where x is a scientific input, such as a sequence, molecule, image, or numerical state, and y is a predicted property, class, structure, or score. These systems are useful as scientific tools, but they usually do not participate in hypothesis generation, experimental planning, or workflow-level reasoning.

Stage 2: Foundation models. Foundation models expand neural learning from narrow prediction to broad representation. Instead of learning one task-specific mapping, they learn general representations:

$$z = F_\theta(x_1, x_2, \ldots, x_n), \tag{3}$$

where inputs may span language, code, molecular structures, scientific images, tables, equations, and other scientific data. This stage enables transfer, adaptation, and cross-domain reuse.

Stage 3: Multimodal representation learning. Scientific reasoning is inherently multimodal. A research problem may require connecting a paper, a figure, a molecular graph, an equation, a protocol, and an experimental result. We represent multimodal alignment as:

$$\begin{aligned} z_m = g_\theta(&x^{\text{text}}, x^{\text{image}}, x^{\text{equation}}, \\ &x^{\text{molecule}}, x^{\text{protocol}}). \end{aligned} \tag{4}$$

where z_m is a shared representation that integrates heterogeneous scientific evidence. This stage transforms neural systems from single-input predictors into scientific representation interfaces.

Stage 4: Tool use and agentic reasoning. Agentic systems extend representation learning into action. A tool-using scientific agent can be represented as:

$$a_t = \pi_\theta(s_t, h_t, T), \tag{5}$$

where s_t is the current scientific state, h_t is the reasoning history, T is the available tool set, and a_t is the selected action, such as searching literature, calling a code interpreter, querying a database, planning an experiment, or revising a hypothesis.

Stage 5: Autonomous AI scientist systems. At the highest stage, the system participates in a research loop:

$$H_t \to E_t \to O_t \to V_t \to H_{t+1}, \tag{6}$$

where H_t is a hypothesis, E_t is an experiment or computational test, O_t is the observed result, V_t is verification or evaluation, and H_{t+1} is the revised hypothesis. This loop captures the transition from passive prediction to active scientific contribution.

3.3. Multimodal Distillation as the Key Mechanism

We define *multimodal distillation* as the process by which heterogeneous scientific signals, expert demonstrations, reasoning traces, and tool-use behaviors are compressed into stable reusable neural representations. Unlike conventional knowledge distillation, which mainly transfers outputs from a teacher model to a student model, multimodal distillation transfers both knowledge and workflow behavior.

Let the scientific evidence set be:

$$\begin{aligned} \mathcal{X} = \{&x^{\text{text}}, x^{\text{figure}}, x^{\text{equation}}, \\ &x^{\text{molecule}}, x^{\text{code}}, x^{\text{protocol}}, x^{\text{tool}}\}. \end{aligned} \tag{7}$$

A multimodal distillation process maps this heterogeneous evidence into a stable representation:

$$z^* = D_\theta(\mathcal{X}, r, a), \tag{8}$$

where r denotes reasoning traces and a denotes tool-use or experimental actions. The objective is not only to preserve predictive information, but also to preserve scientific structure, procedural knowledge, and reproducible workflow behavior.

We formulate this objective as:

$$\begin{aligned} \mathcal{L}_{\text{distill}} = &\mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{reason}} \\ &+ \lambda_3 \mathcal{L}_{\text{action}} + \lambda_4 \mathcal{L}_{\text{stable}}. \end{aligned} \tag{9}$$

where $\mathcal{L}_{task}$ preserves task performance, $\mathcal{L}_{align}$ aligns cross-modal scientific evidence, $\mathcal{L}_{reason}$ preserves reasoning traces, $\mathcal{L}_{action}$ transfers tool-use behavior, and $\mathcal{L}_{stable}$ encourages stable and reusable representations. The coefficients $\lambda_1, \ldots, \lambda_4$ control the relative contribution of each component.

3.4. Stability and Scientific Reliability

A central claim of this paper is that AI scientist systems require stable scientific representations. Scientific workflows

are vulnerable to spurious correlations, hallucinated explanations, unstable reasoning traces, and irreproducible tool-use behavior. We therefore introduce a stability criterion:

$$\Delta z_t = \|z_t - z_{t-1}\|, \tag{10}$$

where z_t is the system representation at step t. A reliable AI scientist should avoid unnecessary representational volatility when the underlying scientific structure remains unchanged.

We define a stability regularization term:

$$\mathcal{L}_{stable} = \sum_{t=1}^{T} w_t \|z_t - z_{t-1}\|^2, \tag{11}$$

where w_t controls the importance of stability at each workflow step. This term does not force the model to ignore real scientific change. Instead, it suppresses unstable or spurious changes while allowing representation updates when new evidence justifies revision.

3.5. Autonomy Assessment Criteria

To distinguish tools, collaborators, and independent contributors, we propose evaluating AI scientist systems through workflow-level autonomy rather than output quality alone. Let the autonomy profile be:

$$\mathcal{A} = (A_h, A_e, A_v, A_t, A_r, A_g), \tag{12}$$

where A_h denotes hypothesis generation, A_e denotes experimental planning, A_v denotes verification behavior, A_t denotes tool-use competence, A_r denotes reproducibility support, and A_g denotes governance and attribution clarity.

We classify systems into three levels:

$$\text{Role}(\mathcal{S}) = \begin{cases} \text{Tool}, & \mathcal{A} \in \Omega_T, \\ \text{Collaborator}, & \mathcal{A} \in \Omega_C, \\ \text{Contributor}, & \mathcal{A} \in \Omega_I. \end{cases} \tag{13}$$

This classification reflects the workshop's central question: whether AI scientists should be understood as tools, collaborators, or founders. Our position is that this distinction should be based on the system's role inside the scientific workflow, not merely on the sophistication of its interface or the fluency of its outputs.

3.6. Summary

In summary, our methodology provides a neural-mechanistic framework for analyzing AI scientist systems. The framework connects five stages of AI evolution, defines multimodal distillation as the key mechanism linking representation and action, and proposes autonomy criteria grounded in scientific workflow participation. This allows AI scientists to be evaluated not only by what they produce, but by how they reason, act, verify, and participate in discovery.

4. Theoretical Statement and Proof

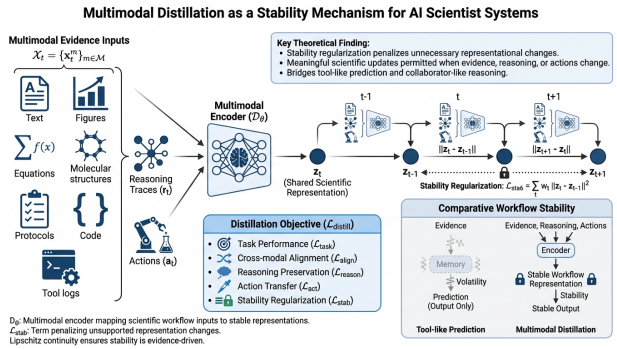

Figure 4. Multimodal distillation as a stability mechanism for AI scientist systems.

Figure 4 visualizes the theoretical role of multimodal distillation in our framework. Scientific workflow inputs are inherently heterogeneous, including papers, figures, equations, molecular structures, protocols, code, tool logs, reasoning traces, and experimental actions. A multimodal encoder D_θ maps these inputs into a shared representation z_t, while the stability regularization term penalizes unsupported changes across workflow steps.

4.1. Theoretical Motivation

The purpose of this section is not to prove full autonomy in AI scientist systems, but to formalize how multimodal distillation can support stable scientific workflow participation. Under a Lipschitz and stability-regularized learning setting, multimodal distillation reduces unnecessary representational volatility while preserving evidence-driven updates. More broadly, the framework can be interpreted as learning compact scientific representations that minimize workflow-level uncertainty across reasoning and tool-use trajectories. Specifically, multimodal distillation can be formulated as:

$$z^* = \arg\min_z \Big(I(X; z) - \beta I(z; Y) \Big), \tag{14}$$

where $I(X; z)$ controls compressed multimodal representations and $I(z; Y)$ preserves scientifically relevant information.

In addition, workflow consistency can be regularized as:

$$L_{\text{workflow}} = \sum_{t=1}^{T} \|\phi_t - \phi_{t-1}\|^2, \tag{15}$$

where ϕ_t denotes workflow representations across reasoning, tool use, verification, and experimental stages. This interpretation connects multimodal distillation with stable and coherent scientific workflow behavior rather than only local representation smoothing.

4.2. Setup

Let $\mathcal{M}$ denote a set of scientific modalities, such as text, figures, equations, molecular structures, protocols, code, and tool logs. At workflow step t, the scientific evidence is represented as

$$\mathcal{X}_t = \{x_t^m\}_{m \in \mathcal{M}}. \tag{16}$$

A multimodal encoder maps this evidence into a shared scientific representation:

$$z_t = D_\theta(\mathcal{X}_t, r_t, a_t), \tag{17}$$

where r_t denotes reasoning traces and a_t denotes tool-use or experimental actions. The distillation objective is defined as

$$\begin{aligned}\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{reason}} \\ + \lambda_3 \mathcal{L}_{\text{act}} + \lambda_4 \mathcal{L}_{\text{stab}},\end{aligned} \tag{18}$$

where $\mathcal{L}_{\text{task}}$ preserves task performance, $\mathcal{L}_{\text{align}}$ aligns cross-modal scientific evidence, $\mathcal{L}_{\text{reason}}$ preserves reasoning traces, $\mathcal{L}_{\text{act}}$ transfers tool-use behavior, and $\mathcal{L}_{\text{stab}}$ regularizes unnecessary representational volatility. We define the stability term as

$$\mathcal{L}_{\text{stab}} = \sum_{t=1}^{T} w_t \|z_t - z_{t-1}\|^2, \tag{19}$$

where $w_t \geq 0$ controls the strength of stability regularization at step t.

4.3. Proposition: Stability Effect of Multimodal Distillation

Proposition 1. Assume that the multimodal distillation mapping D_θ is L-Lipschitz with respect to its input evidence, reasoning traces, and tool-use actions. That is, for two workflow states $(\mathcal{X}_t, r_t, a_t)$ and $(\mathcal{X}_{t-1}, r_{t-1}, a_{t-1})$,

$$\begin{aligned}\|z_t - z_{t-1}\| \leq L\big(d(\mathcal{X}_t, \mathcal{X}_{t-1}) \\ + \|r_t - r_{t-1}\| + \|a_t - a_{t-1}\|\big),\end{aligned} \tag{20}$$

where $d(\mathcal{X}_t, \mathcal{X}_{t-1})$ measures cross-modal evidence variation. If $\lambda_4 > 0$, minimizing $\mathcal{L}_{\text{distill}}$ penalizes representation changes that are not supported by corresponding evidence, reasoning, or action changes. Therefore, multimodal distillation with stability regularization reduces spurious representational volatility while preserving evidence-driven scientific updates.

4.4. Proof

Proof. By definition, the representation at step t is

$$z_t = D_\theta(\mathcal{X}_t, r_t, a_t), \tag{21}$$

and the representation at step $t-1$ is

$$z_{t-1} = D_\theta(\mathcal{X}_{t-1}, r_{t-1}, a_{t-1}). \tag{22}$$

Since D_θ is assumed to be L-Lipschitz, we have

$$\begin{aligned}\|z_t - z_{t-1}\| &= \|D_\theta(\mathcal{X}_t, r_t, a_t) \\ &\quad - D_\theta(\mathcal{X}_{t-1}, r_{t-1}, a_{t-1})\| \\ &\leq L\big(d(\mathcal{X}_t, \mathcal{X}_{t-1}) \\ &\quad + \|r_t - r_{t-1}\| + \|a_t - a_{t-1}\|\big).\end{aligned} \tag{23}$$

This inequality shows that representational variation is bounded by changes in scientific evidence, reasoning traces, and tool-use actions. Therefore, if the scientific evidence remains nearly unchanged, and if reasoning traces and actions also remain stable, then the representation should not change significantly.

The stability regularization term is

$$\mathcal{L}_{\text{stab}} = \sum_{t=1}^{T} w_t \|z_t - z_{t-1}\|^2. \tag{24}$$

Because $\lambda_4 > 0$, the full objective includes the term

$$\lambda_4 \mathcal{L}_{\text{stab}} = \lambda_4 \sum_{t=1}^{T} w_t \|z_t - z_{t-1}\|^2. \tag{25}$$

During minimization, large values of $\|z_t - z_{t-1}\|^2$ increase the objective unless they are compensated by improvements in task performance, cross-modal alignment, reasoning preservation, or action transfer. Thus, unnecessary changes in representation are discouraged.

However, the stability term does not force representations to remain constant. If there is meaningful evidence change, such as a new experimental result, revised hypothesis, updated protocol, or tool-generated observation, then $d(\mathcal{X}_t, \mathcal{X}_{t-1})$, $\|r_t - r_{t-1}\|$, or $\|a_t - a_{t-1}\|$ may increase.

In that case, the Lipschitz bound permits a corresponding update in z_t. Therefore, the model can still revise its representation when scientific evidence warrants revision.

Hence, multimodal distillation with stability regularization suppresses unsupported volatility while allowing evidence-driven scientific updates. This proves the proposition. □

4.5. Implication for AI Scientist Systems

The proposition supports the main position of this paper. A tool-like predictor may produce outputs without stable workflow memory or cross-modal consistency. In contrast, a multimodal distillation-driven AI scientist system can preserve reusable scientific representations across workflow steps. Stability regularization reduces spurious shifts, while multimodal evidence, reasoning traces, and tool-use behavior still allow meaningful revision. This provides a theoretical basis for treating multimodal distillation as a bridge from isolated prediction toward reproducible, collaborator-like scientific workflow participation.

5. Discussion

This research conceptualizes AI scientific systems as a neural-mechanistic transition rather than merely a product-level evolution from chatbots to agents. Current systems exhibit different stages of this transition: AlphaFold and AlphaFold 3 focus on biological structure prediction, GNoME and A-Lab support materials discovery and autonomous synthesis, Coscientist combines language-model reasoning with chemical experimentation, while AlphaGeometry and AlphaProof demonstrate structured mathematical reasoning capabilities (Jumper et al., 2021; Abramson et al., 2024; Merchant et al., 2023; Szymanski et al., 2023; Boiko et al., 2023; Trinh et al., 2024). To improve operationalization, Table 1 compares representative AI-for-science systems across the proposed autonomy dimensions, including hypothesis generation, tool use, verification, reproducibility, and workflow participation. These examples suggest that scientific AI is gradually moving beyond passive assistance toward workflow-level participation.

A key implication is that AI scientist systems should not be evaluated solely by output quality or benchmark accuracy. For tool-like systems, prediction performance may be sufficient. For collaborator-like systems, evaluation should additionally consider hypothesis generation, experimental planning, verification, reproducibility, and scientific documentation. This motivates a transition from static benchmark evaluation to workflow-level assessment. Techniques such as Chain-of-Thought, zero-shot reasoning, and ReAct further demonstrate that neural systems can expose intermediate reasoning processes and coordinate actions with external tools (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023). In scientific environments, these capabilities must also be evaluated with provenance tracking, auditability, and human oversight.

Multimodal distillation is important because scientific knowledge exists across heterogeneous modalities including papers, figures, equations, software, protocols, molecular structures, instruments, and experimental procedures. Existing multimodal systems align language, vision, and structured scientific signals into reusable representations (Bommasani et al., 2021; Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Incremental distillation further shows that reasoning traces can also be transferred rather than only final outputs (Hsieh et al., 2023). Unlike standard knowledge distillation or multimodal representation learning, we define multimodal distillation as a workflow-oriented mechanism that compresses cross-modal evidence, reasoning traces, expert demonstrations, and tool-use behavior into stable representations that support scientific workflow participation.

This perspective also suggests that autonomy is not a binary property. A system may autonomously search literature while still requiring human supervision for experimental design; it may generate hypotheses but still depend on human verification; or it may operate tools without maintaining accountability for scientific claims. Therefore, the distinction between tool, collaborator, and contributor should be defined according to workflow role rather than interface sophistication alone.

5.1. Limitations

This work is primarily conceptual and does not introduce a new benchmark, dataset, or implemented AI scientist architecture. The proposed framework is intended as a structured perspective for analyzing the evolution from predictive neural networks to workflow-oriented scientific agents. In addition, the proposed autonomy dimensions remain qualitative and require future operationalization through measurable evaluation metrics and standardized protocols. While the theoretical discussion provides intuition regarding representation stability and workflow consistency, the framework currently lacks empirical validation and large-scale comparative experimentation.

6. Conclusion

The paper contended that AI scientific systems must be perceived as more than a mere transition from passive tools to autonomous agents. Their rise signifies a profound change in neural mechanisms: transitioning from task-specific predictors to foundational models, then to multimodal representation learning, followed by tool-utilizing agents, and ultimately to systems adept at engaging in scientific workflows.

We recognized multimodal distillation as a crucial process in this transformation, since it consolidates diverse scientific information, expert demonstrations, reasoning pathways, and tool-utilization behaviors into stable, reusable representations.

The primary assertion is that the autonomy of AI scientists should not be evaluated just based on the quality of their output or benchmark performance. A technology that generates coherent scientific prose or attains high predictive accuracy does not inherently qualify as a scientific collaborator. Autonomy should be evaluated based on workflow-level competencies, encompassing hypothesis development, experimental design, verification practices, tool dependability, reproducibility, clarity of attribution, and the necessity for human oversight. This role-based perspective enables the classification of scientific AI systems as tools, collaborators, or contributors based on their real functions within the research process.

This research establishes a conceptual framework for assessing the forthcoming generation of AI scientist systems by linking neural network evolution with issues of scientific agency, multimodal distillation, and governance. Future endeavors must establish definitive benchmarks, audit trails, reproducibility processes, and institutional norms to ascertain the conditions under which AI functions as an assistant, transitions to a collaborator, and initiates autonomous scientific contributions.

7. Future Work

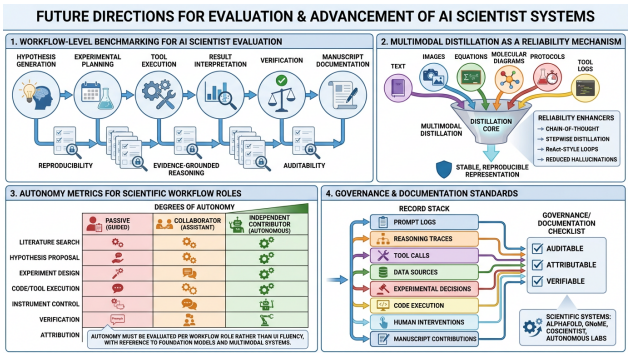

Figure 5. Future directions for evaluating and advancing AI scientist systems.

Figure 5 summarizes the main future directions that follow from our framework. We argue that AI scientist systems require evaluation and governance beyond conventional benchmark accuracy.

A concrete workflow-level evaluation protocol would strengthen the practical relevance of the proposed framework. Future evaluation should measure not only prediction quality, but also hypothesis generation, experimental planning, tool use, verification, reproducibility, and human oversight across scientific workflows.

Initially, benchmarks for AI scientists should progress beyond static question answering and isolated predictive tasks to encompass comprehensive scientific workflow evaluation, which includes hypothesis formulation, experimental planning, tool execution, result interpretation, verification, and manuscript-level documentation. Current advancements in autonomous chemistry, autonomous laboratories, and formal reasoning systems indicate that AI systems are beginning to integrate into certain aspects of the research process; however, standardized workflow-level benchmarks are still inadequately developed (Boiko et al., 2023; Szymanski et al., 2023; Trinh et al., 2024; Musaelian et al., 2023; Cheng et al., 2024). These criteria should assess not only ultimate accuracy but also the reproducibility, evidence-based nature, and auditability of the reasoning process.

Secondly, multimodal distillation ought to be examined as a technique for reliability. Contemporary distillation techniques frequently emphasize compression or performance transfer, whereas AI scientific systems necessitate the transfer of scientific methodologies, reasoning pathways, tool-utilization behaviors, experimental limitations, and cross-modal evidence. Chain-of-thought reasoning, sequential distillation, and ReAct-style reasoning-action loops establish preliminary foundations for this trajectory (Wei et al., 2022; Kojima et al., 2022; Hsieh et al., 2023; Yao et al., 2023). Future studies ought to investigate the impact of various modalities of distillation on stability, hallucination mitigation, repeatability, and scientific validity.

Third, enhanced autonomy metrics are necessary. Future research should assess varying degrees of autonomy in AI systems across particular workflow parameters, including literature search, hypothesis formulation, experimental design, code execution, instrument control, verification, and attribution, rather than categorizing them as just autonomous or non-autonomous. Foundation models and multimodal systems exhibit extensive representational capabilities across language, vision, and structured data; however, their scientific autonomy should be evaluated based on their role in workflows rather than their interface proficiency (Bommasani et al., 2021; Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Yan et al., 2025a).

As systems like AlphaFold, AlphaFold 3, GNoME, Coscientist, and autonomous laboratories assume more significant scientific functions, there is a pressing need for clearer standards to verify, attribute, and regulate AI involvement in discovery processes (Jumper et al., 2021; Abramson et al., 2024; Merchant et al., 2023; Boiko et al., 2023; Szymanski et al., 2023).

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and constructive feedback. Their suggestions have significantly improved the quality and clarity of this manuscript. The authors are also grateful to the AI and scientific discovery communities for the open exchange of ideas that continues to advance research on foundation models, multimodal intelligence, autonomous agents, and AI scientist systems.

No external funding was received for this work.

References

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*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

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

Cheng, W., Wu, Y., Wu, Z., Ling, H., and Hua, G. Toward high quality multi-object tracking and segmentation without mask supervision. *IEEE Transactions on Image Processing*, 33:3369–3384, 2024.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

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*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023. doi: 10.1038/s41467-023-36329-y.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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 novel materials. *Nature*, 624 (7990):86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. doi: 10.1038/s41586-023-06747-5.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yan, J., Cheng, Y., Wang, Q., Liu, L., Zhang, W., and Jin, B. Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2827–2842, 2024a.

Yan, J., Wang, Q., Cheng, Y., Su, Z., Zhang, F., Zhong, M., Liu, L., Jin, B., and Zhang, W. Optimized single-image super-resolution reconstruction: A multimodal approach based on reversible guidance and cyclical knowledge distillation. *Engineering Applications of Artificial Intelligence*, 133:108496, 2024b.

Yan, J., Cheng, Y., Zhang, F., Li, M., Zhou, N., Jin, B., Wang, H., Yang, H., and Zhang, W. Research on multimodal techniques for arc detection in railway systems with limited data. *Structural Health Monitoring*, pp. 14759217251336797, 2025a.

Yan, J., Cheng, Y., Zhang, F., Zhou, N., Wang, H., Jin, B., Wang, M., and Zhang, W. Multi-modal imitation learning for arc detection in complex railway environments. *IEEE Transactions on Instrumentation and Measurement*, 2025b.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.