Bridging Scientific Goals and AI Execution: A Multi-Agent Framework with Hierarchical Self-Evolution

Anonymous authors
Paper under double-blind review

Abstract

Scientific discovery underpins progress in chemistry, biomedicine, and materials research. AI increasingly supports this work through scientific modeling, candidate design, and experiment selection. However, applying AI to a scientific task still requires domain scientists and AI specialists to jointly design the investigation, choose suitable methods, and revise them as results emerge. This repeated coordination is demanding, and experience from one investigation is not readily available to the next. We propose a multi-agent framework that addresses these difficulties through hierarchical self-evolution. The framework first generates a research workflow from the scientist's task description, then formulates an AI problem and execution plan for each scientific subproblem. Scientific and AI-method reviews guide intra-task evolution of the workflow, methods, and execution strategies. After the investigation, the framework extracts workflow playbooks, policy priors, executable skills, and scientific intuitions for reuse in subsequent tasks through inter-task evolution. Experiments across chemistry, biomedicine, and materials science show that our hierarchical self-evolving multi-agent framework supports scientific discovery by translating scientific goals into executable AI plans and improving task outcomes through intra-task and inter-task evolution. In comparisons with expert-team baselines on the evaluated tasks, the framework also achieves outcomes comparable to or better than those of collaborating domain scientists and AI experts.

Keywords: AI-assisted scientific discovery; multi-agent systems; hierarchical self-evolution; experience reuse; human–AI collaboration.

1 Introduction

Scientific discovery is essential to advances in chemistry, drug development, and materials design. AI is playing an increasingly important role in this process: models predict scientific properties, generative methods propose new candidates, and learning algorithms help select experiments. These capabilities have supported applications including autonomous inorganic synthesis, the design of new materials, and compound screening (Szymanski et al., 2023; Zeni et al., 2025; Tosh et al., 2025).

Two complementary directions advance AI-assisted discovery. Self-driving laboratories and scientific agents organize research through planning, execution, and feedback (Szymanski et al., 2023; Pilon et al., 2026; Gao et al., 2026; Bran et al., 2024; Boiko et al., 2023; Rothfarb et al., 2026). Task-specific methods use AI, including active learning, to address particular scientific discovery problems, while platforms make these methods reusable (Dodds et al., 2024; Tosh et al., 2025; Zeni et al., 2025; Hickman et al., 2025; Ma et al., 2024; Cao et al., 2026b).

However, applying AI to a new scientific goal still requires domain and AI experts to jointly formulate subproblems, select methods, and interpret feedback. Repeating this coordination is costly, and decisions made for one task are difficult to reuse in another. We address this gap with a hierarchical self-evolving multi-agent framework (Figure 1). It generates scientific workflows and stage-specific AI solutions, then supports two scales of self-evolution: intra-task evolution revises the ongoing investigation through scientific and AI-method review; inter-task evolution turns completed work into

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

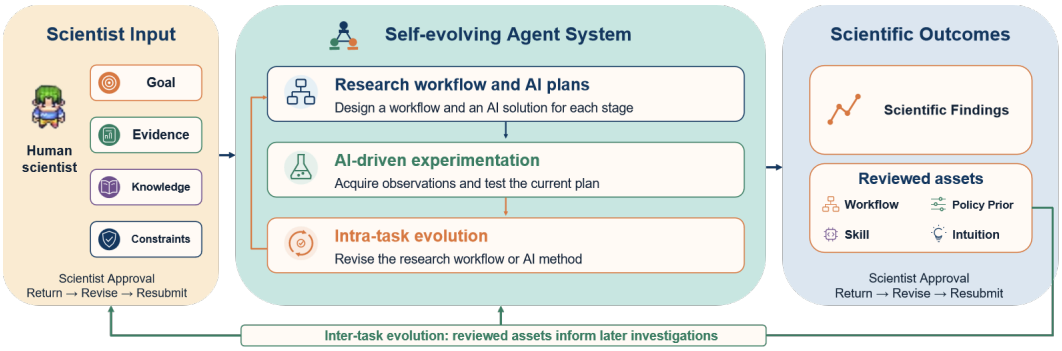

Figure 1: Scientific input, AI-assisted scientific discovery, and scientific outcomes. Intra-task evolution revises the current investigation; inter-task evolution supplies reviewed assets to later tasks.

assets usable in subsequent tasks. Human scientists retain the highest decision authority, supervising the process and reviewing or revising key task information. Our contributions are:

- A multi-agent framework that transforms scientific goals into executable AI investigations by jointly organizing scientific subgoal decomposition, the generation and execution of stage-specific AI plans, and iterative experiment refinement.
- A hierarchical self-evolution method operating within and between tasks: experimental feedback revises current workflows and AI plans, while curated assets support planning and execution in subsequent investigations.
- Experiments across three domains validate the framework's effectiveness in scientific discovery and the value of both evolution mechanisms. Compared with collaborative investigations by domain and AI experts, the framework achieves comparable or better research outcomes.

2 Related Work

Scientific agents and self-driving laboratories. Building on early closed-loop systems (Kusne et al., 2020; Cao et al., 2021), recent laboratories coordinate autonomous synthesis, reaction optimization, and device fabrication (Szymanski et al., 2023; Pilon et al., 2026; Gao et al., 2026); multi-stage optimization models dependencies between experiments (Torresi & Friederich, 2026). Scientific agents extend tool coordination through chemical reasoning, catalyst search, drug discovery, and hypothesis exploration (Bran et al., 2024; Boiko et al., 2023; Rothfarb et al., 2026; Ock et al., 2026a; Kim et al., 2025; Nguyen & Ji, 2026; Pu et al., 2025). Experience also supports reuse: ChemAgent decomposes chemical reasoning problems, retrieves planning and execution memories, and refines solutions or downstream subtasks (Tang et al., 2025). STELLA combines biomedical planning, critical feedback, model training, and tool creation with evolving template and tool libraries (Jin et al., 2025); related systems also accumulate scientific skills (Huang et al., 2025).

AI for specific scientific problems and platforms. Task-specific AI methods address defined scientific problems through modeling and experimental selection (Lookman et al., 2019; Li et al., 2019; Xin et al., 2021; Erps et al., 2021). Their applications now span alloys, molecular design, compound combinations, framework materials, and materials generation (Cao et al., 2024; Dodds et al., 2024; Tosh et al., 2025; Zhang et al., 2025; Zeni et al., 2025; Cao et al., 2026a). To adapt these methods to task feedback, recent approaches adjust representations, kernels, acquisition rules, or tool-guided edits (Rajabi-Kochi et al., 2025; Suwandi et al., 2025; Ngo et al., 2026; Wang et al., 2026; Yang et al., 2025). Scientific platforms and human–AI learning make such methods accessible across investigations (Hickman et al., 2025; Ma et al., 2024; Cao et al., 2026b; Masouleh et al., 2025).

Positioning. Scientific agents and self-driving laboratories focus on coordinating research and tools, but connecting scientific subgoal decomposition to executable, revisable AI plans across domains remains insufficiently addressed. For example, ChemAgent does not formulate chemical reasoning subtasks as executable experimental AI plans (Tang et al., 2025). Task-specific AI methods and platforms focus on particular scientific problems and therefore rely on upstream problem formulation and user-configured optimization tasks, as exemplified by Bgolearn for materials discovery

(Cao et al., 2026b). Our framework addresses these gaps by decomposing scientific goals, generating executable AI plans for each stage, and revising workflows and methods using experimental feedback. On the one hand, our method can implement research-planning and AI-execution agents in broader scientific-agent systems or self-driving laboratories. On the other hand, task-specific AI methods and platforms can serve as its callable skills.

3 Problem Formulation

A scientific task is

$$\mathcal{T} = (G, I_0, C), \tag{1}$$

where G specifies the scientific goal and its outcome criterion, I_0 contains available observations and knowledge, and C specifies admissible operations and scientific requirements. For possible outcomes $\Omega(\mathcal{T})$ and goal-specific utility $U_G(o)$, the objective is

$$o^* \in \arg\max_{o \in \Omega(\mathcal{T})} U_G(o) \quad \text{subject to} \quad C(o) = \text{true}. \tag{2}$$

Achieving this objective requires a research process that produces the evidence needed to evaluate U_G. Our framework constructs that process as a workflow, translates its stages into executable AI plans, and evaluates the resulting observations against G and C.

4 Framework

Studies of reaction optimization, compound screening, adsorption search, and MXene aerogel design, among other applications (Shields et al., 2021; Corsello et al., 2020; Lan et al., 2023; Shrestha et al., 2024), suggest a common process for AI-assisted scientific discovery: formulate the task, design the research workflow, conduct AI-driven experiments, and curate the resulting experience. Mathematically,

$$(\mathcal{T}, M_q) \longrightarrow W^{(v)} \longrightarrow \{(Q_n, P_n) : s_n^{(v)} \in S^{(v)}\} \longrightarrow (O_q, \tau_q) \longrightarrow M_{q+1}. \tag{3}$$

Here, M_q is the asset library before task q, $W^{(v)}$ is workflow version v, and $S^{(v)}$ is its stage set. Q_n and P_n are a stage's AI problem and execution plan; O_q collects the scientific outcomes, evaluated using U_G, and τ_q records the decisions and observations.

We implement this process through the multi-agent framework in Figure 2, following the division of work between domain scientists and AI experts. The Research Designer decomposes goals and specifies stage dependencies; the Experiment Designer formulates AI problems and selects methods and parameters; the Lab Operator executes plans. Scientific and AI-Method Reviewers assess evidence and method performance, with the Review Chair coordinating revisions. The Research Curator extracts reusable assets. Human scientists review task information, workflows, consequential revisions, and assets, retaining final authority. The Method section explains how these agents produce and revise the decisions in Equation 3.

5 Method

5.1 Task Workflow Design

A scientific goal usually needs to be decomposed into subgoals with sequential dependencies. A flat tool sequence does not specify these scientific dependencies or when the evidence is sufficient to proceed. We therefore represent each workflow version as a directed acyclic graph (DAG):

$$W^{(v)} = (S^{(v)}, E^{(v)}), \quad S^{(v)} = \{s_n^{(v)}\}_{n=1}^{N_v}, \quad s_n^{(v)} = (g_n, I_n, O_n, c_n). \tag{4}$$

Each stage specifies a subgoal g_n, inputs I_n, outputs O_n, and completion condition c_n; edges $E^{(v)}$ identify prerequisite outputs. AI methods can iterate within a stage while its dependencies remain

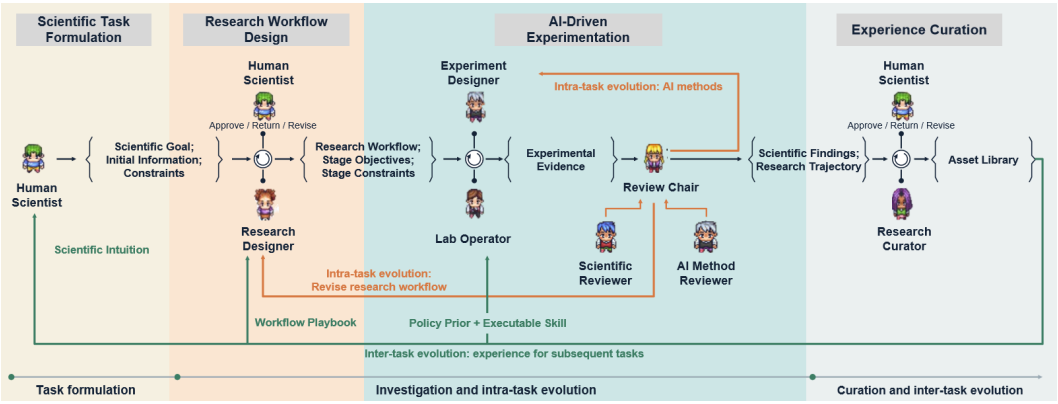

Figure 2: Agents along the research process. Orange paths revise the current workflow or AI plan; green paths reuse reviewed assets. Human scientists retain final decision authority.

unchanged. Workflow revision (Section 5.2) redesigns downstream stages or the entire workflow as a new DAG, preserving acyclic dependencies within each version.

To construct this workflow, the Research Designer receives $\mathcal{T}$, available operations, and applicable workflow and intuition assets. It specifies each stage's scientific subgoal, required inputs, expected outputs, and completion condition, and identifies the prerequisite evidence passed between stages. It returns stage requirements as structured tuples (g_n, I_n, O_n, c_n), together with dependencies and transition conditions. The execution harness checks the output format, acyclic dependencies, input references, and operation availability; failed checks require replanning. Human scientists review the decomposition and scientific requirements; only an approved workflow enters execution.

5.2 AI-Driven Experimentation and Intra-Task Evolution

Extensible AI execution. Building on the task decomposition, we design the Experiment Designer to turn each scientific subgoal into an AI problem Q_n and an executable plan P_n. Its objective is to produce the stage outputs O_n that satisfy c_n, using the available inputs I_n. The plan is defined as

$$P_n = \big(Q_n, \{(k_j, \theta_j)\}_{j=1}^{m_n}, V_n\big). \tag{5}$$

Here, k_j is a registered skill, including a trainable AI algorithm or a callable pretrained scientific model; θ_j denotes its parameters, m_n the number of planned skill invocations, and V_n the validation procedure.

Scientific subproblems differ in their data, models, and experimental interfaces. A skill registry separates implementations from plan design; entries declare problem types, input/output schemas, parameters, and validation requirements. Using the stage goal, available data, and required outputs, the Experiment Designer first identifies the AI problem type, such as regression, classification, or optimization, and specifies its objective. It then selects compatible skills and parameters using task requirements and prior experience. The Lab Operator invokes these skills to execute the plan, including model fitting and numerical computation; failures such as incompatible data or invalid parameters return the plan for correction.

Two revision needs, three review agents. Scientific discovery requires iterative assessment of experimental observations against the goal. Feedback may require scientific-plan changes, such as collecting missing evidence, or AI-method changes, such as replacing a poorly fitting model. Both can arise within one investigation. We therefore design three review agents to diagnose these needs and coordinate a response. The Scientific Reviewer receives the scientific goal, required outputs, available evidence, and stage dependencies. It checks whether the evidence answers the scientific question and recommends completion, further measurement, or workflow replanning. The AI-Method Reviewer receives the executed model, parameters, validation diagnostics, selection history, and progress. It evaluates model suitability and candidate selection, recommending continuation, parameter adjustment, or method reselection. Both reports must identify observations supporting their diagnoses.

The Review Chair receives both reports, available observations, and previous evolution attempts. Through LLM judgment, it reconciles disagreements and outputs a next action with its rationale, evidence references, and responsible agent. Numeric adjustments go to adaptive skills, method reselection to the Experiment Designer, and workflow revision to the Research Designer. Consequential changes require human-scientist approval. When no revision is proposed, the harness checks action–agent consistency and evidence references; accepting completion additionally requires the stage goal to be met. Failed checks return the investigation to review and evolution. Appendix A, Algorithm 1, specifies this loop and its decision requirements.

5.3 Asset Generation and Inter-Task Evolution

To reuse completed research, we synthesize its execution history into workflow playbooks, policy priors, executable skills, and scientific intuitions, drawing on scientific-discovery and agent self-evolution research (Wang et al., 2025; Jin et al., 2025; Zheng et al., 2025; Tang et al., 2025; Huang et al., 2025).

To perform this synthesis, we design a Research Curator that receives workflows, executed plans, outcomes, and decision traces. Type-specific prompts extract stage dependencies and transition conditions into Workflow Playbooks; evidence-supported selection and adaptation decisions into Policy Priors; existing executable models or procedures and invocation requirements into Skill Assets; and scoped, testable hypotheses with counterexamples and suggested tests into Scientific Intuitions. Asset generation and library update are

$$A_q = \text{Curate}(W_q, P_q, O_q, \tau_q), \qquad M_{q+1} = M_q \cup A_q^{\text{approved}}. \tag{6}$$

Here, P_q collects executed plans. Each asset links its content to source observations or an executable artifact and records applicability and limitations. The harness checks these sources and artifacts. The assets are then submitted for human approval; only assets that pass both checks are retained.

Retrieval matches the new task and interface requirements. Workflow and intuition assets guide the Research Designer; intuitions also support scientific review. Policy and skill assets guide the Experiment Designer, with accepted priors applied during initialization and skills executed by the Lab Operator. Each receiving agent records reuse, adaptation, or rejection, its reason, and the affected stage or plan; execution records actual skill use. Appendix A, Algorithm 2, details this process.

6 Experiments

6.1 Setup

We evaluate task planning, execution, and evolution in three settings, using published measurements as experimental feedback:

- **HTE (chemistry):** high-throughput reaction measurements support yield prediction and condition optimization, the two tasks considered here (Ahneman et al., 2018; Shields et al., 2021; Schwaller et al., 2021; Zhao et al., 2021; Viet Johansson et al., 2022; Saebi et al., 2023).
- **PRISM (biomedicine):** compound–cell-line response profiles support activity screening and, where required, secondary dose-response confirmation (Yu et al., 2016; Corsello et al., 2017; 2020; Feizi et al., 2022; Carli et al., 2025; Zhang et al., 2026).
- **OC20Dense (materials science):** adsorption structures and reference energies support low-energy configuration search using scientific models such as EquiformerV2 (Chanussot et al., 2021; Lan et al., 2023; Liao et al., 2024; Chen et al., 2025; Rensmeyer et al., 2026; Ock et al., 2026b).

We identify each task by domain, experiment, and task index (e.g., HTE-1.2-1). Appendix B.1 specifies datasets, candidate pools, and metrics; identical controls share one identifier.

Language models. We use Qwen3.8-max (Q), GPT-6 (G6), GPT-5.6 (G5.6), Gemini 3.1 Pro (Ge), DeepSeek V4 Pro (DS), and Kimi K3 (K3).

Metrics. For HTE prediction, we follow Ahneman et al. (2018) and report RMSE (↓) and R^2 (↑). For reaction discovery, we evaluate yield (↑), following Shields et al. (2021). PRISM re-

Table 1: Research questions and experimental designs.

Research question	Comparison
RQ1 Scientific task execution	Individual/continual tasks; task-to-plan evaluation
RQ2 Inter-task evolution	Assets/history/none; type ablations; accumulation
RQ3 Intra-task evolution	Six cases: None, AI only, science only, Full
RQ4 Expert comparison	Domain and AI expert team versus framework
Supplementary (RQ5–RQ6)	Cross-model transfer and case studies (Appendix C)

Table 2: RQ1: individual (A) and continual (B) tasks. Arrows indicate improvement. †Published RF benchmark (Ahneman et al., 2018); otherwise, pool optima (—: unavailable). Units: yield %, RMSE/yield gap percentage points, energy gap eV. PRISM reports hits.

Task / metric	Q	G6	G5.6	Ge	DS	K3	Reference
A. E1.1: individual tasks — prediction							
HTE-1.1-1: RMSE ↓	6.1253	6.0312	7.452	6.4083	5.7448	7.4483	7.8†
HTE-1.1-1: R^2 ↑	0.9505	0.952	0.9267	0.9458	0.9564	0.9268	0.92†
A. E1.1: individual tasks — discovery							
HTE-1.1-2: yield (%) ↑	68.2481	68.2481	68.2481	68.2481	68.2481	68.2481	68.2481
HTE-1.1-2: yield gap ↓	0	0	0	0	0	0	0
PRISM-1.1-1: hits ↑	12	16	12	9	9	10	—
PRISM-1.1-2: hits ↑	17	13	22	10	14	15	—
OC20-1.1-1: gap ↓	0	0	0	0	0	0	0
OC20-1.1-2: gap ↓	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0
B. E1.2: continual tasks							
HTE-1.2-1	88.4464	91.3906	91.3906	91.3906	91.3906	91.3906	91.3906
HTE-1.2-2	78.4679	87.0352	87.0352	78.4679	87.0352	87.0352	87.0352
HTE-1.2-3	74.2893	74.2893	74.2893	74.2893	80.3388	80.3388	80.3388
HTE-1.2-4	87.9792	87.9792	87.9792	87.9792	88.0506	87.9792	88.0506
PRISM-1.2-1: hits ↑	6	10	16	9	10	8	—
PRISM-1.2-2: hits ↑	8	6	3	10	7	5	—

ports hit counts (↑): the pool's lowest-LFC 10% for screening, or dose-confirmed lowest-AUC 20% for PRISM-3-2 and PRISM-4-1. LFC and AUC describe single-dose and dose-response activity (Corsello et al., 2020). For OC20, we follow AdsorbML (Lan et al., 2023) and report the energy gap from the reference-pool minimum (↓, eV). Five quantitative studies and RQ6 case studies evaluate the framework (Table 1).

6.2 RQ1: Scientific Task Execution

E1.1: individual tasks. To evaluate scientific task execution, we test six tasks spanning prediction, reaction search, compound screening, and adsorption discovery (Table 2A). HTE-1.1-2 reaches the pool maximum of 68.2481%; OC20-1.1-1 reaches its reference minimum, and OC20-1.1-2 finishes 0.0006 eV above it. Table 2 also reports HTE prediction metrics alongside the published RF reference (Ahneman et al., 2018), and active-compound hits for both PRISM tasks. These results demonstrate that the proposed framework can support single-task scientific discovery.

E1.2: continual tasks. To test sustained discovery across changing goals and scientific objects, we transfer experience through four HTE high-yield discovery tasks and two PRISM tasks on different cell lines (Table 2B). On the final HTE task, every model is within 0.0714 percentage points of the pool optimum; the PRISM tasks identify 6–16 and 3–10 active compounds. These results demonstrate that the proposed framework can support multi-task scientific discovery. Table 7 in Appendix B records asset use in these sequences.

E1.3: workflow and AI-plan correctness. To assess task-to-plan translation, we adapt published reaction prediction/optimization, compound-screening, and adsorption-search tasks by varying goals and resource constraints (Ahneman et al., 2018; Shields et al., 2021; Corsello et al., 2020; Lan et al.,

2023). Ten tasks per domain are given to six LLMs, yielding 180 proposals. Three domain experts assess whether workflow stages, dependencies, and outputs satisfy the scientific requirements; three AI experts assess whether AI objectives and methods fit those requirements. Each group votes independently by majority; multiple valid solutions are allowed (Appendix B.2). Table 3 reports 179/180 correct initial workflows (99.4%) and 173/180 correct AI plans (96.1%). Harness-triggered system correction resolves six AI-plan errors; one OC20 proposal requires expert correction of both components. These results show that the proposed framework can translate scientific goals into scientifically appropriate workflows and compatible AI plans.

Table 3: E1.3: proposal correctness. Initial: before harness correction; Checked: both components correct after system correction, before expert modification; Human: proposals requiring expert modification.

Domain	N	Initial workflow	Initial AI plan	Checked both	Human revisions
HTE	60	60/60	59/60	60/60	0
PRISM	60	60/60	56/60	60/60	0
OC20	60	59/60	58/60	59/60	1

6.3 RQ2: Inter-Task Evolution

E2.1: asset effectiveness. To test asset effectiveness, we construct source-to-target transitions with targets HTE-2.1-1, PRISM-2.1-1, and OC20-2.1-1. Experience is collected from completed source investigations and supplied to a subsequent target task. On each target, we compare full assets with no assets and read-only history. Read-only history exposes the original records of preceding tasks as context, without extracting typed assets or making an executable transfer. Figure 3 shows two benefits of the complete asset-reuse scheme, including structured experience and executable skills. First, assets improve 15 of 18 model–task comparisons over no assets and tie three. Second, assets improve 15 comparisons over raw history and tie three. These results indicate that extracting reusable assets enables inter-task evolution.

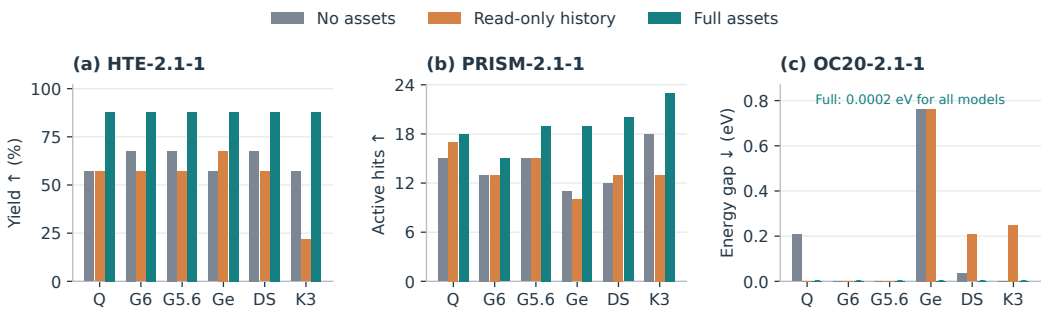

Figure 3: E2.1: assets versus no assets and read-only history. The targets are HTE-2.1-1, PRISM-2.1-1, and OC20-2.1-1, also used in E2.2. Higher yield/hits and lower energy gap are better; full assets reach 0.0002 eV for every OC20 model.

E2.2: asset ablations. To identify which types of experience produce these benefits, we extend the three task settings in E2.1 by removing each asset type and retaining each type alone (Figure 4). Figure 4 shows that the effects differ across tasks. Removing Intuition lowers HTE performance for five models, consistent with scientific hypotheses helping prioritize conditions. Removing Skill reduces PRISM hits for all six models by 3–8, whereas it leaves the OC20 endpoints unchanged. Removing Workflow increases OC20 gaps for Qwen and Kimi, indicating a greater role for experiment organization in these cases. These results show that each asset type can contribute useful experience and that their combined use can improve scientific task outcomes. Their relative importance varies with the task.

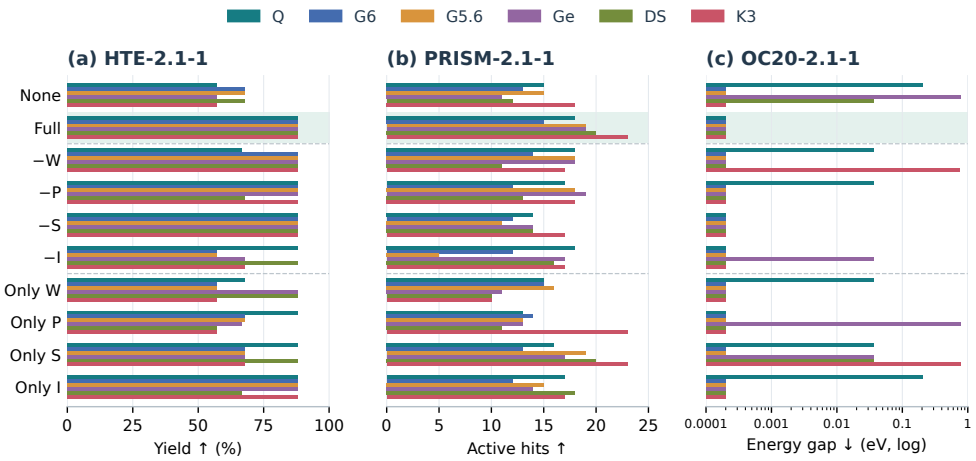

Figure 4: E2.2: asset ablations. W/P/S/I denote Workflow/Policy/Skill/Intuition; "−" removes a type and "Only" retains it alone. Higher yield/hits and lower energy gap are better. OC20 uses a logarithmic axis. Exact values appear in Appendix B.

E2.3: asset accumulation. To examine whether assets accumulated across consecutive tasks can improve a later task, we increase the source history from zero to three tasks. As shown in Figure 5, Qwen and Gemini improve HTE-2.3-1 yield from 65.022% to 68.765%, matching the best observed yield in this comparison; GPT-5.6 improves to 65.3367%. Three of the six models improve, while the other three remain unchanged. Together with E2.1 and E2.2, these results show that inter-task evolution can make past research useful, although the gains may vary with the underlying LLM and task.

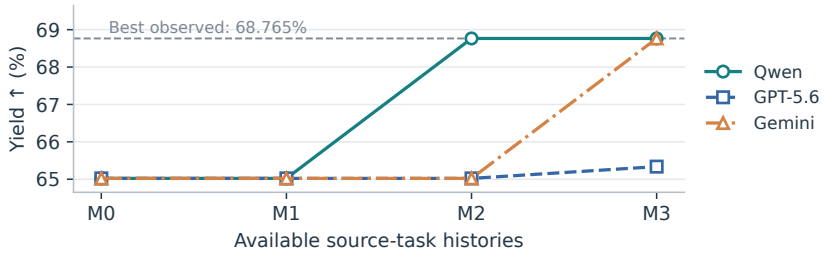

Figure 5: E2.3: asset accumulation on HTE-2.3-1. The dashed line marks the best observed yield, 68.765%. Improving trajectories are shown; unchanged trajectories appear in Table 15.

6.4 RQ3: Intra-Task Evolution

To evaluate intra-task evolution, we run each of six cases from scratch under four configurations and compare None, AI-method-only, scientific-plan-only, and Full branches (Figure 6). Full improves 30 of 36 model–case outcomes over None and ties six. The branch results reveal different revision needs: in HTE-3-1, scientific-plan-only leaves all six models at 57.5399%, whereas Full brings four to 68.2481%. In PRISM-3-2, AI-method-only leaves every model at four hits; enabling scientific-plan evolution raises four models to 8–9 hits. Thus the two forms of evolution address different task requirements, and their combination supports improvement across these settings.

6.5 RQ4: Comparison with Human Experts

To compare with expert collaboration, two domain researchers and two AI researchers jointly complete one investigation per domain. All are master's- or doctoral-level researchers; the same two AI researchers participate across domains (eight experts total). Each human outcome is a fixed reference for six framework LLMs under the same inputs, operations, resources, and evaluation rules. Frame-

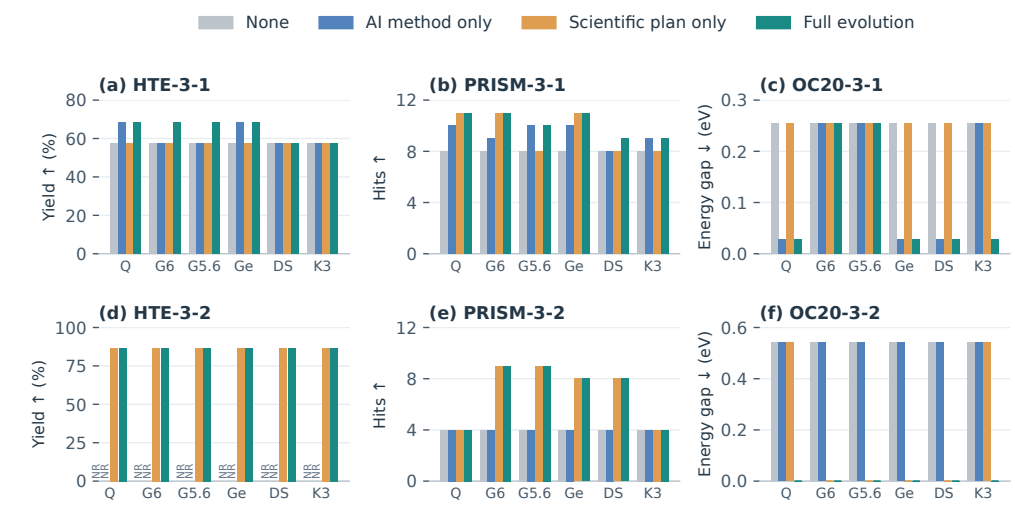

Figure 6: RQ3: four branches in HTE-3-1/2, PRISM-3-1/2, and OC20-3-1/2. PRISM-3-1 counts screening hits; PRISM-3-2 counts dose-confirmed hits. Yield/hits ↑; gap ↓. NR: unmet confirmation; zero gaps are marked at baseline. Exact values: Appendix B.

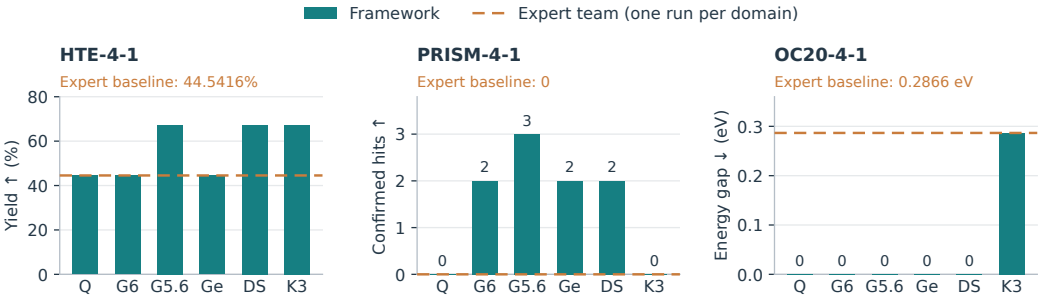

Figure 7: RQ4: six framework conditions (bars) compared with one human-team investigation per domain (dashed line). Yield/hits ↑; energy gap ↓. Each human result is a shared reference, not six independent repetitions.

work approval gates pass automatically, without expert review. Appendix B.3 details the division of work and task conditions.

Figure 7 shows that the framework exceeds the expert baseline in 12 model–task comparisons and ties it in six. Three LLMs attain 67.1813% HTE yield versus 44.5416% for the expert team; four find 2–3 PRISM hits versus zero, and five reach zero OC20 gap versus 0.2866 eV. These results support outcomes comparable to or better than expert collaboration on the evaluated tasks.

7 Discussion and Conclusion

Through workflow design and hierarchical self-evolution, our framework translates scientific goals into AI execution. Across three domains, it supports scientific task execution, benefits from both evolution mechanisms, and matches or exceeds the evaluated expert baselines.

Future work should address three limitations. First, current tasks mainly use short workflows; investigations exceeding four scientific stages or ten AI-plan revisions remain insufficiently tested. Evaluation through existing measurements also leaves performance in prospective discovery with unknown outcomes to be established. Second, agents can hallucinate when generating workflows or AI plans; scientific constraints, tool outputs, and verification agents could strengthen checks on objectives, prerequisites, and method applicability. Third, closer integration of domain-specialized models and human expertise may improve scientific decomposition, method selection, and result interpretation.

AI Use Statement

AI tools assisted literature search and synthesis and manuscript polishing.

References

Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018. doi: 10.1126/science.aar5169.

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

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Bin Cao, Tianhao Su, Shuting Yu, Tianyuan Li, Taolue Zhang, Jincang Zhang, Ziqiang Dong, and Tong-Yi Zhang. Active learning accelerates the discovery of high strength and high ductility lead-free solder alloys. *Materials & Design*, 241:112921, 2024. doi: 10.1016/j.matdes.2024.112921.

Bin Cao, Yin Qin, Yan Luo, Zhehan Ying, Zilin Yan, Lu-Tao Weng, Kaikai Li, and Tong-Yi Zhang. Spatial-adaptive active learning identifies ultra-durable and highly active catalysts for acidic oxygen evolution reaction. *Science Bulletin*, 71(5):1084–1092, 2026a. doi: 10.1016/j.scib.2025.12.021.

Bin Cao, Jie Xiong, Jiaxuan Ma, Yuan Tian, Yirui Hu, Mengwei He, Longhan Zhang, Jiayu Wang, Jian Hui, Li Liu, Dezhen Xue, Turab Lookman, Jun Wang, and Tong-Yi Zhang. Bgolearn: a unified Bayesian optimization framework for accelerating materials discovery. *npj Computational Materials*, 2026b. doi: 10.1038/s41524-026-02226-3.

Liwei Cao, Danilo Russo, Kobi Felton, Daniel Salley, Abhishek Sharma, Graham Keenan, Werner Mauer, Huanhuan Gao, Leroy Cronin, and Alexei A. Lapkin. Optimization of Formulations Using Robotic Experiments Driven by Machine Learning DoE. *Cell Reports Physical Science*, 2(1):100295, 2021. doi: 10.1016/j.xcrp.2020.100295.

Francesco Carli, Pierluigi Di Chiaro, Mariangela Morelli, Chakit Arora, Luisa Bisceglia, Natalia De Oliveira Rosa, Alice Cortesi, Sara Franceschi, Francesca Lessi, Anna Luisa Di Stefano, Orazio Santo Santonocito, Francesco Pasqualetti, Paolo Aretini, Pasquale Miglionico, Giuseppe R. Diaferia, Fosca Giannotti, Pietro Liò, Miquel Duran-Frigola, Chiara Maria Mazzanti, Gioacchino Natoli, and Francesco Raimondi. Learning and actioning general principles of cancer cell drug sensitivity. *Nature Communications*, 16(1):1654, 2025. doi: 10.1038/s41467-025-56827-5.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Junwu Chen, Xu Huang, Cheng Hua, Yulian He, and Philippe Schwaller. A multi-modal transformer for predicting global minimum adsorption energy. *Nature Communications*, 16(1):3232, 2025. doi: 10.1038/s41467-025-58499-7.

Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher C Mader, Aravind Subramanian, and Todd R Golub. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine*, 23(4):405–408, 2017. doi: 10.1038/nm.4306.

Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, Vickie M. Wang, Samantha A. Bender, Evan Lemire, Rajiv Narayan, Philip Montgomery, Uri Ben-David, Colin W. Garvie, Yejia Chen, Matthew G. Rees, Nicholas J. Lyons, James M. McFarland, Bang T. Wong, Li Wang, Nancy Dumont, Patrick J. O'Hearn, Eric Stefan, John G. Doench, Caitlin N. Harrington, Heidi Greulich, Matthew Meyerson, Francisca Vazquez, Aravind Subramanian, Jennifer A. Roth, Joshua A. Bittker, Jesse S. Boehm, Christopher C. Mader, Aviad Tsherniak, and Todd R. Golub. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, 2020. doi: 10.1038/s43018-019-0018-6.

Michael Dodds, Jeff Guo, Thomas Löhr, Alessandro Tibo, Ola Engkvist, and Jon Paul Janet. Sample efficient reinforcement learning with active learning for molecular design. *Chemical Science*, 15(11):4146–4160, 2024. doi: 10.1039/D3SC04653B.

Timothy Erps, Michael Foshey, Mina Konaković Luković, Wan Shou, Hanns Hagen Goetzke, Herve Dietsch, Klaus Stoll, Bernhard von Vacano, and Wojciech Matusik. Accelerated discovery of 3D printing materials using data-driven multiobjective optimization. *Science Advances*, 7(42):eabf7435, 2021. doi: 10.1126/sciadv.abf7435.

Nikta Feizi, Sisira Kadambat Nair, Petr Smirnov, Gangesh Beri, Christopher Eeles, Parinaz Nasr Esfahani, Minoru Nakano, Denis Tkachuk, Anthony Mammoliti, Evgeniya Gorobets, Arvind Singh Mer, Eva Lin, Yihong Yu, Scott Martin, Marc Hafner, and Benjamin Haibe-Kains. PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis. *Nucleic Acids Research*, 50(D1):D1348–D1357, 2022. doi: 10.1093/nar/gkab1084.

Danpeng Gao, Shuaihua Lu, Chunlei Zhang, Ning Wang, Zexin Yu, Xianglang Sun, Rebecca Martin, Francesco Vanin, Liangchen Qian, Nicholas Long, Larry Lüer, Bo Li, Martin Stolterfoht, Junhui Hou, Jun Yin, Yen-Hung Lin, Haipeng Lu, Nan Li, Nicola Gasparini, Christoph Joseph Brabec, Samuel D. Stranks, Xiao Cheng Zeng, and Zonglong Zhu. Autonomous closed-loop framework for reproducible perovskite solar cells. *Nature*, 653(8115):707–714, 2026. doi: 10.1038/s41586-026-10482-y.

Riley J. Hickman, Malcolm Sim, Sergio Pablo-García, Gary Tom, Ivan Woolhouse, Han Hao, Zeqing Bao, Pauric Bannigan, Christine Allen, Matteo Aldeghi, and Alán Aspuru-Guzik. Atlas: A brain for self-driving laboratories. *Digital Discovery*, 4(4):1006–1029, 2025. doi: 10.1039/D4DD00115J.

Xu Huang, Junwu Chen, Yuxing Fei, Zhuohan Li, Philippe Schwaller, and Gerbrand Ceder. CASCADE: Cumulative Agentic Skill Creation through Autonomous Development and Evolution. *arXiv preprint arXiv:2512.23880*, 2025. URL https://arxiv.org/abs/2512.23880.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. STELLA: Self-Evolving LLM Agent for Biomedical Research. *arXiv preprint arXiv:2507.02004*, 2025. URL https://arxiv.org/abs/2507.02004v1.

Hyomin Kim, Yunhui Jang, and Sungsoo Ahn. MT-Mol: Multi Agent System with Tool-based Reasoning for Molecular Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11544–11573, 2025. doi: 10.18653/v1/2025.findings-emnlp.619. URL https://aclanthology.org/2025.findings-emnlp.619/.

A. Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, Albert V. Davydov, Ritesh Agarwal, Leonid A. Bendersky, Mo Li, Apurva Mehta, and Ichiro Takeuchi. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature Communications*, 11(1):5966, 2020. doi: 10.1038/s41467-020-19597-w.

Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M. Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary W. Ulissi. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Computational Materials*, 9(1):172, 2023. doi: 10.1038/s41524-023-01121-5.

Yi-Pei Li, Kehang Han, Colin A. Grambow, and William H. Green. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *The Journal of Physical Chemistry A*, 123(10):2142–2152, 2019. doi: 10.1021/acs.jpca.8b10789.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.12059.

Turab Lookman, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5 (1):21, 2019. doi: 10.1038/s41524-019-0153-8.

Jiaxuan Ma, Bin Cao, Shuya Dong, Yuan Tian, Menghuan Wang, Jie Xiong, and Sheng Sun. MLMD: a programming-free AI platform to predict and design materials. *npj Computational Materials*, 10(1):59, 2024. doi: 10.1038/s41524-024-01243-4.

Shayan Mousavi Masouleh, Corey A. Sanz, Ryan P. Jansonius, Cara Cronin, Jason E. Hein, and Jason Hattrick-Simpers. Human–AI synergy in adaptive active learning for continuous lithium carbonate crystallization optimization. *Digital Discovery*, 4(11):3078–3091, 2025. doi: 10.1039/D5DD00285K.

Giang Ngo, Dat Phan Trong, Dang Nguyen, Sunil Gupta, and Svetha Venkatesh. Adaptive Acquisition Selection for Bayesian Optimization with Large Language Models. *arXiv preprint arXiv:2602.07904*, 2026. URL https://arxiv.org/abs/2602.07904.

Thao Nguyen and Heng Ji. MolLingo: Molecule-Native Representations for LLM-Powered Scientific Agents. *arXiv preprint arXiv:2605.27853*, 2026. URL https://arxiv.org/abs/2605.27853.

Janghoon Ock, Radheesh Sharma Meda, Srivathsan Badrinarayanan, Neha S. Aluru, Achuth Chandrasekhar, and Amir Barati Farimani. Large Language Model Agent for Modular Task Execution in Drug Discovery. *Journal of Chemical Information and Modeling*, 66(4):2055–2068, 2026a. doi: 10.1021/acs.jcim.5c02454.

Janghoon Ock, Radheesh Sharma Meda, Tirtha Vinchurkar, Yayati Jadhav, and Amir Barati Farimani. AdsorbAgent: autonomous identification of stable adsorption configurations via a large language model agent. *Digital Discovery*, 5(2):617–629, 2026b. doi: 10.1039/D5DD00298B.

Simone Pilon, Elia Savino, Oliver M. Bayley, Michael Vanzella, Miguel Claros, Petros Siasiaridis, Junsong Liu, Florian Lukas, Matteo Damian, Vasilis Tseliou, Niccolò Intini, Aidan Slattery, Jesus SanJosé-Orduna, Tim den Hartog, Ron A. H. Peters, Andrea F. G. Gargano, Francesco G. Mutti, and Timothy Noël. A flexible and affordable self-driving laboratory for automated reaction optimization. *Nature Synthesis*, 2026. doi: 10.1038/s44160-026-01053-0.

Yingming Pu, Tao Lin, and Hongyu Chen. PiFlow: Principle-Aware Scientific Discovery with Multi-Agent Collaboration. *arXiv preprint arXiv:2505.15047*, 2025. URL https://arxiv.org/abs/2505.15047.

Mahyar Rajabi-Kochi, Negareh Mahboubi, Aseem Partap Singh Gill, and Seyed Mohamad Moosavi. Adaptive representation of molecules and materials in Bayesian optimization. *Chemical Science*, 16(13):5464–5474, 2025. doi: 10.1039/D5SC00200A.

Tim Rensmeyer, Denis Kramer, and Oliver Niggemann. On-the-fly fine-tuning of foundational neural network potentials: a Bayesian neural network approach. *Digital Discovery*, 5(4):1845–1867, 2026. doi: 10.1039/D5DD00392J.

Samuel Rothfarb, Megan C. Davis, Ivana Matanovic, Baikun Li, Edward F. Holby, and Wilton J. M. Kort-Kamp. Hierarchical Multi-agent Large Language Model Reasoning for Autonomous Heterogeneous Catalyst Discovery. *npj Computational Materials*, 12(1):309, 2026. doi: 10.1038/s41524-026-02139-1.

Mandana Saebi, Bozhao Nan, John E. Herr, Jessica Wahlers, Zhichun Guo, Andrzej M. Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G. Doyle, Nitesh V. Chawla, and Olaf Wiest. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 14(19):4997–5005, 2023. doi: 10.1039/D2SC06041H.

Philippe Schwaller, Alain C. Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2(1):015016, 2021. doi: 10.1088/2632-2153/abc81d.

Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams, and Abigail G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021. doi: 10.1038/s41586-021-03213-y.

Snehi Shrestha, Kieran James Barvenik, Tianle Chen, Haochen Yang, Yang Li, Meera Muthachi Kesavan, Joshua M. Little, Hayden C. Whitley, Zi Teng, Yaguang Luo, Eleonora Tubaldi, and Po-Yen Chen. Machine intelligence accelerated design of conductive MXene aerogels with programmable properties. *Nature Communications*, 15(1):4685, 2024. doi: 10.1038/s41467-024-49011-8.

Richard Cornelius Suwandi, Feng Yin, Juntao Wang, Renjie Li, Tsung-Hui Chang, and Sergios Theodoridis. Adaptive Kernel Design for Bayesian Optimization Is a Piece of CAKE with LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132690–132723, 2025. doi: 10.52202/085713-4421. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c03a2610bca2712b984b331fd4f7bb6f-Abstract-Conference.html.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Xiangru Tang, Tianyu Hu, Muyang Ye, Daniel Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. ChemAgent: Self-updating Memories in Large Language Models Improves Chemical Reasoning. In *International Conference on Learning Representations*, pp. 100922–100958, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa7f64b45970e6a7f8824781e7e01501-Abstract-Conference.html.

Luca Torresi and Pascal Friederich. Multi-stage Bayesian optimisation for dynamic decision-making in self-driving labs. *Digital Discovery*, 5(4):1900–1912, 2026. doi: 10.1039/D5DD00572H.

Christopher Tosh, Mauricio Tec, Jessica B. White, Jeffrey F. Quinn, Glorymar Ibanez Sanchez, Paul Calder, Andrew L. Kung, Filemon S. Dela Cruz, and Wesley Tansey. A Bayesian active learning platform for scalable combination drug screens. *Nature Communications*, 16(1):156, 2025. doi: 10.1038/s41467-024-55287-7.

Simon Viet Johansson, Hampus Gummesson Svensson, Esben Bjerrum, Alexander Schliep, Morteza Haghir Chehreghani, Christian Tyrchan, and Ola Engkvist. Using Active Learning to Develop Machine Learning Models for Reaction Yield Prediction. *Molecular Informatics*, 41(12):2200043, 2022. doi: 10.1002/minf.202200043.

Hongchen Wang, Rafael Espinosa Castañeda, Jay R. Werber, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. Training-free active learning framework in materials science with large language models. *npj Computational Materials*, 12(1):265, 2026. doi: 10.1038/s41524-026-02136-4.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent Workflow Memory. In *International Conference on Machine Learning*, volume 267, pp. 63897–63911, 2025. URL https://proceedings.mlr.press/v267/wang25bx.html.

Rui Xin, Edirisuriya M. D. Siriwardane, Yuqi Song, Yong Zhao, Steph-Yves Louis, Alireza Nasiri, and Jianjun Hu. Active-Learning-Based Generative Design for the Discovery of Wide-Band-Gap Materials. *The Journal of Physical Chemistry C*, 125(29):16118–16128, 2021. doi: 10.1021/acs.jpcc.1c02438.

Zhuo Yang, Yeyun Chen, Jiaqing Xie, Ben Gao, Shuaike Shen, Wanhao Liu, Liujia Yang, Beilun Wang, Tianfan Fu, and Yuqiang Li. MolAct: An Agentic RL Framework for Molecular Editing and Property Optimization. *arXiv preprint arXiv:2512.20135*, 2025. URL https://arxiv.org/abs/2512.20135.

Channing Yu, Aristotle M Mannan, Griselda Metta Yvone, Kenneth N Ross, Yan-Ling Zhang, Melissa A Marton, Bradley R Taylor, Andrew Crenshaw, Joshua Z Gould, Pablo Tamayo, Barbara A Weir, Aviad Tsherniak,

Bang Wong, Levi A Garraway, Alykhan F Shamji, Michelle A Palmer, Michael A Foley, Wendy Winckler, Stuart L Schreiber, Andrew L Kung, and Todd R Golub. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nature Biotechnology*, 34(4):419–423, 2016. doi: 10.1038/nbt.3460.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025. doi: 10.1038/s41586-025-08628-5.

Junming Zhang, Shuwen Xiong, Yugui Xu, and Yongqing Zhang. Prediction of cancer drug response based on heterogeneous graph neural networks and multi-omics data. *Neural Networks*, 193:108001, 2026. doi: 10.1016/j.neunet.2025.108001.

Liang Zhang, Jiahui Du, Zikai Xie, Linjiang Chen, Wenlang Li, Weihang Geng, Yi Zhou, Xinwen Ou, Chengtao Gong, Yijun Gao, Shan He, Chunxing Yan, Chengbin Zhao, Yantian Jiao, Sheng-Yi Yang, Bing Huang, Jacky W. Y. Lam, Jun Qian, Jun Jiang, Ben Zhong Tang, and Hexiang Deng. Discovery of highly fluorescent covalent organic frameworks through AI-assisted iterative experiment–learning cycles. *Nature Chemistry*, 17(11):1645–1654, 2025. doi: 10.1038/s41557-025-01974-x.

Yanan Zhao, Xiaochen Liu, Han Lu, Xuefeng Zhu, Tianhang Wang, Gen Luo, Rencheng Zheng, and Yi Luo. An optimized deep convolutional neural network for yield prediction of Buchwald-Hartwig amination. *Chemical Physics*, 550:111296, 2021. doi: 10.1016/j.chemphys.2021.111296.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. SkillWeaver: Web Agents can Self-Improve by Discovering and Honing Skills. *arXiv preprint arXiv:2504.07079*, 2025. URL `https://arxiv.org/abs/2504.07079`.

A Additional Method Details

Algorithm 1 Experimentation and intra-task evolution

Require: task $\mathcal{T}$, approved workflow W, applicable assets M

Initialize observed data D and research record τ
while unfinished stages remain and execution can continue **do**
  Select a stage with available inputs; retain its approved plan P or obtain an initial plan
  Lab Operator executes P and appends observations and diagnostics to D, τ
  if a stage boundary, stagnation, or missing output triggers review **then**
    Scientific reviewer checks the scientific results; AI reviewer checks the method
    Chair reconciles the reports and chooses continuation, completion, or revision
    if revision is requested **then**
      Route each requested change to the adaptive skill, Experiment Designer, or Research Designer
      Obtain the corresponding parameter update, AI plan, or revised unfinished workflow
      Validate the requested change; obtain required scientist approval after any resubmission
      Apply an accepted change, retain the incumbent otherwise; append the decision to τ
    end if
    if continuation or completion is proposed **then**
      Check action–agent consistency and evidence references with the harness
      For completion, also check that required scientific outputs satisfy the stage goal
      if a check fails **then**
        Retain the current plan; return failed requirements to review and evolution
      else
        Continue the current plan or accept completion as checked
      end if
    end if
  end if
end while
return scientific result O, observed data D, and research record τ

Algorithm 2 Curating experience and transferring it to the next task

Require: completed record $\mathcal{H}_q = (W_q, P_q, O_q, \tau_q)$, library M_q, next task $\mathcal{T}_{q+1}$

$A \leftarrow \text{Curate}(\mathcal{H}_q)$ with source evidence and applicability conditions
Use the harness to check source records, dependencies, and executable artifacts
Submit candidates that pass these checks to the configured approval gate
while human supervision is enabled and the scientist returns candidates for revision **do**
  Curator revises the returned candidates using the feedback
  Recheck the revisions with the harness; resubmit candidates that pass
end while
$M_{q+1} \leftarrow M_q \cup A^{\text{approved}}$; exclude unapproved candidates
$R \leftarrow \text{Retrieve}(\mathcal{T}_{q+1}, M_{q+1})$; $R_{\text{use}}, R_{\text{defer}} \leftarrow \varnothing$
for each asset $a \in R$ **do**
  Route a to its receiving agent; choose direct reuse, adaptation, or rejection
  if the agent rejects a **then**
    Record the reason; **continue**
  end if
  Construct the proposed use of a, adapting it when requested
  if required target observations are unavailable **then**
    Add a to R_{defer} for validation when observations become available
  else if applicability, interface, and required validation checks pass **then**
    Add the accepted use of a to R_{use}
  else
    Reject the proposed use and record the failed check
  end if
end for
Designers build the next workflow and AI plans using R_{use}
Execution rechecks deferred skills on observed target data and logs actual use
return M_{q+1}, next workflow and plans, and the asset-use record

Bounded numeric strategy updates. For a permitted numeric strategy parameter θ, the adaptive service materializes a directional proposal as

$$\theta' = \mathrm{clip}_{[\ell,u]}(\theta + d(u-\ell)\eta s\kappa)\,. \tag{7}$$

Here, $[\ell, u]$ is the registered admissible range, $d \in \{-1, +1\}$ is the proposed direction, $\eta \in (0, 1]$ limits the fractional step, $s \in (0, 1]$ is the proposed strength, and $\kappa \in [0, 1]$ is the proposal's confidence score. The latter is a control input, not a calibrated probability. Three gates check legality and evidence references, an actual change in decisions or behavior, and scientific utility under a registered validation procedure. Missing or incomplete challenger evidence leaves the current strategy unchanged. This rule concerns local numeric adjustments; method and workflow changes use their own compatibility, evidence, and approval checks.

B Task Definitions, Evaluation Protocols, and Complete Results

B.1 Task Identifiers

Each identifier combines the domain, experimental setting, and task number. For example, HTE-1.2-1 is the first HTE task in E1.2. Identifiers distinguish scientific tasks and settings; repeated appearances of an identical comparison use the same identifier. E2.2 reuses the three E2.1 task identifiers and the same no-asset/full-asset controls. A slash (/) denotes information not separately listed in these tables. Candidate-pool outcomes are revealed only through permitted measurements; evaluation may use the full reference pool after execution.

HTE yield aggregation. For discovery tasks, yield is the highest measured value. Confirmation tasks require all six additive measurements for a ligand–base combination c; its score is $r(c) = (y_{c,(1)} + y_{c,(2)})/2$, where $y_{c,(1)} \le \cdots \le y_{c,(6)}$. Tasks requesting one or two confirmed combinations report the highest or second-highest confirmed score, respectively. Pool optima use the same aggregation.

Table 4: Task identifiers: individual and continual studies.

Task ID	Scientific object	Candidate pool / observations	Endpoint
HTE-1.1-1	HTE yield prediction	/	RMSE ↓; R^2 ↑
HTE-1.1-2	Published HTE reaction-search pool	792 candidates; 44 features; initially no yields; 50 measurements	Yield ↑
PRISM-1.1-1	PRISM A549 primary screen	4,683 candidates; 469 screening-active compounds; /	Hits ↑
PRISM-1.1-2	PRISM A549 primary/secondary intersection	1,378 candidates; 138 screening-active compounds; /	Hits ↑
OC20-1.1-1 / OC20-1.1-2	OC20Dense adsorption searches	/	Energy gap ↓
HTE-1.2-1	HTE010; high-yield discovery	72 candidates; /	Yield ↑
HTE-1.2-2	HTE011; one confirmed condition	72 candidates; /	Yield ↑
HTE-1.2-3 / HTE-1.2-4	HTE014 / HTE015; two confirmed conditions	72 candidates each; /	Yield ↑
PRISM-1.2-1 / PRISM-1.2-2	HCC827 → A549	512 candidates each; 64 / 32 measurements	Hits ↑

B.2 Protocol for Workflow and AI-Plan Evaluation

Tasks and evaluation units. The tasks adapt published reaction prediction/optimization, compound-screening, and adsorption-search problems (Ahneman et al., 2018; Shields et al., 2021; Corsello et al., 2020; Lan et al., 2023). For each domain, ten scientist requests $\mathcal{T}_i = (G_i, I_{0,i}, C_i)$ vary in scientific goals and resource constraints. Each request is given to all six LLMs, yielding 30 distinct tasks and 180 task–model proposals, with 60 proposals per domain. Each proposal contains a scientific workflow and its AI plan.

Table 5: Task identifiers: asset studies; E2.2 shares the E2.1 settings.

Task ID	Scientific object	Candidate pool / observations	Endpoint
HTE-2.1-1	HTE015; E2.1 and E2.2 target	72 candidates; 36 measurements	Yield ↑
PRISM-2.1-1	HCC4006; E2.1 and E2.2 target	512 candidates; 60 measurements	Hits ↑
OC20-2.1-1	OC20Dense transfer target	`ood_both::6_10420_2`; 16 measurements; /	Energy gap ↓
HTE-2.3-1	HTE013; asset-accumulation target	288 candidates; six initial observations; 50 total measurements	Yield ↑
HTE-2.3-S1 / S2 / S3	HTE014 / HTE011 / HTE010 source studies	48 observations per source; 144 in total	Source assets

Experts and correctness criteria. Three domain experts and three AI experts independently assess the workflows and AI plans, respectively. All are master's- or doctoral-level researchers in the relevant fields. A workflow is correct when at least two of the three domain experts agree; an AI plan is correct when at least two of the three AI experts agree. Workflow correctness concerns whether the stages, dependencies, and required outputs address the scientific goal and constraints. AI-plan correctness concerns whether the AI objective and proposed solution correspond to those requirements. Multiple valid solutions are admissible; this assessment does not require a unique model choice or optimal parameters.

Versions and corrections. We distinguish the initial proposal, the proposal after harness-triggered system correction, and any subsequent expert modification. Each domain uses the same 60-proposal denominator across versions. Initially, 179 workflows and 173 AI plans are judged correct. System correction resolves six AI-plan errors. The remaining OC20 proposal requires one expert to correct both its workflow and AI plan. Thus, 179/180 proposals have both components correct before human modification, and all 180 are correct after the final expert modification.

The HTE error concerns an incorrectly formulated AI objective. The four PRISM errors concern AI plans, and the two OC20 errors comprise one AI-only error and one joint workflow–AI error.

B.3 Protocol for the Human-Expert Comparison

Participants and collaboration. Each domain is studied by two domain researchers and two AI researchers at master's or doctoral level. The same two AI researchers participate in all three domains, giving six domain researchers and two AI researchers in total. Domain researchers handle workflow design and scientific review; AI researchers handle AI-plan design, execution, and AI-method review. They analyze the current evidence and discuss the next research direction. The experts do not use LLMs.

One human baseline per domain. Each team completes its domain's task once. The resulting human references are 44.5416% yield for HTE-4-1, zero confirmed hits for PRISM-4-1, and 0.2866 eV gap for OC20-4-1. Each reference is compared with the six framework LLM conditions for that task. The experiment therefore contains three human investigations and 18 framework–reference comparisons. Figure 7 displays the references as horizontal dashed lines, and Table 16 reports each human result once.

Task inputs and measurement conditions. The human and framework conditions use the same scientific request, candidate pool, initial observations, permitted operations, resource allowances, and evaluation rule. Unknown responses become available only when the corresponding measurement is selected. Table 6 specifies the task conditions; initial observations count toward the total measurement allowance. Execution stops when the task is completed or no further permitted measurement can be made.

Framework approval setting and outcomes. In RQ4, framework human-approval gates pass automatically; experts do not approve or revise proposals through these gates. This differs from the supervised case demonstrations, where human reviews pass. The framework improves 12 of the

Table 6: RQ4 task definitions and shared measurement allowances. PRISM uses assay-cost units: a primary screen costs one unit and a dose-response confirmation costs eight.

Task	Scientific instance and initial observations	Total allowance and permitted measurements
HTE-4-1	HTE014, 3-bromopyridine; 72 candidates and six initial yield observations	30 reaction measurements; confirm two ligand–base combinations using all six additive conditions
PRISM-4-1	A549 primary/secondary intersection; 1,378 compounds, 20 primary observations and two secondary confirmations	84 assay-cost units (36 already used initially); primary screening and secondary confirmation of screened compounds
OC20-4-1	OC20Dense `ood_both::41_9937_40`; 51 configurations and four initial observations	12 published DFT-result queries; candidate descriptors and cached EquiformerV2 predictions are available

18 model–task comparisons and ties six. These are endpoint comparisons against three single-run references; equal zero-hit results indicate equal endpoints, not successful hit discovery.

B.4 Complete Numerical Outcomes

Table 7: E1.2 asset use in the Qwen sequences. W: Workflow Playbook; P: Policy Prior; S: executable Skill. Entries describe recorded adoption, rather than retrieval alone.

Current task	Source and asset types	Use in the current task
HTE-1.2-1	No earlier task	Initialize high-yield discovery
HTE-1.2-2	HTE-1.2-1: W, P	Adapt search to confirm one condition across six environments
HTE-1.2-3	HTE-1.2-2: W, P; HTE-1.2-1: P	Adapt complete-profile measurement and ranking to two conditions
HTE-1.2-4	HTE-1.2-2/HTE-1.2-3: W, P; HTE-1.2-1: P	Reuse two-condition workflow; adapt selection and ranking policies
PRISM-1.2-1	No earlier task	Initialize compound screening
PRISM-1.2-2	PRISM-1.2-1: W, P, S	Adapt screening workflow and policy; execute the calibrated source predictor

Table 8: RQ2, E2.1: full values for asset and read-only-history comparisons. HTE-2.1-1: yield (%) ↑; PRISM-2.1-1: active hits ↑; OC20-2.1-1: gap (eV) ↓.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
HTE-2.1-1: No assets	56.934	67.6277	67.6277	56.934	67.6277	56.934
HTE-2.1-1: Full assets	88.0506	88.0506	88.0506	88.0506	88.0506	88.0506
HTE-2.1-1: Read-only history	56.934	56.934	56.934	67.6277	56.934	21.659
PRISM: No assets	15	13	15	11	12	18
PRISM: Full assets	18	15	19	19	20	23
PRISM: Read-only history	17	13	15	10	13	13
OC20: No assets	0.2068	0.0002	0.0002	0.7624	0.0357	0.0002
OC20: Full assets	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
OC20: Read-only history	0.0002	0.0002	0.0002	0.7624	0.2068	0.25

Table 9: RQ2, E2.2: leave-one-out and single-type asset ablations on HTE-2.1-1, yield (%) ↑.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
No assets	56.934	67.6277	67.6277	56.934	67.6277	56.934
Full assets	88.0506	88.0506	88.0506	88.0506	88.0506	88.0506
−Workflow	66.7488	88.0506	88.0506	88.0506	88.0506	88.0506
−Policy	87.9792	88.0506	88.0506	88.0506	67.6277	87.9792
−Skill	88.0506	88.0506	88.0506	88.0506	88.0506	87.9792
−Intuition	88.0506	56.934	56.934	67.6277	87.9792	67.6277
Only workflow	67.6277	56.934	56.934	88.0506	87.9792	56.934
Only policy	88.0506	67.6277	67.6277	66.7488	56.934	56.934
Only skill	88.0506	67.6277	67.6277	67.6277	87.9792	67.6277
Only intuition	88.0506	88.0506	88.0506	88.0506	66.7488	88.0506

Table 10: RQ2, E2.2: leave-one-out and single-type asset ablations on PRISM-2.1-1, active hits ↑.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
No assets	15	13	15	11	12	18
Full assets	18	15	19	19	20	23
−Workflow	18	14	18	18	11	17
−Policy	17	12	18	19	13	18
−Skill	14	12	11	14	14	17
−Intuition	18	12	5	17	16	17
Only workflow	15	15	16	11	10	10
Only policy	13	14	13	13	11	23
Only skill	16	13	19	17	20	23
Only intuition	17	12	15	14	18	17

Table 11: RQ2, E2.2: leave-one-out and single-type asset ablations on OC20-2.1-1, energy gap (eV) ↓.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
No assets	0.2068	0.0002	0.0002	0.7624	0.0357	0.0002
Full assets	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−Workflow	0.0357	0.0002	0.0002	0.0002	0.0002	0.748
−Policy	0.0357	0.0002	0.0002	0.0002	0.0002	0.0002
−Skill	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−Intuition	0.0002	0.0002	0.0002	0.0357	0.0002	0.0002
Only workflow	0.036	0.0002	0.0002	0.0002	0.0002	0.0002
Only policy	0.0002	0.0002	0.0002	0.7624	0.0002	0.0002
Only skill	0.0357	0.0002	0.0002	0.0357	0.0357	0.7624
Only intuition	0.2068	0.0002	0.0002	0.0002	0.0002	0.0002

Table 12: RQ3: complete four-branch outcomes across six models for HTE. Yield in both AI and SCI cases, in % and higher-is-better. NR indicates an unmet confirmation requirement, not zero yield.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
HTE-3-1: None	57.5399	57.5399	57.5399	57.5399	57.5399	57.5399
HTE-3-1: AI	68.2481	57.5399	57.5399	68.2481	57.5399	57.5399
HTE-3-1: Sci.	57.5399	57.5399	57.5399	57.5399	57.5399	57.5399
HTE-3-1: Full	68.2481	68.2481	68.2481	68.2481	57.5399	57.5399
HTE-3-2: None	NR	NR	NR	NR	NR	NR
HTE-3-2: AI	NR	NR	NR	NR	NR	NR
HTE-3-2: Sci.	86.7009	86.7009	86.7009	86.7009	86.7009	86.7009
HTE-3-2: Full	86.7009	86.7009	86.7009	86.7009	86.7009	86.7009

Table 13: RQ3: complete four-branch outcomes across six models for PRISM. Hits: screening top-10% (Case 1) and dose-confirmed top-20% (Case 2); higher is better.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
PRISM-3-1: None	8	8	8	8	8	8
PRISM-3-1: AI	10	9	10	10	8	9
PRISM-3-1: Sci.	11	11	8	11	8	8
PRISM-3-1: Full	11	11	10	11	9	9
PRISM-3-2: None	4	4	4	4	4	4
PRISM-3-2: AI	4	4	4	4	4	4
PRISM-3-2: Sci.	4	9	9	8	8	4
PRISM-3-2: Full	4	9	9	8	8	4

Table 14: RQ3: complete four-branch outcomes across six models for OC20. Energy gap (eV); lower is better.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
OC20-3-1: None	0.2547	0.2547	0.2547	0.2547	0.2547	0.2547
OC20-3-1: AI	0.028	0.2547	0.2547	0.028	0.028	0.2547
OC20-3-1: Sci.	0.2547	0.2547	0.2547	0.2547	0.2547	0.2547
OC20-3-1: Full	0.028	0.2547	0.2547	0.028	0.028	0.028
OC20-3-2: None	0.5415	0.5415	0.5415	0.5415	0.5415	0.5415
OC20-3-2: AI	0.5415	0.5415	0.5415	0.5415	0.5415	0.5415
OC20-3-2: Sci.	0.5415	0	0	0	0	0.5415
OC20-3-2: Full	0	0	0	0	0	0

Table 15: RQ2, E2.3: accumulated experience on HTE-2.3-1, yield (%) ↑. M_0–M_3 correspond to zero through three source-task histories.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
M_0	65.022	65.022	65.022	65.022	65.022	65.022
M_1	65.022	65.022	65.022	65.022	65.022	65.022
M_2	68.765	65.022	65.022	65.022	65.022	65.022
M_3	68.765	65.022	65.3367	68.765	65.022	65.022

Table 16: RQ4: one human-team result per domain and six framework conditions. Columns correspond to HTE-4-1 (yield %, ↑), PRISM-4-1 (confirmed hits, ↑), and OC20-4-1 (gap in eV, ↓).

Condition	HTE-4-1	PRISM-4-1	OC20-4-1
Human expert team	44.5416	0	0.2866
Framework (Q)	44.5416	0	0
Framework (G6)	44.5416	2	0
Framework (G5.6)	67.1813	3	0
Framework (Ge)	44.5416	2	0
Framework (DS)	67.1813	2	0
Framework (K3)	67.1813	0	0.2866

C Supplementary Experiments: RQ5 and RQ6

C.1 RQ5: Transfer Across Language Models

To test cross-model asset transfer on PRISM, we equip Qwen3.6-Flash with strong-model assets and compare against weak-model baselines and strong models using their own assets (Figure 8). Hits

rise from five to 6–10; five donors exceed the weak own-asset baseline of six. Four settings match the strong-model reference; the remaining two obtain 10 versus 13 and eight versus nine hits. These results support asset transfer that brings weak-model performance to or near strong-model levels (exact results: Table 17).

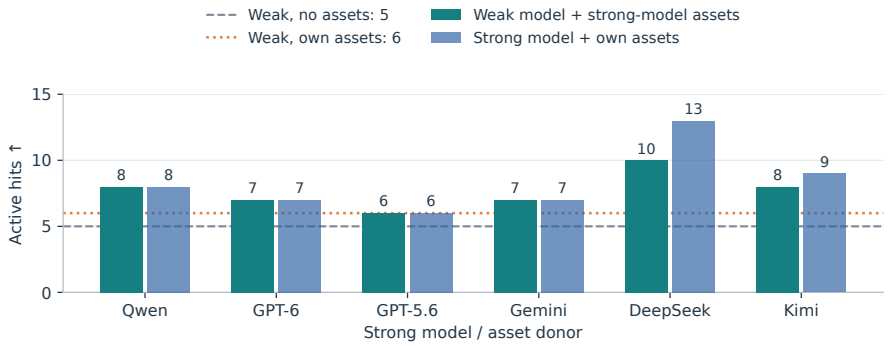

Figure 8: RQ5: strong-to-weak asset transfer (hits ↑). Lines mark weak-model baselines.

RQ5 fixes Qwen3.6-Flash as the weak model on PRISM-5-1. The six columns in Table 17 identify strong-model donors and their corresponding references. Weak-model no-asset and own-asset values repeat the same baseline conditions across columns, not independent runs. Donor assets are supplied to the weak model without updating its weights. Figure 8 presents the comparison; Table 17 provides the exact endpoints.

Table 17: RQ5: active hits on PRISM-5-1 ↑. Columns identify the strong-model asset donor; the weak model is fixed. Repeated weak baselines are references, not independent repetitions.

Metric / condition	Q	G6	G5.6	Ge	DS	K3
Weak, no assets	5	5	5	5	5	5
Weak, own assets	6	6	6	6	6	6
Weak, donor assets	8	7	6	7	10	8
Strong, own assets	8	7	6	7	13	9

C.2 RQ6: Case Studies

The first case is E1.1 task HTE-1.1-2, following the scientist's request through model-guided measurements and method reselection. The second follows E2.3: assets from three completed HTE studies guide active learning on a new substrate. An MP4 walkthrough and the interactive webpage accompany these cases in the supplementary materials; the webpage is available at the anonymous case-study site. The numbered steps specify the responsible agents and their inputs and outputs; symbols instantiate Sections 3–5, and numerical values come from the recorded executions.

C.2.1 Case 1: Discovering High-Yield Reaction Conditions in HTE-1.1-2

This task uses an active-learning loop for reaction optimization: fit a surrogate on measured yields, select informative or promising unmeasured conditions, measure them, and refit. The scientific output is a condition with experimentally observed high yield.

Phase 1: Task Formulation

Step 1: Human scientist—specifying the discovery task.

- **Goal** G_{H}**.** Discover high-yield Buchwald–Hartwig reaction conditions; the recommended condition must have a measured yield.
- **Constraints** C_{H}**.** Cover all levels of the reaction factors in the candidate pool before search; use only queried yields, do not measure the same condition twice, and perform at most 50 measurements.
- **Initial information** $I_{0,\mathrm{H}}$**.** Candidate pool $\mathcal{X}$ with $|\mathcal{X}| = 792$, 44 encoded features per condition, the factor schema, and a measurement interface. The factors comprise three aryl halides, four catalysts, three bases, and 22 additives. Initially $D_0 = \varnothing$; unqueried $y(x)$ is unavailable.
- **Task output.** $\mathcal{T}_{\mathrm{H}} = (G_{\mathrm{H}}, I_{0,\mathrm{H}}, C_{\mathrm{H}})$ is passed to the Research Designer. The observed data, best measured yield, and recommended measured condition are represented by:

$$D_t = \{(x_i, y_i)\}_{i=1}^{n_t}, \qquad b_t = \max_{(x_i,y_i)\in D_t} y_i, \qquad x_t^* \in \arg\max_{x_i:(x_i,y_i)\in D_t} y_i. \tag{8}$$

  The full-pool optimum is used only in evaluation after execution.

Phase 2: Investigation and Intra-Task Evolution

Step 1: Research Designer—constructing the workflow.

- **Input.** $\mathcal{T}_{\mathrm{H}}$, public factor schema, and registered measurement operations; $D_0 = \varnothing$. The agent has no measured yields at this step.
- **Processing.** The coverage requirement must be satisfied before model-guided search. The agent therefore separates initial coverage from subsequent optimization and specifies the observations passed between them.
- **Output.** $W^{(1)} = (S^{(1)}, E^{(1)})$, with two stages:

$$S^{(1)} = \{s_{\mathrm{cov}}, s_{\mathrm{opt}}\}, \qquad E^{(1)} = \{(s_{\mathrm{cov}}, s_{\mathrm{opt}})\}. \tag{9}$$

  For the initial-data collection stage s_{cov}, (g_n, I_n, O_n, c_n) specifies factor-covering measurements, its schema and requirements, measured yields/identifiers/trace, and completion after all required factor levels have been measured. For s_{opt}, the tuple specifies yield improvement, the coverage data and remaining candidates, additional measurements and the search trace, and completion subject to valid nonrepeated measurements and the required trace.

Step 2: Harness and human scientist—reviewing the workflow.

- **Input.** The structured $W^{(1)}$, $\mathcal{T}_{\mathrm{H}}$, and available operations.
- **Processing.** The harness checks schema, dependencies, input references, and available operations. The workflow passed programmatic validation and human scientific review.
- **Output.** The approved workflow $W^{(1)}$ is passed to execution.

Step 3: Experiment Designer and Lab Operator—obtaining initial observations.

- **Input.** s_{cov}, the factor schema, and permitted measurement skills.
- **Processing.** The Experiment Designer uses the factor levels to select 22 conditions that jointly cover all required levels. The Lab Operator measures their yields. Thus, 22 is the number selected by the initial-data collection plan, rather than observations supplied with the task.
- **Output.** The operator returns D_t with $n_t = 22$, candidate identifiers, and the coverage trace. The best measured yield is $b_t = 57.5399\%$. These observations enter s_{opt}.

Step 4: Experiment Designer—selecting the initial AI plan.

- **Input.** s_{opt}, the 22 observed condition–yield pairs, and compatible registered skills.
- **Processing.** The agent identifies Q_{opt} as an optimization problem and uses active learning (AL) to search for high-yield conditions. The loop fits a surrogate to measured yields, selects new condi-

tions using an acquisition rule, measures them, and updates the model. It proposes three surrogate–acquisition combinations: RF-LCB, GP-EI, and bootstrap-SVR-LCB; strategy initialization selects RF-LCB for execution.

- **Output.** $P_{\text{opt}}^{(0)}$: RF-LCB with 150 trees, square-root feature sampling, and batches of four conditions. The plan is passed to the Lab Operator.

Step 5: Lab Operator—executing the active-learning loop.

- **Input.** $P_{\text{opt}}^{(0)}$, observed data D_t, and unmeasured conditions $\mathcal{X} \setminus \mathcal{X}(D_t)$.
- **Processing.** Fit f_t on the measured condition–yield pairs. Predict the mean $\mu_t(x)$ and uncertainty $\sigma_t(x)$ for each unmeasured condition x, and use acquisition score $\alpha_t(x)$ to select the next k_t conditions. The Lab Operator queries only these selected conditions and refits on the enlarged data before the next selection:

$$\begin{aligned} f_t &= \text{Fit}(D_t; P_{\text{opt}}), \quad \mathcal{U}_t = \mathcal{X} \setminus \mathcal{X}(D_t), \\ (\mu_t(x), \sigma_t(x)) &= \text{Predict}(f_t, x), \quad x \in \mathcal{U}_t, \\ \mathcal{B}_t &= \text{Top}_{k_t}\{\alpha_t(x) : x \in \mathcal{U}_t\}, \\ D_{t+1} &= D_t \cup \{(x, y(x)) : x \in \mathcal{B}_t\}. \end{aligned} \tag{10}$$

  Here, $\mathcal{X}(D_t)$ contains measured conditions and Top_{k_t} returns the conditions with the largest acquisition scores. The registered RF-LCB skill accounts for the objective direction: for yield maximization, $\alpha_t(x) = \mu_t(x) + 1.96\sigma_t(x)$ and $k_t = 4$. Predictions favor promising yields, while uncertainty allows exploration. The first guided batch has a maximum acquisition score of 60.6317; its four measured yields are 53.1452%, 55.0122%, 49.3239%, and 50.1038%. Acquisition scores are computed before the measurements and are not observed yields.
- **Output.** Three batches increase the observation count $22 \to 26 \to 30 \to 34$, while b_t remains 57.5399%. Updated observations, selection diagnostics, and trace τ_t are passed to review.

Step 6: Scientific Reviewer—assessing the scientific plan.

- **Input.** G_{H}, required stage outputs, measured yields, and $E^{(1)}$.
- **Processing.** Check whether coverage and workflow structure remain appropriate for high-yield discovery.
- **Output.** R_t^{sci}: retain the workflow and continue collecting evidence.

Step 7: AI-Method Reviewer—assessing the search method.

- **Input.** $P_{\text{opt}}^{(0)}$, selection diagnostics, and three batches without yield improvement.
- **Processing.** Observations increase from 22 to 34, but three batches leave the best yield at 57.5399%, with 16 measurements remaining. The reviewer interprets high RF-LCB uncertainty without yield improvement as a possible exploration–exploitation mismatch. It recommends the registered, untried GP-EI alternative to change both the surrogate and acquisition rule.
- **Output.** R_t^{AI} recommends method reselection based on observed stagnation and available alternatives, proposing GP-EI with bootstrap-SVR-LCB as a backup.

Step 8: Review Chair—assigning a revision.

- **Input.** R_t^{sci}, R_t^{AI}, D_t, and previous attempts in τ_t.
- **Processing.** Reconcile the recommendation to retain the scientific workflow with the recommendation to change the AI method.
- **Output.** Action d_t assigns method reselection to the Experiment Designer, with evidence $\mathcal{E}_t$:

$$d_t = \text{Chair}(R_t^{\text{sci}}, R_t^{\text{AI}}, D_t, \tau_t) = (\text{reselect method}, \text{Experiment Designer}, \mathcal{E}_t). \tag{11}$$

  The stage structure of $W^{(1)}$ remains unchanged. The proposed change passes programmatic validation and human review.

Step 9: Experiment Designer—constructing the revised plan.

- **Input.** d_t, all 34 observations, and compatible alternatives.
- **Processing.** Select GP-EI and specify the executable model settings; retain existing observations.
- **Output.** $P_{\text{opt}}^{(1)}$ selects GP-EI with one condition per iteration. The registered GP uses a Matérn-$5/2$ kernel and normalized targets; expected improvement (EI) replaces confidence-bound acquisition. The GP posterior and current best measured yield b_t determine the next condition through:

$$\alpha_t^{\text{EI}}(x) = (\mu_t(x) - b_t)\Phi(z_t(x)) + \sigma_t(x)\phi(z_t(x)), \qquad z_t(x) = \frac{\mu_t(x) - b_t}{\sigma_t(x)}. \tag{12}$$

  Here, Φ and ϕ are the standard normal CDF and density. The execution skill computes EI for every unmeasured condition and chooses the largest value; the LLM selects the method, while the skill performs its numerical fit and acquisition.

Step 10: Lab Operator—continuing discovery and reporting the outcome.

- **Input.** $P_{\text{opt}}^{(1)}$ and the 34 measured examples.
- **Processing.** Fit the GP on all 34 observations and choose the condition with the highest EI (3.5243). Its subsequently measured yield is 50.8377%; after adding it and refitting, the next EI-maximizing condition (EI 3.4970) yields 68.2481%. The new measurements, selected using fitted predictions, update the incumbent:

$$\max(57.5399, 50.8377) = 57.5399, \qquad \max(57.5399, 68.2481) = 68.2481. \tag{13}$$

- **Output.** O_{H} contains the recommended measured condition x^*, its 68.2481% yield, and the search trace. This value is first reached at measurement 36; execution continues to 50, where the terminal guard reports exhausted measurement allowance. The fitted model therefore guides successive measurements toward the scientific goal.

Table 18: Selected HTE-1.1-2 acquisition steps. Scores precede measurement; the two acquisition functions have different scales. Yield values are percentages.

Observed before	Method / selected	Largest score	Best new yield	Best so far
22	RF / 4	60.6317	55.0122	57.5399
34	GP-EI / 1	3.5243	50.8377	57.5399
35	GP-EI / 1	3.4970	68.2481	68.2481

Phase 3: Experience Curation and Inter-Task Evolution

Step 1: Research Curator interface—identifying the available history.

- **Input $\mathcal{H}_{\text{H}}$.** $\mathcal{H}_{\text{H}} = (W^{(1)}, \{P_{\text{opt}}^{(0)}, P_{\text{opt}}^{(1)}\}, O_{\text{H}}, \tau_{\text{H}})$ contains the workflow, initial and revised plans, measurements, and reviews.
- **Output available in the archive.** This single-task record has no independently archived Research Curator output or asset-transfer result. It documents coverage before search and progress-based method reselection, but these lessons are not counted as generated assets. The next case provides recorded asset generation and use.

C.2.2 Case 2: Historical Assets Guide Active Learning on HTE-2.3-1

This case follows the multi-task web animation and the Qwen3.8-max run in E2.3 (Figure 5). Three completed source studies supply assets for a new reaction-search task, HTE-2.3-1 (HTE013). The target uses 3-chloropyridine and contains 288 candidate conditions.

Preceding studies: what the asset library contains. The source studies concern 3-bromopyridine, 2-bromopyridine, and 2-chloropyridine, recorded as HTE-2.3-S1, HTE-2.3-S2, and HTE-2.3-S3 (HTE014, HTE011, and HTE010), respectively. Each supplies 48 observations, for 144

source observations in total. The Research Curator receives completed workflows, executed plans, and measured outcomes to summarize reusable experience. Exported source predictors and their invocation information are retained as executable skills. The resulting library M_{src} contains four asset types:

- **Workflow** a_W**.** A measure–recommend dependency: scientific recommendations require measured evidence.
- **Policy** a_P**.** The source's executed sampling choices, including a random-sampling policy.
- **Skills** $\mathcal{K}$**.** Previously trained predictors, including a cumulative predictor $h_{1:3}$ fitted on 144 source observations and h_3 fitted on 48 observations for 2-chloropyridine.
- **Intuition** a_I**.** Scoped observations about ligand–base combinations.

These source assets passed programmatic validation and human review. Their use, rejection, and adaptation are traced within the receiving agents below.

Phase 1: Task Formulation

Step 1: Human scientist—defining the new search.

- **Goal** G_{T}**.** Find a reaction condition with a high measured yield for 3-chloropyridine and return that condition with its experimental evidence.
- **Constraints** C_{T}**.** At most 50 measurements, including the six initial observations; no repeated condition; decisions may use only yields already observed. The final recommendation must have a measured yield.
- **Initial information** $I_{0,\text{T}}$**.** The descriptors of 288 candidate conditions, a measurement interface, and $D_6 = \{(x_i, y_i)\}_{i=1}^{6}$. The highest initial yield is 39.1289%. Unmeasured target yields are unavailable to the agents.
- **Output.** $\mathcal{T}_{\text{T}} = (G_{\text{T}}, I_{0,\text{T}}, C_{\text{T}})$, together with the applicable source assets, is passed to the design agents. Source observations are not counted as target measurements.

Phase 2: Investigation and Iterative AI Execution

Step 1: Research Designer—adapting the inherited workflow.

- **Input.** $\mathcal{T}_{\text{T}}$, D_6, the workflow asset a_W, and the conditional intuition a_I.
- **Processing.** Reuse the dependency between evidence and recommendation, while accounting for the six observations already available. The first stage interprets the target observations and relevant source experience; the second acquires further measured evidence. A source-derived preference for a ligand–base combination is recorded as a hypothesis, not a measured target outcome.
- **Output.** The recorded workflow is

$$W_{\text{T}} = (S_{\text{T}}, E_{\text{T}}), \qquad S_{\text{T}} = \{s_{\text{interpret}}, s_{\text{measure}}\}, \qquad E_{\text{T}} = \{s_{\text{interpret}} \to s_{\text{measure}}\}. \tag{14}$$

  The first stage uses existing evidence without additional measurement. Its analysis is passed to the measurement stage; the Experiment Designer determines the executable sampling method.

Step 2: Experiment Designer—selecting source skills and an AL plan.

- **Input.** s_{measure}, D_6, the registered methods, a_P, and available predictor skills $\mathcal{K}$.
- **Processing.** The agent rejects the random-only policy and selects RF-LCB to use the sparse observed yields. It selects $h_{1:3}$ for its broader source coverage and h_3 because 2-chloropyridine is the closest source substrate. It configures 64 target trees, square-root feature sampling, four queries per batch, and zero additional scientific-priority weight. All 44 remaining measurements are assigned to this model-guided stage.
- **Output.** P_{T} specifies the target model, its parameters, and the two predictor invocations. Execution records confirm that the source models are frozen and supply normalized predictions as extra features:

$$\widetilde{x} = \big[x,\ h_{1:3}(x)/100,\ h_3(x)/100\big], \qquad f_t = \text{Fit}_{\text{RF}}\big(\{(\widetilde{x}_i, y_i) : (x_i, y_i) \in D_t\}\big). \tag{15}$$

Only observed target yields supply training labels. Source-model predictions are features, and the source predictors are not refitted during target execution. Thus asset use changes the executable model inputs as well as the proposed research plan.

Step 3: Lab Operator—fitting, scoring, selecting, and measuring.

- **Input.** P_{T}, the current observations D_t, and the unmeasured pool $\mathcal{U}_t = \mathcal{X} \setminus X(D_t)$, where $X(D_t)$ denotes conditions already measured.
- **Processing.** Fit the target RF to all currently observed target labels. For each unmeasured condition, estimate its yield $\mu_t(x)$ and tree-ensemble uncertainty $\sigma_t(x)$. In this yield-maximization execution, the registered RF-LCB skill uses the following positive utility score:

$$a_t(x) = \mu_t(x) + 1.96\sigma_t(x), \qquad Q_t = \mathrm{Top4}_{x \in \mathcal{U}_t}\, a_t(x). \tag{16}$$

  Only the four selected conditions are queried. Their published HTE measurements are then revealed; the true yields are not available when the scores are computed.
- **Output.** Four new condition–yield pairs update the target evidence:

$$D_{t+4} = D_t \cup \{(x, y(x)) : x \in Q_t\}, \qquad b_{t+4} = \max_{(x,y) \in D_{t+4}} y. \tag{17}$$

  The next batch refits the model using this expanded dataset. Eleven batches increase the target observations from six to 50. After each batch, the remaining unmeasured candidates form the choice set for the next iteration.

Step 4: A recorded selection—how AL reaches a better condition.

- **Input.** Before batch 7, $|D_t| = 30$, 258 candidates remain unmeasured, and the best measured yield is 65.0220%.
- **Model decision.** The condition using t-BuXPhos, P2Et, and 4-phenylisoxazole has predicted mean 28.9988, uncertainty 15.8524, and acquisition score 60.0695. This score places it first among the four selected conditions. Its measured yield is still unknown at selection time.
- **Measured output.** Querying this condition reveals 68.7650% yield. The other three selected conditions yield 65.3367%, 8.7766%, and 8.4703%. After the batch, $|D_{t+4}| = 34$ and $b_{t+4} = 68.7650\%$.

Step 5: Lab Operator—returning the result and stopping.

- **Input.** The accumulated observations and the remaining measurement allowance.
- **Processing.** Continue the same fit–select–measure–update loop through 50 observations, then return the highest-yield measured condition. The stopping trigger is exhaustion of the measurement allowance; the best condition was already found in the batch ending at 34 observations.
- **Output.** O_{T} recommends t-BuXPhos/P2Et/4-phenylisoxazole at 68.7650%, with 238 of 288 candidates still unmeasured. With the same six initial observations and 50-measurement limit, the no-asset run reaches 65.0220% (Table 19). The comparison concerns the complete asset-supported plan: assets influence features, parameters, and planning, so it does not isolate AL against random sampling.

Table 19: Recorded best yield (%) in the E2.3 case after completed batches. Both runs start from the same six observations. Higher is better.

Measured conditions	6	14	18	26	34	46	50
No assets	39.1289	40.4824	40.4824	40.4824	40.4824	65.0220	65.0220
Three source studies	39.1289	43.9333	65.0220	65.0220	68.7650	68.7650	68.7650