SWE-BENCH SCIENCE: CAN CODING AGENTS RESOLVE ENGINEERING TASKS IN SCIENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Software increasingly functions as part of the scientific instrument itself, making failures in scientific code capable of compromising not only program behavior but also the evidence underlying scientific conclusions. Yet existing evaluations of coding agents largely emphasize aggregate task success, providing limited insight into why agents fail when repairing scientific software. We introduce **SWE-bench Science**, a repository-level benchmark for scientific software engineering comprising 119 tasks from 98 GitHub repositories across 20 scientific domains. Each task is organized into one of three paradigms: Issue-driven, Expert-exploratory, and Engineering-integration. The best-performing configuration, **GPT-6 Astra with Codex at max reasoning depth, achieves 50.42% Pass@1**, leaving 59 of the 119 tasks unresolved and highlighting the substantial challenges posed by scientific software engineering. We identify four recurring failure mechanisms: deficits in scientific knowledge or abstraction, misguided exploration or surface-level repair, incomplete repair coverage or system integration, and failures to generalize scientific knowledge beyond observed cases in our analysis. We further conduct a paired ablation that removes explicit scientific guidance while preserving the repository and executable engineering context. The results show that scientific knowledge is not uniformly beneficial: well-grounded information can constrain repair and improve average performance and token efficiency, whereas poorly aligned guidance can induce anchoring and does not necessarily improve exact repair success. Across paired conditions, the value of additional information reverses with the configuration's standalone capability, turning negative once deeper reasoning already recovers the scientific contract without guidance. Together, SWE-bench Science provides a broad testbed for studying both the capabilities and failure mechanisms of coding agents in scientific software engineering.

1 INTRODUCTION

Software is no longer merely an aid to science; it is part of the instrument through which scientific claims are produced. Research groups rely on code to run simulations, process instrument outputs, train surrogate models, search chemical or biological candidates, manage high-throughput experiments, and reproduce published results. A defective patch can therefore corrupt not only a program output but also the evidence behind a scientific conclusion.

Understanding why coding agents fail on scientific software is as important as measuring whether their patches pass tests. A failed repair may reflect an incorrect scientific abstraction, superficial exploration of an observed symptom, incomplete integration across a software system, or an inability to generalize a scientific principle beyond visible cases. Aggregate test scores cannot distinguish these sources and therefore provide limited guidance for improving scientific coding agents. Existing benchmarks measure function synthesis and repository-level software repair (Chen et al., 2021; Austin et al., 2021; Jimenez et al., 2024), while scientific coding benchmarks cover curated research problems, paper-grounded implementations, scientific workflows, repository execution, and production scientific codebases (Tian et al., 2024; Xia et al., 2026; Chen et al., 2025; Bogin et al., 2024; Bragg et al., 2026; Duston et al., 2025). However, repository-level coverage across scientific domains and the analysis of failure mechanisms remain limited, leaving broad cross-domain scientific software engineering underexplored.

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.

We introduce **SWE-bench Science**, a repository-level benchmark containing 119 tasks from 98 unique GitHub repositories and spanning 20 scientific domains. Each task is manually inspected to verify its scientific contract, reproducibility, and evaluation validity. We organize the benchmark into three scientific task paradigms—**Issue-driven**, **Expert-exploratory**, and **Engineering-integration**—that capture complementary requirements in scientific software maintenance. Beyond measuring task success, we manually analyze unsuccessful attempts through four recurring scientific failure mechanisms: scientific-knowledge or abstraction deficits, misguided exploration or surface-level repair, incomplete repair coverage or system integration, and failures of scientific-knowledge generalization. Runtime or evaluation-path failures are recorded separately. We further construct a paired comparison that removes explicit scientific auxiliary information while preserving the repository and executable engineering context. The resulting analysis shows that scientific knowledge is not uniformly beneficial: its effect depends on how the knowledge is organized and grounded in executable evidence, because well-targeted information can constrain the repair while poorly aligned information can induce anchoring or substitute for independent validation.

Our contributions are:

1. We introduce a broad-coverage scientific software engineering benchmark with 119 tasks from 98 unique GitHub repositories, spanning 20 scientific domains, and evaluate representative frontier coding agents on these tasks.
2. We analyze unsuccessful scientific software repairs and identify four recurring failure mechanisms, revealing concrete directions for improving scientific abstraction, disciplined exploration, system-wide repair, and scientific-knowledge generalization.
3. We provide a paired scientific-knowledge ablation that removes explicit scientific guidance while preserving the executable engineering context. The results show that the organization and grounding of scientific knowledge matter in scientific software engineering: additional information can improve average scores and token efficiency, but does not automatically improve exact repair success. Its value further depends on the configuration's reasoning depth: the same model gains from the guidance at high depth but is hurt at max depth, where the guidance more than quintuples total input without improving repair success.

2 Related Work

2.1 Code and Software Engineering Benchmarks

Early code-generation benchmarks evaluate whether models can synthesize short functions from natural-language prompts. HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) are representative examples, focusing on unit-test correctness for self-contained programming problems. Later benchmarks increase realism by requiring repository-level modification and issue resolution. SWE-bench (Jimenez et al., 2024) evaluates whether language models can resolve real GitHub issues by generating patches against project test suites. More recent benchmarks extend this setting toward longer-horizon agentic engineering: SWE-Bench Pro targets complex enterprise-level repository tasks, DeepSWE uses original tasks and hand-written verifiers to reduce contamination and oracle ambiguity, and Terminal-Bench evaluates agents on hard, realistic tasks in command-line environments (Deng et al., 2025; Huang et al., 2026a; Merrill et al., 2026). Together, these benchmarks measure increasingly realistic software engineering capabilities, including repository inspection, fault localization, multi-file modification, tool use, and test-validated repair.

2.2 LLMs for Scientific Workflows

Scientific-code benchmarks mainly study two settings. SciCode and MMSciCode evaluate self-contained scientific programming and paper-grounded function implementation, respectively, making them closer to non-agentic programming evaluation than to repository-level software engineering (Tian et al., 2024; Xia et al., 2026). A second line evaluates coding within broader research workflows: ResearchCodeBench and LMR-BENCH study research-paper code implementation and research-code reproduction (Hua et al., 2025; Yan et al., 2025); ScienceAgentBench, SUPER, and CSR-Bench emphasize data-driven scientific discovery, research-repository execution, and research-system deployment (Chen et al., 2025; Bogin et al., 2024; Xiao et al., 2025); and RExBench, Au-

toMat, and AstaBench evaluate research-code extension, scientific workflow recovery, and multi-stage scientific tasks (Edwards et al., 2026; Huang et al., 2026b; Bragg et al., 2026). AInstein-Bench is closest to our setting because it targets scientific computing through issue-derived and synthesized feature tasks in production repositories (Duston et al., 2025). SWE-Bench 5G studies specification-dependent telecom repair (Chen et al., 2026). SWE-bench Science instead centers scientific software engineering as an applied engineering problem: coding agents must inspect production repositories, modify interacting components, and preserve scientific validity across **Issue-driven**, **Expert-exploratory**, and **Engineering-integration** tasks. This focus differs from solving standalone scientific programming problems or completing broader research workflows by evaluating the maintenance and extension of scientifically valid software systems.

Table 1: Comparison of coding-agent benchmarks for science. SWE-bench Science covers 20 scientific domains.

Benchmark	Task count	Scientific domains	GitHub repos	Scientific focus	Task source
SciCode	80	16	–	Function programming	expert
MMSciCode	624	6	–	Function programming	expert
ResearchCodeBench	212	1	20	Research code	expert
LMR-BENCH	28	1	23	Research code	expert
ScienceAgentBench	102	4	30	Research code	expert
SUPER	801	–	694	Research code	expert
CSR-Bench	100	5	100	Research code	–
RExBench	12	–	–	Research code	expert
AutoMat	85	1	–	Research code	expert
AstaBench	>2,400	–	–	Research code	expert
AInsteinBench	244	6	6	Scientific computing	issue/expert
SWE-Bench 5G	210	1	3	Scientific software engineering	issue
SWE-bench Science (ours)	119	20	98	Scientific software engineering	issue/expert/engineer

3 SWE-BENCH SCIENCE

We cover 98 unique GitHub repositories and construct 119 tasks across 20 scientific domains, with the domain and task-paradigm composition shown in Figure 1(a). The five largest domains contain 76 tasks (63.9%), while six domains contribute one task each. The task set contains 52 Issue-driven, 49 Expert-exploratory, and 18 Engineering-integration tasks. The complete scientific-domain counts are reported in Appendix Table 4. The benchmark also spans substantial variation in repository context and repair size. Figure 1(b) shows that non-empty input code averages 80,600.12 lines and ranges from 174 to 2,029,051 lines. Reference patches add 117.81 lines on average, with a range of 1–1,035, and delete 44.53 lines on average, with a range of 0–458. Direct labels preserve the exact values despite the large scale differences among these quantities.

3.1 TASK SCHEMA

The following fields are visible to the agent in the standard condition:

- **Repository snapshot**: the exact pre-task state with locked dependencies and runnable entry points, stripped of Git history, remotes, future changelogs, build caches, and solution-linked artifacts.
- **Problem statement**: a specification frozen before test and reference-patch authors inspect one another's artifacts.
- **Required scientific context** c_i^{req}: local, versioned definitions and constraints required for a well-posed task; it is held fixed in every condition.
- **Public tests**: visible software and scientific assertions that support interactive debugging.

Evaluator-only fields include private tests, contract labels, reference and alternative-valid patches, scientific-rationale and localization support blocks, expected files, source and difficulty metadata,

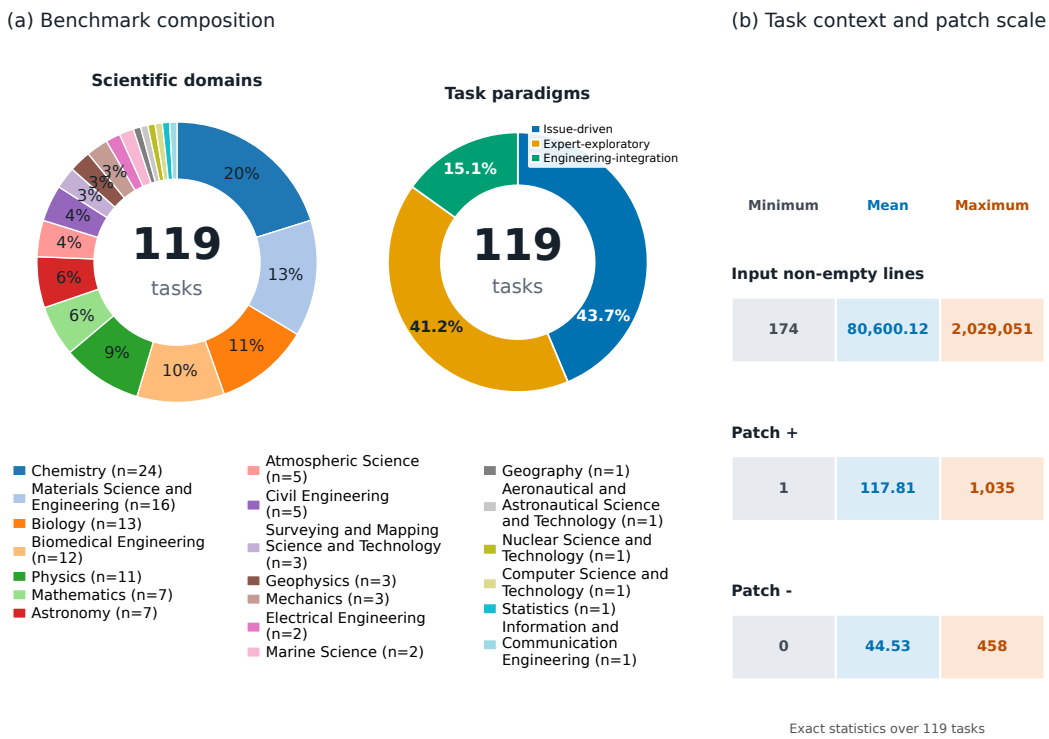


Figure 1: Benchmark coverage and scale over 119 tasks. (a) Scientific-domain distribution and task-paradigm composition. The domain legend reports the task count for each of the 20 domains; the task-paradigm chart contains 52 Issue-driven tasks (43.7%), 49 Expert-exploratory tasks (41.2%), and 18 Engineering-integration tasks (15.1%). Chemistry is the largest domain with 24 tasks, followed by Materials Science and Engineering with 16, Biology with 13, Biomedical Engineering with 12, and Physics with 11. (b) Minimum, mean, and maximum input and reference-patch sizes, shown directly in horizontal summary bands. Input size is measured as non-empty physical lines in the task code, while Patch + and Patch − count additions and deletions in the standard reference diff.

contamination tier, and post-hoc annotations. Private tests are mounted only after patch submission in a separate evaluator container. None of fields is available through the agent workspace, environment variables, logs, or task metadata.

3.2 Example.

Figure 2 illustrates an evaluation pipeline in the benchmark for an example task. The task asks the agent to reconcile two periodic representations of the same crystal while preserving their physical energy. The example shows the agent-visible problem statement, scientific context, and workspace snapshot as the input to the agent-driven coding loop and the clean evaluation phase. Public checks support debugging, whereas private scientific cases determine whether the submitted patch completes the intended repair.

4 Benchmark Construction

To evaluate agent performance across different capability dimensions in real scientific software development and maintenance, we establish a unified Chain-of-Evidence Protocol and derive three task paradigms with clearly differentiated evaluation goals: **Issue-driven**, **Expert-exploratory**, and **Engineering-integration** tasks. Issue-driven tasks emphasize localized repair and regression avoidance, Expert-exploratory tasks require autonomous investigation of scientific discrepancies,

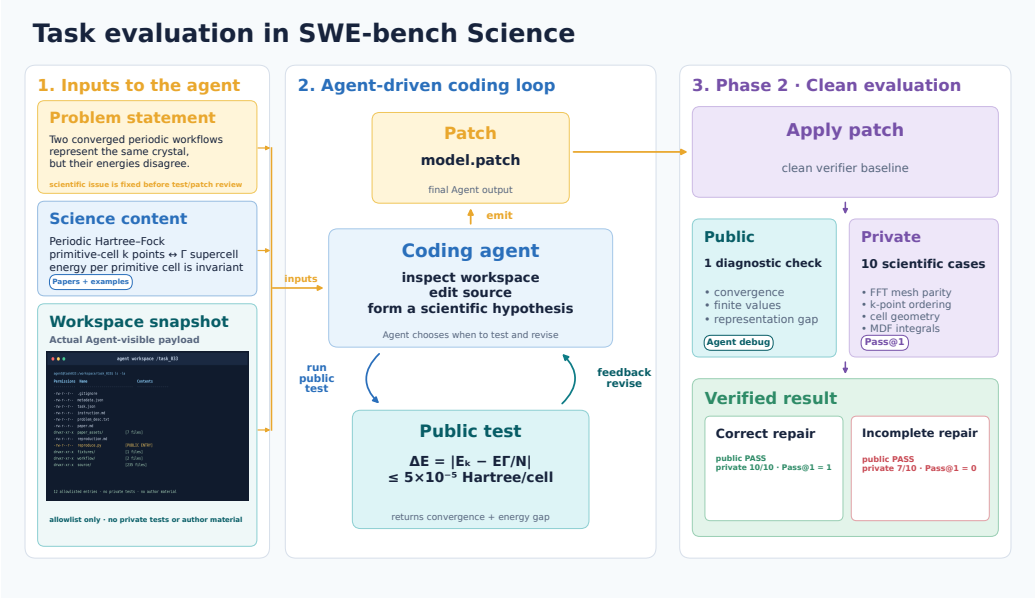


Figure 2: Evaluation pipeline in SWE-bench Science for an example task. The diagram shows the frozen agent-visible inputs, the agent-driven coding loop, and the clean evaluation phase with separate public diagnostics and private scientific cases.

and Engineering-integration tasks evaluate cross-module completion of end-to-end scientific workflows.

Raw tasks pass through four stages: source sampling and screening, snapshot freezing and reproduction, public-material abstraction and information isolation, and hidden oracle and counter-calibration. This process controls environmental interference and information leakage before tasks are re-designed for the three paradigms. The full protocol and construction diagram are in Appendix A.

4.1 Class-Specific Task Re-design

Through differentiated construction strategies, we re-design raw tasks to benchmark three capabilities: repairing known defects (Issue-driven), autonomously reasoning about unknown root causes in scientific scenarios (Expert-exploratory), and connecting a complete multi-module capability chain across files (Engineering-integration). Detailed definitions and construction procedures are given in Appendix B.

4.2 Scientific Auxiliary Information Separation

For 91 of 119 tasks, we compare paired conditions with and without auxiliary scientific information while holding the repository snapshot, execution environment, public reproduction entry points, hidden validators, and required context c_i^{req} fixed. Engineering cues remain available in both conditions; only auxiliary scientific information differs. Appendix C defines the information boundary, Appendix E identifies the 91 tasks, and Section 6.2 reports the results.

5 Experiments

Coding Agents. We evaluate nine agent configurations in our benchmark: GPT-6 Astra and GPT-5.6-sol with Codex at **max** reasoning depth, Claude-Opus-5 with Claude Code at **max**, Kimi-K3 with Kimi Code at **max**, GLM-5.2 with Codex at **max**, Nex N2 with Codex, DeepSeek-V4-flash-0731 with Claude Code at **max**, Qwen3.5-397B with Codex, and DeepSeek-V4-Pro-0813 with Claude Code at **max**. The comparison is shown in Table 2.

Evaluation. We evaluate each submitted patch with public and private tests while keeping private tests and evaluator-only metadata outside the agent workspace. For each task–attempt, PUBLICSCORE is the mean score across all applicable public test cases, and PRIVATESCORE is the corresponding mean across private test cases. FAIL2PASS is the fraction of previously failing private tests that pass after the repair, while PASS2PASS is the fraction of previously passing private tests that remain passing; we set PASS2PASS to 1 when the previously passing set is empty. PASS@1 is a binary exact-private-success measure that equals 1 only when every applicable private test passes. Table 2 reports task-level means for the public score, private score, Fail2Pass, Pass2Pass, and overall Pass@1, together with Pass@1 for Issue-driven, Expert-exploratory, and Engineering-integration tasks.

Table 2: Main experimental results. Scores are task-level means over the common 119 tasks. The best value in each column is bold and underlined; ties are retained.

LLM	Harness	Public score	Private score	Fail2Pass	Pass2Pass	Overall Pass@1	Issue Pass@1	Expert Pass@1	Engineering Pass@1
GPT-6 Astra (max)	Codex	**100.00%**	**83.18%**	**80.11%**	95.95%	**50.42%**	**46.15%**	57.14%	**44.44%**
GPT-5.6-sol (max)	Codex	98.32%	75.57%	69.98%	96.30%	40.34%	34.62%	46.94%	38.89%
Claude-Opus-5 (max)	Claude Code	96.64%	75.11%	68.60%	97.37%	47.90%	38.46%	**65.31%**	27.78%
DeepSeek-V4-Pro (max)	Claude Code	**100.00%**	73.16%	65.77%	96.58%	42.02%	26.92%	57.14%	**44.44%**
Kimi-K3 (max)	Kimi Code	98.32%	66.34%	57.55%	94.94%	35.29%	25.00%	44.90%	38.89%
GLM-5.2 (max)	Codex	94.12%	63.61%	53.81%	**97.53%**	31.93%	17.31%	46.94%	33.33%
Nex N2	Codex	93.28%	61.89%	51.09%	94.92%	24.37%	11.54%	36.73%	27.78%
DeepSeek-V4-flash (max)	Claude Code	98.32%	61.41%	52.34%	95.74%	23.53%	19.23%	26.53%	27.78%
Qwen3.5-397B	Codex	96.64%	51.79%	38.33%	95.16%	14.29%	5.77%	24.49%	11.11%

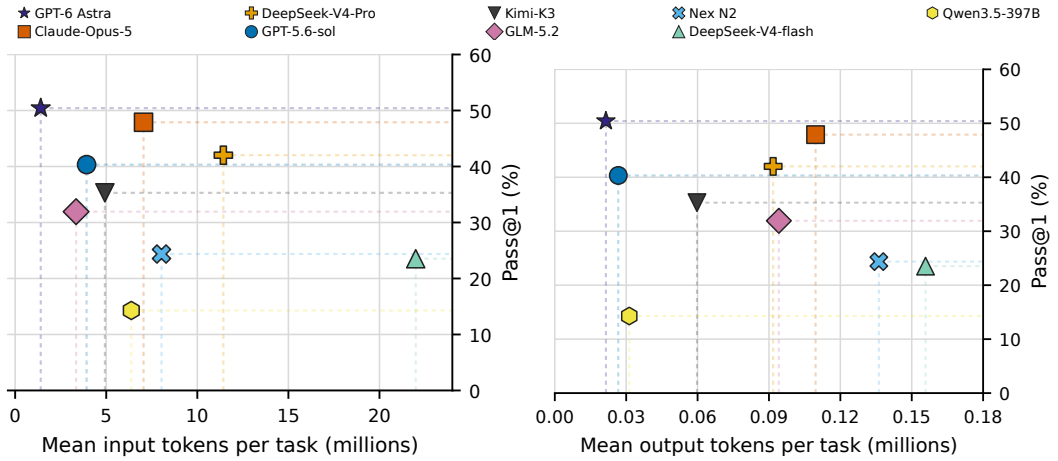
5.1 PERFORMANCE AND TOKEN CONSUMPTION

Table 2 shows that no single model leads every metric. GPT-6 Astra obtains the best private score and FAIL2PASS, and it leads both overall PASS@1 and Issue-driven Pass@1. It also ties DeepSeek-V4-Pro for the best public score and Engineering-integration Pass@1. Claude-Opus-5 remains strongest on Expert-exploratory tasks, while GLM-5.2 attains the best PASS2PASS at preserving tests that already pass. Although GPT-6 Astra is the first evaluated configuration to exceed 50% overall Pass@1, its 50.42% result leaves 59 of 119 tasks unresolved, underscoring the challenge of SWE-bench Science. Figure 3 further shows the Pass@1 score versus token consumption for those agents. GPT-6 Astra attains the highest PASS@1 while reporting a mean of 1.40 million input tokens and 21.48 thousand output tokens per task. Claude-Opus-5 ranks second in exact repair success, while DeepSeek-V4-Pro and GPT-5.6-sol follow with different context and generation budgets. Kimi-K3 and GLM-5.2 achieve lower performance with intermediate token use. For smaller models under 400B, Nex-N2 achieves the best performance–token balance, ahead of DeepSeek-V4-flash and Qwen3.5-397B. Because token accounting differs across harness interfaces, these values should be interpreted as reported consumption for each model–harness configuration rather than as perfectly standardized computational cost.

6 ANALYSIS

6.1 OBSERVED FAILURE MECHANISMS

We use four recurring scientific failure mechanisms in the audit reports. **Scientific-knowledge or abstraction deficit** denotes a repair based on an incorrect or incomplete scientific object, mathematical definition, or domain abstraction. **Misguided exploration or surface-level repair** denotes a patch that addresses the visible symptom or public metric without tracing the failure to an independent oracle or the underlying scientific contract. **Incomplete repair coverage or system integration** denotes a locally plausible repair that does not satisfy the requirements of the full software system: for example, one module is corrected while its interactions, data flow, shared invariants, or compatibility with other modules remain unpreserved. **Failure of scientific-knowledge generalization** denotes a repair that handles the observed scientific case but does not extend the same scientific principle to unseen conditions, equivalent representations, boundary regimes, or other variants that



(a) Mean input-token consumption. **(b)** Mean output-token consumption.

Figure 3: Pass@1 versus mean token consumption per task over the common 119-task evaluation. Dashed guides connect each observation to the corresponding axes.

Table 3: Error-count breakdown by coding agents over the common 119 tasks. The total column records attempts assigned to the four mutually exclusive scientific failure mechanisms defined in the text; for Claude-Opus-5, the parenthesized +4 denotes additional runtime or evaluation-path failures. The four category counts sum to the main total for each row. The lowest count in each column is bold and underlined.

LLM	Harness	Total errors	Knowledge/ abstraction	Exploration/ surface repair	Repair coverage/ system integration	Scientific generalization
GPT-6 Astra (max)	Codex	59	23	**1**	**19**	16
GPT-5.6-sol (max)	Codex	71	18	13	21	19
Claude-Opus-5 (max)	Claude Code	**58 (+4)**	24	2	21	11
DeepSeek-V4-Pro (max)	Claude Code	69	**15**	12	**19**	23
Kimi-K3 (max)	Kimi Code	77	20	14	22	21
GLM-5.2 (max)	Codex	81	20	14	24	23
Nex N2	Codex	90	31	14	23	22
DeepSeek-V4-flash (max)	Claude Code	91	23	14	48	**6**
Qwen3.5-397B	Codex	102	26	15	33	28

require scientific generalization. Claude-Opus-5 additionally has four **safety-guardrail interceptions**: the agent was stopped by a safety guardrail before producing a valid code delivery, so the intended evaluation path could not be completed.

Claude-Opus-5 produces the lowest categorized scientific-error count (58); four additional attempts were stopped by safety guardrails before a valid code delivery was produced. GPT-6 Astra records the fewest misguided-exploration or surface-level-repair errors (1) and ties DeepSeek-V4-Pro for the fewest incomplete-repair or system-integration errors (19). DeepSeek-V4-flash (max) records the fewest scientific-knowledge generalization errors (6), while DeepSeek-V4-Pro records the fewest scientific-knowledge or abstraction errors (15).

6.2 How Scientific Information Affects Performance

On the 91-task separable subset, we compare three configurations with and without scientific information: GPT-5.6-sol (xhigh) with Codex, and DeepSeek-V4-flash with Claude Code at two reasoning depths, high and max, which let the same model be probed at distinct standalone capability levels. Figure 4 summarizes their mean scores and token use.

For GPT-5.6-sol, scientific information slightly increases the mean public and private scores from 96.70% and 73.23% to 97.80% and 74.06%, but decreases Pass@1 from 36.26% to 31.87%. Mean

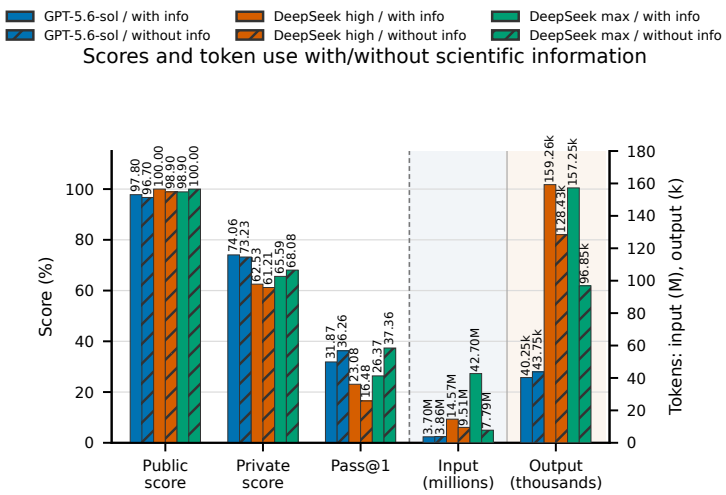
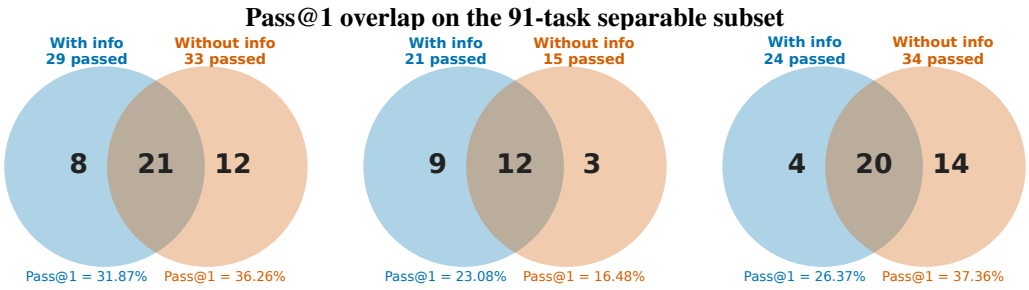


Figure 4: Scores and token consumption over the 91 tasks whose scientific-information content differs between conditions. GPT-5.6-sol and DeepSeek-V4-flash, the latter at high and max reasoning depth, are each shown with and without scientific information on the same task subset. Input and output token use are shown in millions and thousands, respectively; token-bar heights are comparable within each metric.

input and output tokens also decrease from 3.86 million and 43.75 thousand to 3.70 million and 40.25 thousand. For DeepSeek-V4-flash at high reasoning depth, the same intervention raises the public score, private score, and Pass@1 from 98.90%, 61.21%, and 16.48% to 100.00%, 62.53%, and 23.08%, while increasing input and output tokens from 9.51 million and 128.43 thousand to 14.57 million and 159.26 thousand. At max reasoning depth, the direction reverses: introducing the same material lowers the public score, private score, and Pass@1 from 100.00%, 68.08%, and 37.36% to 98.90%, 65.59%, and 26.37%, while raising input and output tokens from 7.79 million and 96.85 thousand to 42.70 million and 157.25 thousand. These paired descriptive differences show model-specific score–cost responses but do not establish statistical significance or a causal effect.



(a) GPT-5.6-sol. **(b)** DeepSeek-V4-flash (high). **(c)** DeepSeek-V4-flash (max).

Figure 5: Task-level overlap of Pass@1 success on the 91-task separable subset. (a) GPT-5.6-sol passes 29 tasks with scientific information and 33 tasks without it; 21 pass in both conditions, 8 only with information, and 12 only without it. (b) DeepSeek-V4-flash at high reasoning depth passes 21 tasks with scientific information and 15 tasks without it; 12 pass in both conditions, 9 only with information, and 3 only without it. (c) At max reasoning depth, DeepSeek-V4-flash passes 24 tasks with scientific information and 34 tasks without it; 20 pass in both conditions, 4 only with information, and 14 only without it.

Figure 5 shows how these aggregate changes arise from task-level transitions. GPT-5.6-sol solves eight tasks only with scientific information and twelve only without it, DeepSeek-V4-flash at high reasoning depth solves nine only with information and three only without it, and at max reasoning depth the asymmetry reverses, with four tasks solved only with information and fourteen only

without it. Inspection of the GPT-5.6-sol cases suggests that scientific information can supply semantic constraints absent from local code symptoms, such as limiting cases, coordinate consistency, independent observables, and authoritative interfaces. However, it can also induce anchoring, scope spillover, or premature reliance on a supplied explanation, underscoring the need for executable evidence and independent validation.

For the max-depth configuration, the failure-mechanism audit points to the same redirection. With scientific information, misguided-exploration or surface-repair failures decrease from 18 to 10 in the categories of Table 3, while incomplete repair coverage or system-integration failures increase from 18 to 35: the supplied material tends to redirect the agent toward the diagnosed subsystem without carrying the fix through the full module chain. These shifts must be read together with verifier outcomes, because among the 53 tasks that fail under both conditions, 27 change failure category between the two conditions and nine of those retain the same private mask and the same passed and failed tests, so a category change alone does not imply a change in capability.

Read together across the three configurations, the response to scientific information tracks each configuration's standalone capability: the configuration weakest without information, DeepSeek-V4-flash at high reasoning depth with 16.48% Pass@1, gains 6.60 points from the material, while GPT-5.6-sol and the max-depth configuration, both above 36% without information, lose 4.39 and 10.99 points respectively. The paired depths of the same model further indicate that deeper reasoning partially substitutes for supplied guidance: moving from high to max more than doubles Pass@1 without scientific information (16.48% to 37.36%) yet yields only 3.29 additional points when the information is present (23.08% to 26.37%). This pattern suggests that weaker models may benefit more from scientific auxiliary information, whereas deeper reasoning and stronger standalone capability may reduce its value when the agent can already recover the scientific contract through its own exploration.

7 Conclusion

We introduced **SWE-bench Science**, a repository-level benchmark for evaluating whether coding agents can repair scientific software while preserving its scientific contracts. The benchmark contains 119 tasks from 98 GitHub repositories across 20 scientific domains and organizes them into **Issue-driven**, **Expert-exploratory**, and **Engineering-integration** paradigms. Its Chain-of-Evidence Protocol combines separated public and private tests with metrics for repair progress, exact success, and regression preservation, making it possible to distinguish visible-test performance from complete private-test correctness.

Across nine coding-agent configurations, the strongest result is GPT-6 Astra's 50.42% Pass@1, while the same configuration attains a 100.00% public score. This gap, together with the four recurring scientific failure mechanisms identified in unsuccessful repairs, shows that repository-level scientific software engineering requires scientific abstraction, disciplined exploration, system-wide integration, and generalization beyond visible cases. Paired comparisons on the scientific-information subset show that additional scientific context does not have a uniform effect across models or reasoning settings. It tends to reduce repair performance for GPT-5.6-sol, while DeepSeek-V4-flash benefits at high reasoning depth but performs worse at max reasoning depth, despite substantially higher input cost. These findings indicate that supplying scientific information alone does not guarantee a better repair; its value depends on how effectively an agent connects that information to repository evidence and validates it through execution. SWE-bench Science provides a foundation for measuring this capability, while broader model coverage is needed to understand how scientific information affects agent performance on scientific software engineering tasks.

AI use statement

In this work, we used generative AI writing assistants for proofreading and polishing the English of the manuscript. We have not used generative AI tools for generating research ideas, hypotheses, or experimental designs, and no reported number was produced by a generative AI tool. We have reviewed and revised all AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

SWE-bench Science draws on publicly accessible scientific-software repositories, their issue and pull-request records, and scientific publications, including arXiv preprints. Copyright in third-party source code, documentation, and papers remains with the respective rightsholders. Our benchmark release will identify upstream sources and applicable licenses, and will redistribute third-party material only where the relevant terms permit; otherwise, it will provide provenance links and instructions for obtaining the material from its original source. We will review released task records for unnecessary personal or sensitive information. The release will keep evaluator-only tests and solution-linked artifacts separate from the materials visible to coding agents.

REPRODUCIBILITY STATEMENT

The main paper specifies the task schema, construction protocol, evaluation metrics, and model–harness configurations. Appendix A details the construction process and task paradigms; Appendix C defines the paired scientific-information conditions; Appendix D reports domain counts; and Appendix E lists task-level metadata for all 119 tasks, including upstream repositories, issue or pull-request references when available, knowledge scope, and scientific focus. We provide an anonymous repository at `https://anonymous.4open.science/r/swebench-science-D281` with task artifacts, environment setup instructions, evaluation scripts, and instructions for reproducing the reported metrics, subject to third-party license terms. The released materials will distinguish agent-visible inputs from evaluator-only materials.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. SUPER: Evaluating agents on setting up and executing tasks from research repositories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12622–12645, 2024.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jiao Chen, Jianhua Tang, Xiaotong Yang, and Zuohong Lv. SWE-Bench 5G: Benchmarking AI coding agents on telecom network engineering tasks, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. SWE-Bench Pro: Can AI agents solve long-horizon software engineering tasks?, 2025. URL https://arxiv.org/abs/2509.16941.

Titouan Duston, Shuo Xin, Yang Sun, Daoguang Zan, Aoyan Li, Shulin Xin, Kai Shen, Yixiao Chen, Qiming Sun, Ge Zhang, et al. Ainsteinbench: Benchmarking coding agents on scientific repositories, 2025.

Nicholas Edwards, Yukyung Lee, Yujun Audrey Mao, Yulu Qin, Sebastian Schuster, and Najoung Kim. RExBench: Can coding agents autonomously implement AI research extensions? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 16380–16417, 2026. doi: 10.18653/v1/2026.acl-long.745.

Tianyu Hua, Harper Hua, Violet Xiang, Benjamin Klieger, Sang Truong, Weixin Liang, Fan-Yun Sun, and Nick Haber. ResearchCodeBench: Benchmarking LLMs on implementing novel machine learning research code. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Wenqi Huang, Charley Lee, Leonard Tng, and Serena Ge. DeepSWE: Measuring frontier coding agents on original, long-horizon engineering tasks, 2026a. URL https://arxiv.org/abs/2607.07946.

Ziyang Huang, Yi Cao, Ali K. Shargh, Jing Luo, Ruidong Mei, Mohd Zaki, Zhan Liu, Wyatt Bunstine, William Jurayj, Somdatta Goswami, Tyrel McQueen, Michael Shields, Jaafar El-Awady, Paulette Clancy, Benjamin Van Durme, Nicholas Andrews, William Walden, and Daniel Khashabi. Can coding agents reproduce findings in computational materials science?, 2026b.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems*, 2024.

Xue Xia, Zheyuan Yang, Arman Cohan, and Yilun Zhao. MMSciCode: Real-world evaluation of multilingual multi-discipline scientific research coding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 33981–33999, 2026. doi: 10.18653/v1/2026.acl-long.1566.

Yijia Xiao, Runhui Wang, Luyang Kong, Davor Golac, and Wei Wang. CSR-bench: Benchmarking LLM agents in deployment of computer science research repositories. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 12705–12723, 2025.

Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, Juntong Ni, George Michalopoulos, Yue Zhang, Ziyang Zhang, Mian Zhang, Zhiyu Chen, and Xinya Du. LMR-BENCH: Evaluating LLM agents' ability on reproducing language modeling research. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6164–6186, 2025. doi: 10.18653/v1/2025.emnlp-main.314.

A Benchmark Construction Details

All raw tasks are first generated according to a unified and auditable four-stage construction process. The process is designed to control environmental interference and prevent information leakage.

1. **Source Sampling and Screening.**
   We collect real issues, pull requests (PRs), commit records, and related literature from candidate open-source scientific-computing repositories. We then apply predefined screening criteria to remove overly simple fixes, tasks with unstable environment dependencies, tasks with severe solution leakage, and projects that substantially overlap with existing samples.
2. **Snapshot Freezing and Reproduction.**
   We freeze the source-code snapshot immediately before the target defect or missing capability occurred, denoted by $\mathcal{S}_{\mathrm{bug}}$. The snapshot is executed in an isolated container to verify that the target anomaly or capability gap can be reproduced reliably. This step ensures that the observed failure is caused by the core algorithm or logical semantics rather than by environmental noise.
3. **Public-Material Abstraction and Information Isolation.**
   Based on the intended behavioral contract, we extract the scientific invariants and data-flow constraints evaluated by each task. We construct a public evaluation package containing the public repository, a coarse-grained description of the observed phenomenon, a reproduction script, and the necessary domain-background documentation. The public materials do not reveal the patch location or hidden assertions.

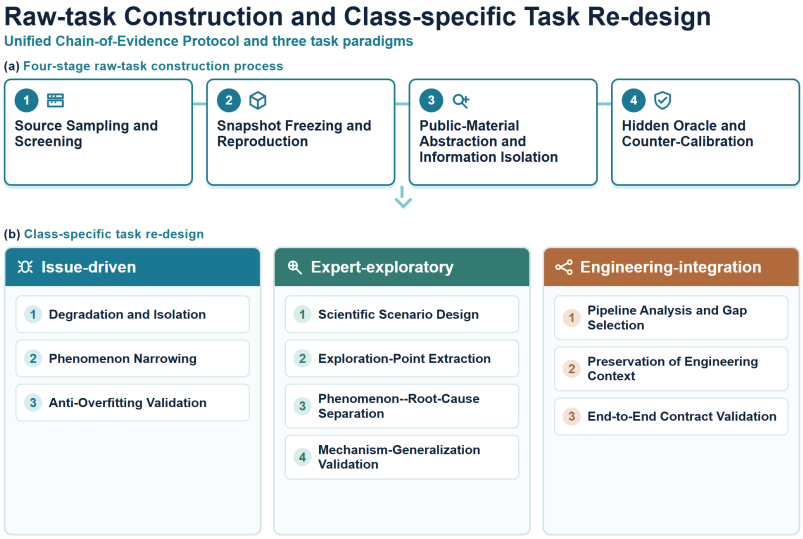


Figure 6: Combined overview of raw-task construction and class-specific task re-design.

4. **Hidden Oracle and Counter-Calibration.**
   We construct hidden validation suites based on semantic equivalence and boundary conditions. The validators examine not only the execution results under normal inputs, but also include reverse checks for hard-coded solutions, heuristic pseudo-fixes, and incomplete repairs.

Raw tasks are further re-designed according to their fit with the three task paradigms. Figure 6 summarizes the four-stage raw-task construction process and the class-specific task re-design for the three paradigms.

B Class-Specific Task Re-design Details

- **Issue-driven tasks** focus on repairing known defects;
- **Expert-exploratory tasks** focus on autonomously reasoning about unknown root causes in scientific scenarios;
- **Engineering-integration tasks** focus on understanding architecture across files and connecting a complete multi-module capability chain.

This task taxonomy avoids conflating different agent capabilities and provides a structured experimental basis for evaluating large language models in realistic scientific software engineering settings.

B.1 Issue-Driven Tasks

Definition and paradigm. Issue-driven tasks evaluate an agent's ability to repair real bugs under low-distortion conditions. These tasks take a confirmed historical defect as their direct definition starting point. Their central design principle is to narrow the observable phenomenon while isolating the patch information.

Construction process.

1. **Degradation and Isolation.**

We analyze the historical issue and PR discussion, roll the source code back to the state before the problematic commit, and construct a minimal reproducible example (MRE).

2. **Phenomenon Narrowing.**
   We reduce the upstream error to a coarse-grained and auditable phenomenon, such as inconsistent results across different computational paths or data loss under a particular format. The public materials remove any specific hints about the solution patch.
3. **Anti-Overfitting Validation.**
   Hidden validators cover different data scales, boundary parameters, and input-order transformations. These validators test whether the repaired code genuinely restores the scientific semantics rather than merely fitting the public script.

B.2 Expert-Exploratory Tasks

Definition and paradigm. Unlike Issue-driven tasks, which directly reproduce a known defect, Expert-exploratory tasks simulate a real scientific exploration process. Their definition starts from an exploration space containing a scientific scenario. Some tasks may be inspired by historical issues or PRs, but those issues or PRs do not serve as the definition starting point. These tasks evaluate how an agent uses domain knowledge to perform autonomous black-box or gray-box reasoning when the root cause of a complex phenomenon is unknown.

Construction process.

1. **Scientific Scenario Design.**
   We first define a high-level scientific use case, such as molecular representation consistency calibration, measurement-chain bias analysis, medical-image coordinate alignment, or experimental-workflow verification.
2. **Exploration-Point Extraction.**
   From theoretical materials, methodological literature, experimental phenomena, or discrepancies between results, we identify a core difference that is worth actively exploring. Upstream issues or PRs may provide optional background evidence, but they do not define the task.
3. **Phenomenon–Root-Cause Separation.**
   We construct a runnable public workflow that preserves the relevant domain background and observable anomaly, such as abnormal convergence or biased results, while hiding the precise location of the defective source code. The agent must identify the root cause through autonomous observation, controlled comparisons, and scientific reasoning.
4. **Mechanism-Generalization Validation.**
   Hidden validators change parameter scales, physical topologies, coordinate orderings, or boundary conditions. These tests determine whether the agent has genuinely inferred the scientific mechanism behind the observed phenomenon.

B.3 Engineering-Integration Tasks

Definition and paradigm. Engineering-integration tasks go beyond single-point bug repair. They focus on evaluating an agent's ability to understand system architecture in a realistic and complex codebase, connect different modules, and complete a full capability chain.

Construction process.

1. **Pipeline Analysis and Gap Selection.**
   We analyze the repository's end-to-end call chain, including data loading, parameter interpretation, intermediate-representation (IR) construction, operator assembly, and numerical solving. We then select a functionality gap with an appropriate scope.
2. **Preservation of Engineering Context.**

The public source code preserves the complete real package structure and neighboring modules. Public reproduction scripts show state comparisons at multiple stages and identify where the functionality chain fails to connect correctly. This design requires the agent to navigate across files and develop a system-level understanding.

3. **End-to-End Contract Validation.**
   Hidden validators include alternative execution paths, state-reset tests, and inter-module behavioral-contract tests. These validators ensure that the proposed repair achieves architecture-level integration across the entire module capability chain rather than merely making a single public test pass.

C Scientific Auxiliary Information Separation Details

The scientific-information factor measures the contribution of externally supplied domain knowledge, instantiated as scientific-rationale support. Of the 119 tasks, 91 allow this information to be separated from the supplied materials. For each eligible task, we compare paired conditions that differ only in auxiliary scientific information; the source-code snapshot, execution environment, public reproduction entry points, hidden validators, and required scientific context c_i^{req} remain fixed. Scientific auxiliary information denotes evidence relevant to scientific validity that is not directly available from source code or observable runtime behavior, including scientific rationales, upstream repairs, audit findings, paper excerpts, and expert guidance.

Withholding this information does not remove repository-intrinsic engineering cues, such as code structure, interfaces, traces, tests, and incomplete implementations, because doing so would change the software task itself. The ablation retains the executable context, task objective, observable symptoms, input data, and minimal interface documentation while removing scientific principles, equations or assumptions, expected properties, domain-specific diagnoses, and scientifically motivated repair strategies. It therefore estimates the marginal contribution of scientific auxiliary information given the same repository evidence, rather than performance in a clue-free setting. Results are analyzed in Section 6.2.

D Scientific Domain Coverage

Table 4: Distribution of the 119 SWE-bench Science tasks across 20 scientific domains.

Scientific domain	**Tasks**
Chemistry	24
Materials Science and Engineering	16
Biology	13
Biomedical Engineering	12
Physics	11
Mathematics	7
Astronomy	7
Atmospheric Science	5
Civil Engineering	5
Surveying and Mapping Science and Technology	3
Geophysics	3
Mechanics	3
Electrical Engineering	2
Marine Science	2
Geography	1
Aeronautical and Astronautical Science and Technology	1
Nuclear Science and Technology	1
Computer Science and Technology	1
Statistics	1
Information and Communication Engineering	1
Total	**119**

E TASK-LEVEL REPOSITORY AND KNOWLEDGE INVENTORY

This appendix provides the complete task-level annotation used to construct the benchmark. Each row links a task identifier to one of the 20 scientific domains used throughout the paper, together with its upstream repository, issue or pull request (when available), knowledge-domain scope, and scientific focus. The table contains all 119 tasks (IDs 001–119). The Boolean label 91-task subset indicates whether a task belongs to the paired scientific-information analysis in Section 6.2.

Table 5: Task-level scientific-domain, provenance, and scientific-knowledge inventory. A missing issue or pull-request reference indicates that no reference was provided in the source annotation.

Task 001 — Chemistry (91-task subset: False)

Upstream repository: `Auto-Mech/autochem`

Issue/PR: PR #637

Knowledge scope: TS linear segment cap selection, rotational symmetry number, reaction edge relabeling invariance, rotor observation

Task focus: The rotational symmetry of the TS graph must be determined by the real linear segment and reaction edge topology; changing the atom number or writing it in reverse cannot change the rotor/symmetry observation.

Task 002 — Chemistry (91-task subset: True)

Upstream repository: `aditisingh4812/simplified_tuned_range_separated`

Issue/PR: Not specified

Knowledge scope: range-separated tuning, local density structure, bohr-based grid scale and size consistency

Task focus: In range-separated tuning, the units of the density cutoff are not converted together with the grid representation, and the local thresholds are therefore misaligned.

Task 003 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/dpdata`

Issue/PR: Issue #973 / PR #983

Knowledge scope: stress/virial tag preservation, ASE compatibility, read and write consistency in extxyz/QUIP-GAP parsing

Task focus: extxyz parsing must convert stress into virial and be compatible with tensor components and field aliases, otherwise the stress semantics will be distorted.

Task 004 — Biology (91-task subset: True)

Upstream repository: `arq5x/bedtools2`

Issue/PR: Issue #750

Knowledge scope: split-read exon overlap filtering, interval intersection semantics, hit determination on splicing boundaries

Task focus: For split-read interval intersection, block overlap should be accumulated first and then fraction gate should be performed. Each block cannot be treated as an independent hit.

Task 005 — Biomedical Engineering (91-task subset: True)

Upstream repository: `schorschinho/osprey`

Issue/PR: Issue #750 / PR #751

Knowledge scope: edited MRS subspectrum polarity correction, generalization from two subspectra to four subspectra, phase-cycle control propagation

Task focus: Edited MRS requires polarity correction and phase-cycle control to be propagated through every subspectrum path; the fix cannot be confined to a single loader.

Task 006 — Biology (91-task subset: True)

Upstream repository: `arq5x/bedtools2`

Issue/PR: Issue #954 / PR #959

Knowledge scope: Query database tail count after EOF, same-chrom / later-chrom tails, Fisher right tail p-value consistency

Task focus: Fisher statistics require full database counts, and query EOF cannot truncate chromosome sweeps early.

Task 007 — Biomedical Engineering (91-task subset: True)

Upstream repository: `openmrslab/suspect`

Issue/PR: Issue #33 / PR #34

Knowledge scope: spectral registration initial value symbol, warm-start convention drift, frequency shift and phase shift consistency

Task focus: The drift sign of spectral registration must be consistent with the optimization residual agreement, otherwise the compensation direction will be reversed.

Task 008 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: first usable vacuum-plus-wall capability chain, outer-region operator payload, pseudo-vacuum metrics

Task focus: In the vacuum-wall link, the outer-region geometry and coefficients must be transferred together, and different wall shapes cannot be pressed into the same diagnosis.

Task 009 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: projection-wall geometry construction, external-region path, downstream reduced payload propagation

Task focus: The projection wall cannot degenerate into a scaled wall, and the boundary derivatives and geometric parameters must be entered into the outer-region calculation together.

Task 010 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: alternate perturbation-representation support, FOURIN, packed trig lookup, matrix assembly

Task focus: The PARFAC-driven alternate perturbation representation must not fall back to the baseline; parity, phase, and coefficient assembly must all follow the alternate representation.

Task 011 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: response-gated sideband capability, localized exchange propagation, energy diagnostics

Task focus: Sideband response must propagate through the support domain, coefficient assembly, and energy diagnostics; changing only a local cutoff is insufficient.

Task 012 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: 3-D cross-spectral coupling, coefficient assembly, energy channel, mode-family generalization

Task focus: 3-D mode coupling must retain the effective support of the difference/sum two branches, and cross-modal coupling cannot be collapsed into a unified weight.

Task 013 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: free / soft-fixed / hard-fixed boundary regimes, operators and energy closure ladder

Task focus: The boundary response should be closed to coefficients and energies respectively by interval and cannot be smoothed out by global averaging.

Task 014 — Physics (91-task subset: True)

Upstream repository: `PlasmaControl/DESC`

Issue/PR: Not specified

Knowledge scope: prescribed wall geometry, vacuum-plus-wall chain, outer-region operator payload, reduced response blocks

Task focus: The TERPSICHORE link must be closed end-to-end and cannot rely on a single fixture or plasma-only fallback.

Task 015 — Biomedical Engineering (91-task subset: True)

Upstream repository: `IlyaKuprov/Spinach`

Issue/PR: Issue #22

Knowledge scope: spin-dynamics thermalization, formalism-specific identity/projector construction, unit-state representation consistency

Task focus: thermalize should use formalism-specific identity/projector, and unit_state cannot be treated as two objects at the same time.

Task 016 — Biology (91-task subset: True)

Upstream repository: `macs3-project/MACS`

Issue/PR: Issue #735 / PR #736

Knowledge scope: HMMRATAC semantics, blacklist filtering, query EOF post-count consistency

Task focus: HMMRATAC refinement must first guard empty peak_content, and cannot close empty peaks on sparse bedGraph.

Task 017 — Astronomy (91-task subset: True)

Upstream repository: `gammapy/gammapy`

Issue/PR: Issue #6754 / PR #6757

Knowledge scope: J-factor line-of-sight impact parameter, b = D sin(theta), r = b cosh(t) stable integration

Task focus: The J-factor line integral must stabilize the branch according to the impact parameter, and cannot use tan(theta) to push directly to the singular point.

Task 018 — Biology (91-task subset: True)

Upstream repository: `macs3-project/MACS`

Issue/PR: Issue #680 / PR #689

Knowledge scope: blacklist filtering, HMMRATAC training/inference path, peak call statistical consistency

Task focus: The blacklist should be filtered before training and segmentation, and blacklist segments cannot be allowed to enter the model and counting of HMMRATAC.

Task 019 — Biology (91-task subset: True)

Upstream repository: `aertslab/pycisTopic`

Issue/PR: PR #196

Knowledge scope: Discretization of topic probability matrix to region / cell set, legacy object API, Regions_in_binarized_topic maintenance

Task focus: Topic binarization needs to support both matrix/object entries and stable threshold semantics, and cannot just cover the old object API.

Task 020 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/deepmd-kit`

Issue/PR: PR #5581

Knowledge scope: neighbor graph, periodic-cell mapping, segment reductions, force / atom virial scatter

Task focus: The neighbor graph needs to unify PBC mapping, empty segment reduction and force/virial scatter, and guard edges cannot be regarded as effective contributions.

Task 021 — Chemistry (91-task subset: True)

Upstream repository: `deepmodeling/dpti`

Issue/PR: PR #103

Knowledge scope: coupling endpoints, pair-specific epsilon, three-stage HTI/FEP input generation, thermo_style consumption

Task focus: Soft-core LJ FEP must generate a real `compute fep`; energy divided by lambda cannot substitute for derivative observations.

Task 022 — Biomedical Engineering (91-task subset: True)

Upstream repository: `nilearn/nilearn`

Issue/PR: Issue #5168 / PR #5169

Knowledge scope: Volume-to-surface atlas projection, discrete region identity preservation under nearest sampling, missing samples should return NaN

Task focus: Discrete brain area projections should adopt mode sampling, and NaN should be retained for missing samples. Labels cannot be averaged numerically.

Task 023 — Chemistry (91-task subset: True)

Upstream repository: `MDAnalysis/mdanalysis`

Issue/PR: Issue #4257 / PR #4258

Knowledge scope: NPT trajectory without jump expansion, box dimensions update, first frame displacement history initialization

Task focus: NoJump's frame bookkeeping must preserve the first cross-boundary displacement and cannot exit early at frame 0.

Task 024 — Statistics (91-task subset: True)

Upstream repository: `pymc-devs/pymc`

Issue/PR: Issue #6871 / PR #6872

Knowledge scope: ZeroSumNormal's log-density, normalization and pseudo-determinants on zero-sum manifolds

Task focus: The logp of ZeroSumNormal must be calculated on the zero-sum manifold and cannot directly reuse ambient Gaussian normalization.

Task 025 — Biomedical Engineering (91-task subset: True)

Upstream repository: `schorschinho/osprey`

Issue/PR: Issue #447 / PR #448; Issue #657 / PR #658

Knowledge scope: acquisition metadata consistency in the edited-MRS workflow, consistent calibration reports and representations

Task focus: Edited-MRS acquisition metadata and the water-calibration time scale must remain consistent; TE/TR must not be converted a second time.

Task 026 — Atmospheric Science (91-task subset: True)

Upstream repository: `Unidata/MetPy`

Issue/PR: Issue #3315 / PR #3324

Knowledge scope: Isentropic interpolation on sigma coordinates, pressure column selection, dimension semantic correction

Task focus: Isentropic interpolation must respect vertical coordinate semantics such as sigma/eta, and cannot treat them directly as pressure columns.

Task 027 — Chemistry (91-task subset: True)

Upstream repository: `little1d/NMRTrans`

Issue/PR: Issue #4

Knowledge scope: Masking and representation semantics of NMR peak-set encoding, batch padding, and set transformer

Task focus: The mask semantics of the peak encoder are reversed, the padding is retained, the real peaks are masked, and the set representation is distorted.

Task 028 — Biology (91-task subset: True)

Upstream repository: `JialongJiang/DSPIN`

Issue/PR: Not specified

Knowledge scope: control-referenced relative responses, perturbation response normalization and fallback path integrity

Task focus: The control sample index and relative response baseline cannot be misaligned, otherwise the disturbance response will shift as a whole.

Task 029 — Biology (91-task subset: True)

Upstream repository: `biopython/biopython`

Issue/PR: Issue #4611 / PR #4637

Knowledge scope: Feature coordinates and sequence semantics after circular record orientation normalization and reverse-complement

Task focus: After circular feature reverse-complement, the part order must be reversed, and the original cross-origin order cannot be retained.

Task 030 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `materialsproject/pymatgen`

Issue/PR: Issue #3677 / PR #3779

Knowledge scope: SHELX/AIRSS.res export, SFAC ordered species column, stable round-trip serialization

Task focus: The SFAC of.res is an ordered list, and species cannot be written as an unordered set.

Task 031 — Biomedical Engineering (91-task subset: True)

Upstream repository: `mne-tools/mne-python`

Issue/PR: Issue #13913 / PR #13920

Knowledge scope: average reference across multiple sensor types, classification type re-reference and joint / projection path consistency

Task focus: The average reference of mixed sensor types should be done separately according to channel type and cannot be directly combined into a union mean.

Task 032 — Chemistry (91-task subset: True)

Upstream repository: `openmm/openmm`

Issue/PR: Issue #5025 / PR #5026

Knowledge scope: GROMACS topology import, NBFIX / combination rule semantics, total energy representation maintenance

Task focus: GROMACS topology import needs to unify the translation semantics of NBFIX, combination rule and 1-4 exception.

Task 033 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `pyscf/pyscf`

Issue/PR: PR #2946

Knowledge scope: consistency between k-point and Gamma-supercell representations in periodic HF, Coulomb wrap-around semantics

Task focus: Periodic HF k-point and Gamma-supercell paths must implement equivalent Coulomb semantics; wrap-around handling must not diverge between them.

Task 034 — Mathematics (91-task subset: True)

Upstream repository: `patrick-kidger/diffrax`

Issue/PR: Issue #598 / PR #746

Knowledge scope: strong order, reverse-oriented SDE integration, boundary conditions and hidden state processing

Task focus: The step size and Levy area parity of the reverse time SDE must be processed according to the direction, and the forward SRK convention cannot be followed.

Task 035 — Mathematics (91-task subset: True)

Upstream repository: `scipy/scipy`

Issue/PR: Issue #14480 / PR #14552

Knowledge scope: Numerical consistency of LSODA event processing, dense output, stiff kinetics

Task focus: The dense output of LSODA needs to take the history of the last successful step, and cannot use the next candidate state for event positioning.

Task 036 — Marine Science (91-task subset: True)

Upstream repository: `Parcels-code/Parcels`

Issue/PR: Issue #2521 / PR #2598

Knowledge scope: Problems with spherical curvilinear grids, latitude and longitude grid index, and downstream parameters still follow the old path

Task focus: Curved grid point-in-cell must be calculated according to spherical geometry and cannot continue to be approximated by longitude and latitude planes.

Task 037 — Biomedical Engineering (91-task subset: True)

Upstream repository: `nipy/nibabel`

Issue/PR: PR #1340

Knowledge scope: Enhanced multiframe DICOM reconstruction, frame/direction semantics and image plane consistency

Task focus: The frame order of multi-frame DICOM must be consistent with the physical slice order, and cannot be reconstructed solely by index.

Task 038 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `spglib/spglib`

Issue/PR: Issue #584 / PR #589

Knowledge scope: Classification stability, primitive/conventional transformation and symmetry search under equivalent lattice representation

Task focus: The operation route of symmetry search must be updated synchronously with the base transformation and cannot be disconnected from the structure route.

Task 039 — Geography (91-task subset: True)

Upstream repository: `landlab/landlab`

Issue/PR: Issue #2408 / PR #2409

Knowledge scope: parcel transport, reach-scale flux consistency, segmentation and external representation consistency

Task focus: The speed of parcel transport must be consistent with reach-level flux conversion, and the number of particles cannot be allowed to change the physical transport rate.

Task 040 — Geophysics (91-task subset: True)

Upstream repository: `simpeg/simpeg`

Issue/PR: Issue #1099 / PR #1575

Knowledge scope: magnetic-dipole forward modeling, source representation equivalence, numerical accuracy and theoretical fit

Task focus: Different formulations of magnetic dipole forward modeling must fall into the same discrete field space, and the representation method cannot change the results.

Task 041 — Geophysics (91-task subset: True)

Upstream repository: `sunpy/sunpy`

Issue/PR: Issue #8406 / PR #8415

Knowledge scope: Earth-centered coordinate transformations, GEI/GSE path consistency and vector conservation

Task focus: The GEI/GSE transformation must follow a purely rotational path and cannot go through an intermediate map with additional corrections.

Task 042 — Astronomy (91-task subset: True)

Upstream repository: `sunpy/sunpy`

Issue/PR: Issue #8490 / PR #8548

Knowledge scope: Sun coordinate objects maintain frame, observer, obstime, representation, units and numerical meanings in table serialization

Task focus: The solar coordinate table round-trip must retain frame, observer and obstime, and cannot compress the coordinate object into a pure numerical value.

Task 043 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `PMEAL/porespy`

Issue/PR: Issue #966 / PR #1174

Knowledge scope: Two-point correlation function, aperiodic boundaries, overlap counting and radial binning on finite images

Task focus: Finite image two-point correlation needs to perform overlap and radial bin statistics according to finite-image semantics, and cannot be treated as a periodic domain.

Task 044 — Chemistry (91-task subset: True)

Upstream repository: `pybamm-team/PyBaMM`

Issue/PR: PR #5600

Knowledge scope: In regularized power operations, distinguish between state-independent fitting coefficients and dynamic state-dependent expressions

Task focus: The fractional-power stabilization in the exchange current expression must distinguish between constants and state variables, and the parameters themselves cannot be rewritten.

Task 045 — Marine Science (91-task subset: True)

Upstream repository: `xgcm/xgcm`

Issue/PR: Issue #712 / PR #713

Knowledge scope: Connection surface direction, axis order, consistency of pad and transpose

Task focus: cubed-sphere halo/padding must first align the dimensions and surface directions of the source slice, and cannot read the adjacent surfaces incorrectly after transpose.

Task 046 — Astronomy (91-task subset: True)

Upstream repository: `StingraySoftware/stingray`

Issue/PR: PR #965

Knowledge scope: Coherence/power spectrum calculation under effective domain, boundary state and invalid value determination

Task focus: The coherence estimate must correctly deduct the noise deviation and count as averaged spectra, and cannot return a finite value in the noise-dominated frequency band.

Task 047 — Information and Communication Engineering (91-task subset: True)

Upstream repository: `scikit-rf/scikit-rf`

Issue/PR: Issue #654 / PR #655

Knowledge scope: Conversion and parameter consistency of different microwave network representations under complex reference impedance

Task focus: Network connections under complex reference impedance must first undergo wave-definition transformation, and the traveling-wave convention cannot be used.

Task 048 — Chemistry (91-task subset: True)

Upstream repository: `pycalphad/pycalphad`

Issue/PR: Issue #310 / PR #311

Knowledge scope: consistency between phase energy, Gibbs energy, and ordered-phase representation

Task focus: Gibbs-energy bookkeeping for ordered/disordered phases must project onto the non-ordering sublattice; database encoding must not change phase stability.

Task 049 — Mathematics (91-task subset: True)

Upstream repository: `zwicker-group/py-pde`

Issue/PR: PR #839

Knowledge scope: Multi-step physical process semantics and incombinability in the evolution of discrete random fields

Task focus: The noise terms and volume factors of discrete random fields must be processed hierarchically, and cell-volume scaling cannot be placed in the wrong location.

Task 050 — Physics (91-task subset: True)

Upstream repository: `jcmgray/quimb`

Issue/PR: Issue #402

Knowledge scope: Scale-dependent norm reconstruction, result consistency of different tensor representations

Task focus: The norm of the tensor network must remain unchanged after gauging/rescaling, and the approximate contraction cannot change the exact norm.

Task 051 — Geophysics (91-task subset: True)

Upstream repository: `SHTOOLS/SHTOOLS`

Issue/PR: PR #397

Knowledge scope: Spherical harmonic gradient, polar boundary and planetary magnetic field calculation

Task focus: The gravity/magnetic force second-order tensor must be consistent from the local frame to the Cartesian frame, and recursion and boundary constraints cannot be allowed to destroy the rotation.

Task 052 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `usnistgov/fipy`

Issue/PR: Issue #1190 / PR #1192

Knowledge scope: Modular gradients and physical geometry consistency on non-uniform orthogonal grids

Task focus: Non-uniform Grid3D needs to unify cell widths, face sizes and modular gradient semantics, and cannot only retain a single spacing.

Task 053 — Biomedical Engineering (91-task subset: True)

Upstream repository: `Project-MONAI/MONAI`

Issue/PR: Issue #4103 / PR #6681

Knowledge scope: Physical spacing, subvoxel surface area approximation and surface Dice

Task focus: Surface Dice should be calculated based on physical surface area, and anisotropic voxels cannot simply be regarded as equal-area boundaries.

Task 054 — Chemistry (91-task subset: True)

Upstream repository: `matchms/matchms`

Issue/PR: PR #863

Knowledge scope: Screening, merging and entropy-based spectral similarity of highly similar peaks

Task focus: Entropy similarity cannot allow overlapping candidate windows to count the same peak repeatedly, and the pair/matrix paths must share the same peak semantics.

Task 055 — Astronomy (91-task subset: True)

Upstream repository: `astropy/specutils`

Issue/PR: Issues #817 / #1059 / #1300; PRs #1060 / #1305

Knowledge scope: Boundary extension, partially overlapping bins and finite interval filling

Task focus: Spectral resampling must propagate flux and variance according to limited bin overlap, and bins cannot be used as point samples.

Task 056 — Aeronautical and Astronautical Science and Technology (91-task subset: True)

Upstream repository: `poliastro/poliastro`

Issue/PR: Issue #1563; PR #1583

Knowledge scope: Consistency between orbital state and binding epoch, history sampling semantics

Task focus: Each state and epoch of Orbit→Ephem must point to the same orbital state, and the anomaly cannot drift with the timeline.

Task 057 — Chemistry (91-task subset: True)

Upstream repository: `mdtraj/mdtraj`

Issue/PR: Issue #1925; PR #1929

Knowledge scope: Molecular reconstruction under periodic boundaries, bonded geometry and coordinate order independence

Task focus: make_molecules_whole restores the periodic image along the bond graph in one pass, and cannot let the bond order determine the translation results.

Task 058 — Nuclear Science and Technology (91-task subset: True)

Upstream repository: `openmc-dev/openmc`

Issue/PR: Issue #3852; PR #3853

Knowledge scope: Rectangular lattice distance, boundary and geometric determination in particle transport

Task focus: The boundary distance and transport event of the repeated rectangular lattice must be calculated correctly, and the legal geometry cannot be misjudged as terminated.

Task 059 — Computer Science and Technology (91-task subset: True)

Upstream repository: `scikit-image/scikit-image`

Issue/PR: Issue #5112; PR #6296

Knowledge scope: voxel spacing, area and morphometric calibration of regionprops

Task focus: regionprops should treat spacing as a physical scale and cannot treat anisotropic sampling as a lattice index.

Task 060 — Biomedical Engineering (91-task subset: True)

Upstream repository: `SpikeInterface/spikeinterface`

Issue/PR: Issue #3328; PR #3517

Knowledge scope: non-equidistant sampling, cross-band motion correction, and interpolation correction

Task focus: Motion interpolation must align AP/LF streams on the physical time axis; assuming uniform motion from the sampling rate is invalid.

Task 061 — Astronomy (91-task subset: True)

Upstream repository: `astropy/reproject`

Issue/PR: Issue #199; PR #589

Knowledge scope: Numerical elimination, reprojection and energy conservation of small sphere area

Task focus: The extremely small sky pixel in high-resolution reproject will have numerical cancellation, and overlap geometry and flux conservation must be stable.

Task 062 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `materialsproject/atomate2`

Issue/PR: Issue #333 / PR #415

Knowledge scope: Green-Lagrange strain-stress fitting, independent shear response reconstruction, and tensor symmetry in a high-throughput crystal elasticity workflow

Task focus: Elastic tensor reconstruction must bring the complete shear response back to the final result, and the reduce/reconstruct link cannot lose symmetry information.

Task 063 — Biology (91-task subset: True)

Upstream repository: `tskit-dev/tskit`

Issue/PR: Issue #3053, #3175 / PR #3059, #3176

Knowledge scope: Distinguishing between ”no lineage” and ”multi-rooted trees with no common ancestors” under window normalization, and normalizing the observation span

Task focus: Window normalized pair-coalescence needs to distinguish between non-lineage trees and multi-rooted trees without common ancestors, and the boundary semantics of the two types cannot be confused.

Task 064 — Chemistry (91-task subset: True)

Upstream repository: `cclib/cclib`

Issue/PR: Issue #660 / PR #745

Knowledge scope: Signed energy extraction, unit consistency and ccData exposure for DFT-D dispersion correction in ORCA output

Task focus: ORCA's DFT-D dispersion correction should be written out as an independent signed physical quantity and not just left inside the parser.

Task 065 — Astronomy (91-task subset: True)

Upstream repository: `yt-project/yt`

Issue/PR: Issue #4788 / PR #4939

Knowledge scope: SPH-based three-dimensional projection, oblique slicing, real sight integral and kernel function geometric semantics

Task focus: SPH line-of-sight integration and oblique slicing need to share consistent kernel geometry and normalization, and the resolution cannot be allowed to change the observation results.

Task 066 — Atmospheric Science (91-task subset: True)

Upstream repository: `SciTools/iris`

Issue/PR: Issue #4830 / PR #5548

Knowledge scope: Equivalent identification of Rotated Mercator and Oblique Mercator under CF grid mapping, projection object and site matching semantics

Task focus: The CF grid mapping of rotated Mercator and oblique Mercator should be recognized as equivalent and cannot be linked to the wrong projection reference.

Task 067 — Atmospheric Science (91-task subset: True)

Upstream repository: `Ouranosinc/xclim`

Issue/PR: Issue #700 / PR #704

Knowledge scope: Overwintering precipitation and Drought Code state transfer, recursive boundary between active season and dormant phase when seasons switch

Task focus: The seasonal state machine of the fire danger index should separate the winter precipitation and active-day updates, and the state transfer cannot be connected to the wrong stage.

Task 068 — Biology (91-task subset: True)

Upstream repository: `samtools/htslib`

Issue/PR: Issue #1820 / PR #1897

Knowledge scope: Actual coverage interval, rlen, END / symbolic allele / reference block semantics recorded on the reference genome

Task focus: The reference occupancy interval of VCF/BCF should be consistent on paths such as symbolic allele, END and SVLEN.

Task 069 — Biology (91-task subset: True)

Upstream repository: `open2c/cooltools`

Issue/PR: Issue #456 / PR #501

Knowledge scope: Contact-frequency and smoothed P(s) propagation, normalization and multi-region aggregation consistency

Task focus: expected cis smoothing needs to propagate the numerators and denominators of raw contact and weighted contact separately, and observation apertures cannot be mixed.

Task 070 — Biology (91-task subset: True)

Upstream repository: `levitsky/pyteomics`

Issue/PR: Issue #202 / PR #203

Knowledge scope: ProForma local modification and global ion-carrying charge calculation, charge_state consistency

Task focus: ProForma's global loads, local modifications, and neutral masses/m/z require unified charge accounting, and electrons cannot be counted repeatedly.

Task 071 — Physics (91-task subset: True)

Upstream repository: `silx-kit/pyFAI`

Issue/PR: Issue #2773 / PR #2772

Knowledge scope: Transformation from detector direction to sample coordinate system q vector, exit-angle and reciprocal represent consistency

Task focus: The q vector transformation must maintain the consistency of the detector to sample coordinates, and the orientation encoding cannot change the reciprocal space results.

Task 072 — Biomedical Engineering (91-task subset: True)

Upstream repository: `pydicom/pydicom`

Issue/PR: Issue #2210 / PR #2212

Knowledge scope: Bit depth, sampling, limited-range luma/chroma conversion of YBR_FULL / YBR_FULL_422 / YBR_PARTIAL_420

Task focus: The YBR_* pixel format must unify bit depth and sampling extension semantics, and cannot just modify the limited-range path.

Task 073 — Biology (91-task subset: True)

Upstream repository: `biotite-dev/biotite`

Issue/PR: Issue #777 / PR #781

Knowledge scope: Conformational connectivity, bond map preservation and bond mapping after alternate location selection

Task focus: The alternate location filtering must be completed before the bond graph, and the selected conformation cannot be disconnected from the struct_conn.

Task 074 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/DeePTB`

Issue/PR: Issue #332 / PR #320

Knowledge scope: Generalized eigenvalue problems, ill-conditioned overlap matrix processing, finite band data and validity masks

Task focus: The generalized eigenvalue problem of non-orthogonal tight binding must first deal with the ill-posed overlap matrix, and cannot directly Cholesky on the invalid subspace.

Task 075 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/deepmd-kit`

Issue/PR: Issue #5593 / PR #5601

Knowledge scope: Boundary interpolation, first-order sensitivity, layout and derivative consistency of DeePMD compressed tables

Task focus: Compressed table extrapolation must maintain both value and derivative continuation semantics, and C1 continuation cannot be truncated outside the boundary.

Task 076 — Mathematics (91-task subset: True)

Upstream repository: `cvxpy/cvxpy`

Issue/PR: Issue #3092 / PR #3093

Knowledge scope: Coordinate mapping of N-dimensional parameters to sparse coefficient matrix after indexing, broadcasting, and matrix multiplication

Task focus: Canonicalization of the Parameter tensor requires first folding repeated positions and then mapping to a sparse coefficient matrix.

Task 077 — Mathematics (91-task subset: True)

Upstream repository: `shapely/shapely`

Issue/PR: Issue #1670, #1929 / PR #1708, #1939

Knowledge scope: Numerical stability fallback, candidate edge enumeration, NumPy ufunc batching and output buffering compatibility

Task focus: Oriented envelope must truly select the smallest area among candidate edges and is compatible with NumPy batch processing.

Task 078 — Astronomy (91-task subset: True)

Upstream repository: `astropy/photutils`

Issue/PR: Issue #2287 / PR #2288

Knowledge scope: Bidirectional transformation, direction consistency and round-trip constraints of aperture geometry under WCS

Task focus: The conversion from sky aperture to pixel aperture must keep the WCS local base direction consistent, and the aperture mirror cannot be shifted.

Task 079 — Chemistry (91-task subset: True)

Upstream repository: `aspuru-guzik-group/selfies`

Issue/PR: Issue #120 / PR #125

Knowledge scope: kekulization of SMILES aromatic system to SELFIES, charged aromatic carbon electron counting

Task focus: When converting the SMILES aromatic system to SELFIES, the electron count of the charged aromatic carbon must also be dealt with, and the neutral zero-hydrogen path cannot only be repaired.

Task 080 — Surveying and Mapping Science and Technology (91-task subset: True)

Upstream repository: `xarray-contrib/xarray-spatial`

Issue/PR: Issue #1086 / PR #1087

Knowledge scope: footprint, coordinate direction normalization, nodata propagation and cubic boundary interpolation

Task focus: The footprint, coordinate direction and nodata propagation during reprojection must be bound to the interpolation order, and the boundaries cannot be allowed to degenerate into copied values.

Task 081 — Biology (91-task subset: True)

Upstream repository: `scverse/anndata`

Issue/PR: Issue #1854 / PR #2122

Knowledge scope: backed HDF5/Zarr re-indexing, outer join, and sparse/dense annotation consistency

Task focus: Disk-backed concatenation must write non-concatenated-axis annotations back under the new index; preserving only the main matrix is insufficient.

Task 082 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/dpgen`

Issue/PR: Issue #1639 / PR #1801

Knowledge scope: DP-GEN active-learning input generation, LAMMPS template revise, model-deviation control items, relative error and electron temperature propagation

Task focus: DP-GEN must propagate the same set of LAMMPS model-deviation settings in both direct generation and template revision paths, and the relative, cluster or electron temperature items cannot be lost in the template path.

Task 083 — Chemistry (91-task subset: False)

Upstream repository: `deepmodeling/reacnetgenerator`

Issue/PR: Issue #2165 / PR #2221

Knowledge scope: bond-order trajectory, atom-bond graph equivalence, isomer merge mode, cyclic structure species naming and reaction network report

Task focus: The species report of ReacNetGenerator must define the same molecule according to chemical graph equivalence and merge mode, and cannot separate bond order changes or traversal order into different species.

Task 084 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `deepmodeling/dpdata`

Issue/PR: Issue #951 / PR #957

Knowledge scope: ABACUS relaxation-output parsing, frame-wise coordinates and cell labels, archive/reload round-trip, complete trajectory recovery

Task focus: The ABACUS relaxation round-trip must preserve the complete relaxation history and frame-wise labels, rather than retaining only the final structure or losing cell/coordinate semantics.

Task 085 — Chemistry (91-task subset: False)

Upstream repository: `dftbplus/dftbplus`

Issue/PR: Issue #1069 / PR #1462

Knowledge scope: DFTB+ COSMO/GB archived observations, atomic order invariance, equivalent atomic records and solvent surface adjacencies

Task focus: DFTB+ implicit solvent observations should be insensitive to equivalent atoms and input order, and should not allow the order of atom records to change COSMO-like reporting.

Task 086 — Materials Science and Engineering (91-task subset: True)

Upstream repository: `materialyzeai/matgl`

Issue/PR: Issue #642 / PR #819

Knowledge scope: MatGL dataset element list, pre-trained model element vocabulary, graph construction atom type encoding, fine-tuning compatibility check

Task focus: MatGL fine-tuning must first verify that the data set element list is consistent with the model element vocabulary, and the graph construction cannot be allowed to silently map atoms to wrong embeddings.

Task 087 — Biomedical Engineering (91-task subset: False)

Upstream repository: `dipy/dipy`

Issue/PR: Issue #3968 / PR #3976

Knowledge scope: Patch2Self small batch denoising, trailing-volume flush, b0 volume data processing, patch radius and baseline shift

Task focus: Patch2Self wants to keep the diffuse volume consistent with b0 semantics on mini-batch and tail volume data, and cannot leave finite but offset denoised observations.

Task 088 — Atmospheric Science (91-task subset: False)

Upstream repository: `pydata/xarray`

Issue/PR: Issue #9108 / PR #9116

Knowledge scope: CFTimeIndex resample, calendar and has_year_zero reserved, year-zero near month/year-end boundary aggregation

Task focus: CF time aggregation must retain calendar and has_year_zero semantics, and cannot aggregate legal short sequences into failure paths at the year-zero boundary.

Task 089 — Chemistry (91-task subset: False)

Upstream repository: `cclib/cclib`

Issue/PR: Issue #46 / PR #1085

Knowledge scope: ORCA excited-state absorption spectrum, oscillator strength, multiple transition attribute table, etoscs representation selection

Task focus: cclib must retain the correct oscillator-strength representation when parsing ORCA excited state spectra, and must not allow subsequent or substitute spectra to overwrite semantic values.

Task 090 — Chemistry (91-task subset: True)

Upstream repository: `jjhelmus/nmrglue`

Issue/PR: Issue #187 / PR #189

Knowledge scope: Bruker NUS compressed FID stream, nuslist sparse coordinates, quadrature/acquisition order, missing FID zero-fill

Task focus: nmrglue needs to expand the sparse FID stream into a multi-dimensional NMR grid according to the nuslist and collection order, and cannot treat the compressed stream as an ordinary dense 2D sequence.

Task 091 — Materials Science and Engineering (91-task subset: False)

Upstream repository: `materialsproject/pymatgen-core`

Issue/PR: Issue #4607 / PR #87

Knowledge scope: Gaussian cube nonzero origin, Bohr-to-angstrom unit conversion, voxel origin and structural coordinate placement

Task focus: When reading Gaussian cube, the non-zero volume data origin and unit conversion must be included in the structure placement. The first voxel cannot be the default coordinate origin.

Task 092 — Chemistry (91-task subset: False)

Upstream repository: `chemprop/chemprop`

Issue/PR: Issue #426 / PR #427

Knowledge scope: Condensed Graph of Reaction, atom mapping, reaction-mode options, and reactant/product feature propagation

Task focus: Chemprop reaction featurization must construct a CGR according to the reaction mode and atom map; mapped atoms must be retained and reactant/product changes must not be encoded in reverse.

Task 093 — Biomedical Engineering (91-task subset: False)

Upstream repository: `Project-MONAI/MONAI`

Issue/PR: Issue #6985 / PR #7000

Knowledge scope: ITKReader, MetaTensor, ITK image round-trip, affine/spacing/origin/direction coordinate convention

Task focus: MONAI's reader → MetaTensor → ITK round-trip must preserve patient-space coordinates and orientation, image geometry cannot be silently changed between equivalent paths.

Task 094 — Physics (91-task subset: False)

Upstream repository: `materialsproject/pymatgen`

Issue/PR: PR #1706

Knowledge scope: magnetic space group operations, BNS/Jones-Faithful setting transform, fractional translation, time reversal

Task focus: Magnetic space group archive records must be converted according to setting to preserve translation and time-reversal semantics, and the old coordinate convention cannot be read directly as the current convention.

Task 095 — Materials Science and Engineering (91-task subset: False)

Upstream repository: `PMEAL/OpenPNM`

Issue/PR: Issue #2843 / PR #2849

Knowledge scope: throat conductance field, model builder, operator assembly, advection-diffusion solver, repeated stage transfer

Task focus: Porous network transmission requires consistent transmission of conductance representation from model output to solver and flux calculations. Object fields and array fields cannot be mixed in repeated solves.

Task 096 — Atmospheric Science (91-task subset: False)

Upstream repository: `CLIMADA-project/climada_python`

Issue/PR: Issue #831 / PR #839

Knowledge scope: antimeridian longitude normalization, track bounds, centroid filtering, sparse hazard matrix, equivalent date-line encoding

Task focus: Tropical cyclone hazard calculations need to unify longitude wrapper semantics in paths, centroid ranges, and sparse maps, and cannot miss or misplace wind fields near ± 180 degrees.

Task 097 — Physics (91-task subset: True)

Upstream repository: `deepmodeling/deepmd-kit`

Issue/PR: Issue #5663 / PR #5854

Knowledge scope: Spin virtual types, negative placeholder atoms, dense/lower layouts, dp-model/PT parity, padded mixed-size batches

Task focus: DeePMD spin model preprocessing requires negative type padding to remain inert in real/virtual type expansion and backend layout, and cannot change placeholder atoms into valid spin types.

Task 098 — Surveying and Mapping Science and Technology (91-task subset: True)

Upstream repository: `pytroll/satpy`

Issue/PR: Issue #3182 / PR #3217

Knowledge scope: geostationary satellite position units, TLE fallback, cloud-top-height parallax, observer-look boundary, inverse resampling

Task focus: Satpy parallax correction needs to unify internal satellite height/radius units, only convert at external library boundaries, and cannot mix meters/km between reader metadata, TLE and Cartesian geometry.

Task 099 — Surveying and Mapping Science and Technology (91-task subset: True)

Upstream repository: `Stanford-NavLab/gnss_lib_py`

Issue/PR: PR #121

Knowledge scope: WLS pseudorange positioning, Earth-rotation/Sagnac correction, satellite ECEF convention, simulator direct satellite states

Task focus: GNSS WLS point positioning requires rotating the satellite position to the receive measurement frame and being compatible with simulator input. Transmitted-frame ECEF cannot be directly solved as receive-frame.

Task 100 — Chemistry (91-task subset: True)

Upstream repository: `DCoupry/autografs`

Issue/PR: Issue #174/#197/#198/#207 / PR #208

Knowledge scope: AuToGraFS relaxed embedding, topology slot motion, closure regression, cell volume scaling, pinned-net control

Task focus: Relaxed framework embedding must allow low symmetry slot movement while maintaining topology, closure and volume scale, the objective cannot be adjusted to allow individual nets to pass.

Task 101 — Chemistry (91-task subset: True)

Upstream repository: `pybamm-team/PyBaMM`

Issue/PR: Issue #5414 / PR #5540

Knowledge scope: half-cell Doyle-Fuller-Newman model, surface-form electrolyte potential, boundary-condition registration, thermal Ohmic-heating coupling

Task focus: PyBaMM half-cell DFN must consistently register electrolyte/solid phase potential boundary conditions for surface-form and thermal sub-models, and cannot allow half-cell domain splitting to cause gradient shape mismatch.

Task 102 — Materials Science and Engineering (91-task subset: False)

Upstream repository: `marrink-lab/polyply_1.0`

Issue/PR: Issue #380 / PR #381

Knowledge scope: polymer residue graph, template/volume identity, coordinate generation, library loading, residue equivalence

Task focus: Polymer workflows must pass template, volume, and residue graph identities all the way to coordinate generation without losing graph structure semantics between loading and generation.

Task 103 — Mechanics (91-task subset: False)

Upstream repository: `SteveDoyle2/pyNastran`

Issue/PR: Issue #584 / PR #85

Knowledge scope: symmetric PCOMP expansion, ABD equivalent, D block calculation, explicit layup and symmetric layup consistency

Task focus: The ABD of the symmetrical layup must be consistent with the explicit layup, and the expanded directional stiffness and D matrix cannot deviate.

Task 104 — Chemistry (91-task subset: False)

Upstream repository: `pybamm-team/PyBaMM`

Issue/PR: Issue #3526 / PR #3979

Knowledge scope: upwind/downwind edge states, cell-centered and edge-centered flux, Dirichlet FV boundary handling, mesh-size invariance

Task focus: The three-layer semantics of cell/edge/boundary in a finite volume must be consistent, and upwind/downwind cannot be mismatched in boundary processing.

Task 105 — Chemistry (91-task subset: False)

Upstream repository: `rdkit/rdkit`

Issue/PR: Issue #9338 / PR #9339

Knowledge scope: tautomer canonicalization, stereo preservation, V3000 structure serialization, atropisomer bond validity

Task focus: Tautomeric normalization must preserve chirality and output serializable molecules, and the stereochemistry cannot be broken in canonicalize.

Task 106 — Chemistry (91-task subset: False)

Upstream repository: `MLCIL/scikit-fingerprints`

Issue/PR: Issue #519 / PR #525 / PR #528

Knowledge scope: MAP4 / MAPC retrieval protocol, chirality invariance, representation consistency, benchmark panel, version pinning

Task focus: Molecular retrieval protocols need to maintain the same semantics across chirality, representation, and batch processing paths, and cannot be true only for visible panels.

Task 107 — Mathematics (91-task subset: False)

Upstream repository: `sympy/sympy`

Issue/PR: Issue #28186 / PR #28189

Knowledge scope: rational integration, exact root handling, log_to_real, derivative identity, real antiderivative semantics

Task focus: The original function returned by rational integration must really be able to convert back to the original integrand, and root processing cannot break the derivative identity.

Task 108 — Mathematics (91-task subset: False)

Upstream repository: `mpmath/mpmath`

Issue/PR: Issue #1104 / PR #1107; Issue #930 / PR #1111

Knowledge scope: theta3/theta4 fixed-precision paths, near-unit nome stability, tiny-nome cancellation, derivative propagation, modular transform

Task focus: Jacobi theta must also be stable on legal nomes. It cannot be rejected by mistake, nor can it explode on the path of small nomes.

Task 109 — Mechanics (91-task subset: False)

Upstream repository: `elhajjar1/wrinkleFE`

Issue/PR: Issue #252 / PR #285

Knowledge scope: multi-wrinkle field composition, fiber-angle propagation, phase offsets, FE analysis, RSS shortcut

Task focus: The multi-wrinkle field cannot be approximated by single wrinkle RSS, and the fiber angle must follow the synthesized wrinkle field.

Task 110 — Civil Engineering (91-task subset: False)

Upstream repository: `robbievanleeuwen/section-properties`

Issue/PR: Issue #272 / PR #287

Knowledge scope: first-moment result, outline quantity semantics, cross-section property consistency, reported vs constructed geometry

Task focus: The first-order moment result cannot be mistakenly reported as an outline quantity, and the report semantics and construction geometry must be consistent.

Task 111 — Mechanics (91-task subset: True)

Upstream repository: `GazzolaLab/PyElastica`

Issue/PR: Issue #81 / PR #87

Knowledge scope: tapered slender filaments, inverse-time material-response, cross-section variation, translational observations, free-response consistency

Task focus: The free response of the slender body cannot jump randomly with the change of cross section, and the material response and geometric changes must be coordinated.

Task 112 — Civil Engineering (91-task subset: False)

Upstream repository: `robbievanleeuwen/section-properties`

Issue/PR: Issue #460 / PR #462

Knowledge scope: hybrid-section plastic capacity, constituent force/centroid preservation, split geometry pieces, plastic neutral axis

Task focus: The plastic load-bearing of the composite section must preserve the force and centroid of each material segment, and cannot just report a seemingly reasonable number.

Task 113 — Electrical Engineering (91-task subset: False)

Upstream repository: `pvlib/pvlib-python`

Issue/PR: Issue #2054 / PR #2055

Knowledge scope: timezone-aware timestamps, calendar bins, day-of-year, extraterrestrial irradiance, SPA, equivalent aware representations

Task focus: Solar resource calculations must give consistent results for ”different time representations of the same moment”, and time zone normalization cannot be aliased.

Task 114 — Electrical Engineering (91-task subset: True)

Upstream repository: `PyPSA/PyPSA`

Issue/PR: Issue #1360 / PR #1371

Knowledge scope: storage history, planning-epoch boundary, cyclic vs initial state, Store/StorageUnit boundary semantics, state recurrence

Task focus: The energy storage state must abide by the cycle boundary priority rules before and after the planning boundary, and the initial state cannot be mistaken for the previous period.

Task 115 — Civil Engineering (91-task subset: False)

Upstream repository: `JWock82/Pynite`

Issue/PR: Issue #255, #275 / PR #323

Knowledge scope: Global stiffness singularity, structural stability determination, sparse/dense solution consistency, zero right end boundary

Task focus: Legal frames cannot return finite but distorted displacements when unstable. Global stiffness singularities, zero load boundaries, and sparse/dense path consistency must be determined together.

Task 116 — Civil Engineering (91-task subset: False)

Upstream repository: `JWock82/Pynite`

Issue/PR: Issue #255, #275 / PR #323

Knowledge scope: Structural response reports, equilibrium equations, matrix storage modes, physical interpretation of control studies

Task focus: Response reporting by the same framework in sparse / dense matrix storage modes must remain physically interpretable and consistent with equilibrium equations and API conventions.

Task 117 — Civil Engineering (91-task subset: False)

Upstream repository: `JWock82/Pynite`

Issue/PR: Issue #317 / PR #319

Knowledge scope: inactive member results, node displacement chord, physical member segmentation, local/global coordinate transformation

Task focus: The inactive member still needs to be linearly interpolated along the chord after displacement to ensure that the scalar / sampled results are compatible with the geometry.

Task 118 — Chemistry (91-task subset: False)

Upstream repository: `tblite/tblite`

Issue/PR: PR #326 / PR #343

Knowledge scope: GFN2-xTB multipole electrostatics, 3-D periodic Ewald decomposition, energy/force/stress consistency, unit cell translation and derivative cache

Task focus: Different legitimate periodic unit cell representations of the same molecule should give consistent energies, forces, and stresses; multipolar Ewald's real/reciprocal paths and cached derivatives must be aligned.

Task 119 — Chemistry (91-task subset: False)

Upstream repository: `MolSSI/QCEngine`

Issue/PR: PR #349

Knowledge scope: Turbomole nprhessian/hessian selection, 3N×3N Cartesian Hessian, projected/unprojected curvature, QCSchema result consistency

Task focus: Turbomole Hessian analysis must recognize nprhessian/hessian records, infer the full 3N×3N dimensions, and pass the unprojected Cartesian curvature to downstream reaction path/vibration analysis.