MECHBENCH: CAN AI SCIENTIFIC AGENTS DISCOVER MECHANISMS BEYOND PHENOMENAL LAWS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific discovery requires not only recovering mathematical laws that describe observable behavior, but also identifying the mechanisms that generate them. Existing benchmarks for symbolic regression and scientific agents primarily evaluate phenomenal-law recovery, leaving mechanism discovery largely untested. We introduce MECHBENCH, a benchmark that explicitly separates these two capabilities. Each task is defined by a mechanistic model, a structured set of scientifically meaningful relations whose joint consequences entail an observable phenomenal law, while agents receive only observational data and scientific context. We evaluate mechanism recovery through *mechanism probes*, which query internal scientific consequences that cannot be inferred from the phenomenal law alone. To reduce reliance on memorized textbook mechanisms, we construct unfamiliar variants through controlled, scientifically interpretable mutations of canonical mechanisms, and screen for *mechanistic indistinguishability* to exclude ambiguous instances admitting comparable competing mechanisms. Experiments across representative scientific agents reveal a substantial phenomenal–mechanism recovery gap: for Codex with GPT-5.6-sol, phenomenal-law accuracy reaches 35.00% on the Core-set while mechanism accuracy is only 13.75%, with mechanism recovery failing in 64.29% of cases where the phenomenal law is correctly recovered. The gap widens as mechanisms become increasingly mutated, and even providing the correct phenomenal law leaves mechanism recovery below 50%. These results reveal a substantial generalization gap in mechanistic reasoning and establish mechanism discovery as a distinct challenge beyond recovering observable scientific laws.

1 INTRODUCTION

Discovering compact mathematical laws from observations has long been a central objective of science and, more recently, of automated scientific discovery. Yet recovering a law that describes observable behavior is not the same as recovering the mechanism that produces it. In the philosophy of science, *phenomenal models* characterize the observable inputs, modulators, and outputs of a system while leaving its relevant internal causal structure unspecified, whereas *mechanistic models* further describe the components, relations, and organization responsible for generating the phenomenon (Machamer et al., 2000; Kaplan & Craver, 2011; Craver & Kaplan, 2020). The distinction is exemplified by the transition from Kepler's empirical laws of planetary motion to Newtonian mechanics: the former compactly characterize regularities among observable quantities, while the latter provides an underlying structure from which those regularities can be derived. Mechanistic models therefore make commitments beyond observational adequacy: they seek to explain how a phenomenon is generated and imply properties of scientifically meaningful internal quantities that are not specified by the observable law itself. Assessing whether AI systems can recover such mechanisms is thus a distinct and stronger test of scientific discovery than accurate equation recovery.

Large language models and scientific agents have recently shown rapidly improving capabilities for discovering mathematical regularities from data, progressing from LLM-guided symbolic regression to autonomous systems that analyze observations, conduct experiments, invoke scientific tools, and iteratively refine hypotheses (Shojaee et al., 2025a; Xia et al., 2026; Yang et al., 2026). Accordingly, recent benchmarks have substantially strengthened the evaluation of scientific law discovery: they introduce unfamiliar or counterfactual laws to reduce memorization, allow agents to actively

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.

interrogate simulated systems, and increasingly ground discovery in realistic scientific contexts and data (Shojaee et al., 2025b; Zheng et al., 2026; Cerrato et al., 2026; Huang et al., 2026). Despite these advances, their principal target remains the recovery of mathematical relations that reproduce observable system behavior. Such evaluations establish whether an agent can discover an observationally adequate law, but do not require it to reconstruct any internal mechanism from which that law follows. A recent work argues that a central missing capability in current LLMs is the abductive "jump" from observations to explanatory premises, distinguishing it from induction over observed patterns and deduction from given premises (Zahavy, 2026). This perspective highlights a capability that existing law-discovery benchmarks do not directly test: whether AI systems can infer explanatory internal structure beyond observable regularities.

Extending evaluation from law discovery to mechanism discovery is nontrivial. A benchmark must specify mechanistic structure in a form that can be evaluated objectively, distinguish genuine reconstruction from recall of familiar textbook mechanisms, and account for *mechanistic indistinguishability*, where distinct scientifically plausible mechanisms entail the same observable law. These challenges make it difficult to determine, under controlled conditions, whether the growing law-discovery capability of scientific agents extends to recovering underlying mechanisms.

To address these challenges, we introduce MECHBENCH, a benchmark that explicitly separates the recovery of observable laws from the recovery of the mechanisms that generate them, as illustrated in Figure 1. We represent a mechanism as a structured set of scientifically meaningful relations whose joint consequences entail an observable phenomenal law, allowing an agent to be evaluated not only on whether its proposed mechanism reproduces the observed input–output relationship, but also on whether it makes the correct additional commitments about internal scientific quantities. We operationalize the latter through *mechanism probes*, which query scientifically defined internal consequences that follow from the submitted mechanism but cannot be inferred from the phenomenal law and provided scientific context alone. To distinguish reconstruction from memorization, we construct unfamiliar mechanism variants through controlled, scientifically interpretable mutations of canonical mechanisms. Finally, we explicitly screen for mechanistic indistinguishability and revise or exclude instances admitting readily constructible competing mechanisms with the same phenomenal law. Together, these components turn mechanism discovery from an informal notion of "deeper understanding" into a controlled and objectively testable discovery task.

Experiments across representative scientific agents reveal a substantial gap between phenomenal-law and mechanism recovery. Taking Codex with GPT-5.6-sol as an example, phenomenal-law accuracy reaches 35.00% on the Core-set while mechanism accuracy is only 13.75%, and mechanism recovery still fails in 64.29% of cases where the phenomenal law is recovered correctly. This separation grows as mechanisms depart further from their canonical forms: phenomenal-law and mechanism recovery decline together with increasing mutation count, but mechanism recovery deteriorates more sharply. Moreover, even when the correct phenomenal law is directly provided, mechanism accuracy remains below 50% and drops rapidly on increasingly mutated mechanisms. These results expose a substantial generalization gap in mechanistic reasoning and suggest that mechanism discovery constitutes an additional, qualitatively more difficult challenge beyond observable-law recovery.

2 RELATED WORK

LLMs for Scientific Equation Discovery. Recent LLM-based symbolic-regression and scientific-agent systems increasingly combine scientific prior knowledge with program synthesis, data analysis, tool use, and iterative hypothesis refinement. LLM-SR integrates language-model proposals with evolutionary search (Shojaee et al., 2025a); SR-Scientist equips agents with data-analysis and equation-evaluation tools for iterative discovery (Xia et al., 2026); and systems such as KeplerAgent and A-SR further incorporate structural reasoning, tool orchestration, and longer agentic search trajectories (Yang et al., 2026; Zhao et al., 2026). These approaches substantially extend symbolic regression beyond blind expression search, but their primary target remains the recovery of mathematical relations that characterize observable system behavior. Our setting instead evaluates whether an agent can recover the scientific internal relations from which those observable relations follow.

Benchmarks for Scientific Law Discovery. Classical symbolic-regression benchmarks, including the Feynman dataset and SRBench, evaluate the recovery of compact equations from predefined observations (Udrescu & Tegmark, 2020; La Cava et al., 2021; Matsubara et al., 2022). More recent

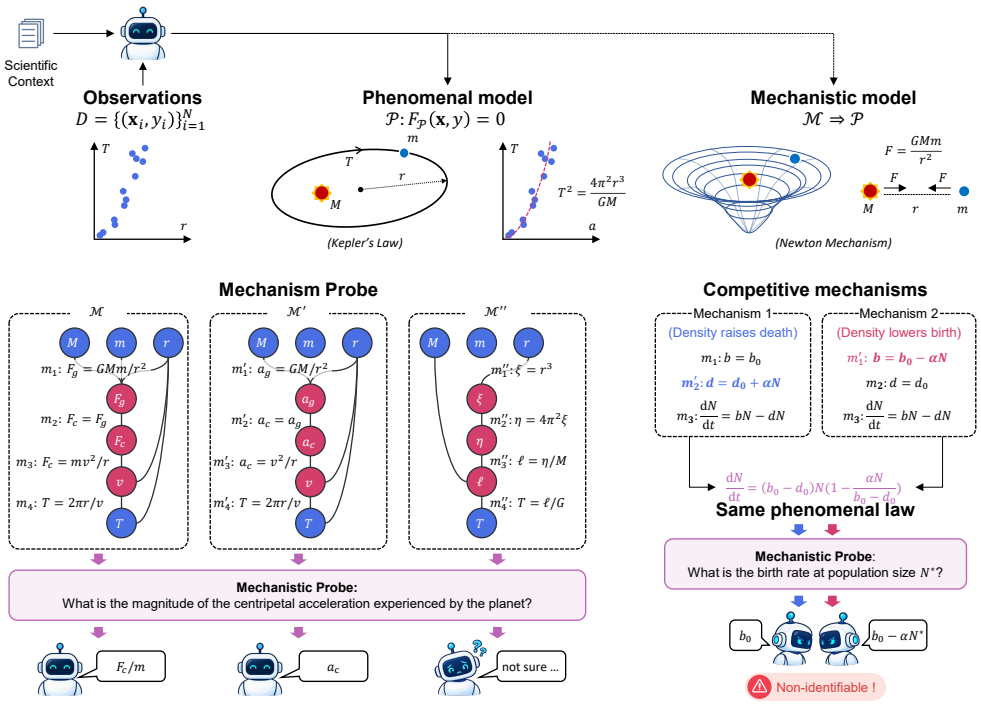


Figure 1: From phenomenal-law recovery to mechanistic discovery.

benchmarks strengthen this setting for LLM-based scientific discovery: LLM-SRBench uses transformed and synthetic equations to reduce direct memorization (Shojaee et al., 2025b); NewtonBench places altered physical laws in interactive simulated systems (Zheng et al., 2026); Science-Gym evaluates autonomous experimentation and data collection (Cerrato et al., 2026); and SciLaws-Bench combines real scientific observations with controlled, actively queryable scientific worlds (Huang et al., 2026). Together, these benchmarks increase unfamiliarity, interactivity, and scientific grounding, while their primary object of evaluation remains the recovered observable law. MechBench instead targets mechanism discovery: it evaluates whether agents can recover structured mechanistic models whose internal scientific consequences agree with those of the reference mechanism, using controlled variants, mechanism probes, and screening for mechanistic indistinguishability.

3 From Phenomenal Laws to Mechanistic Discovery

3.1 Formalizing Mechanistic Discovery

Let $\mathbf{x}$ denote a set of observable input variables and y the target variable. A phenomenal law $\mathcal{P}$ of the form $F_{\mathcal{P}}(\mathbf{x}, y) = 0$ represents their observable relationship. A mechanism $\mathcal{M} = \{m_1, \ldots, m_J\}$, on the other hand, consists of a collection of scientifically meaningful relations involving the observable variables and internal quantities $\mathbf{h}$, where $\mathbf{h}$ denotes quantities that cannot be directly observed. By eliminating $\mathbf{h}$ from $\mathcal{M}$, we can derive the phenomenal law $\mathcal{P}$:

$$\mathcal{M}(\mathbf{x}, y, \mathbf{h}) \Longrightarrow \mathcal{P}(\mathbf{x}, y). \tag{1}$$

The additional internal structure encoded by $\mathcal{M}$ allows a mechanistic model to determine scientifically meaningful quantities that cannot be inferred from $\mathcal{P}$ alone. For example, the phenomenal relation for Kepler's third law can be written as $T^2 = 4\pi^2 r^3/(GM)$. A Newtonian mechanism instead specifies gravity, orbital acceleration, force balance, and velocity–period coupling.

In this work, we formulate mechanistic discovery as the task of reconstructing a mechanistic model from observational data and scientific context. A reference mechanism $\mathcal{M}$ induces a phenomenal law $\mathcal{P}$, from which observations $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ can be sampled. As demonstrated in Figure 1, the agent receives D and scientific context C that specifies the descriptions of observable variables

and other background information, and is asked to reconstruct the underlying mechanistic model without access to the reference mechanism or its internal quantities. Mechanistic discovery is therefore formulated as

$$(C, D) \longrightarrow \hat{\mathcal{M}}. \tag{2}$$

3.2 Mechanism Probes

The same mechanism may be represented in different but scientifically equivalent ways. For example, as illustrated in Figure 1, one formulation may explicitly reason through force, whereas another may eliminate force and express the same physical relations in terms of acceleration. Mechanism recovery therefore cannot be assessed through literal agreement with the equations used to represent the reference mechanism, as exact matching of equations, intermediate symbols, or dependency graphs would conflate representational and mechanistic differences. On the other hand, evaluating a submitted mechanism only by eliminating its internal quantities and comparing the resulting phenomenal law with the reference $\mathcal{P}$ is also insufficient, because scientifically distinct mechanisms may entail the same phenomenal law.

Inspired by the mechanistic view that explaining a phenomenon requires specifying scientifically meaningful internal relations that generate the observable regularity and support correct reasoning about internal quantities that are not determined by the phenomenal law itself, we introduce *mechanism probes* to assess these additional mechanistic commitments. Specifically, we select a set $\mathcal{Z} = \{z_k\}_{k=1}^{K}$ of scientifically defined quantities such that, for a generic probe $z \in \mathcal{Z}$,

$$\mathcal{M} \Rightarrow z = g_{\mathcal{M}}(\mathbf{x}), \qquad \mathcal{P}, C \not\Rightarrow z = g_{\mathcal{M}}(\mathbf{x}). \tag{3}$$

The first condition requires each probe to capture an internal consequence of the reference mechanism, while the second ensures that it cannot be inferred from the phenomenal law and the provided scientific context alone. A properly chosen probe can then be used to assess whether $\hat{\mathcal{M}}$ reproduces the scientifically equivalent mechanistic commitments. Given its submitted mechanism, the agent is asked to derive

$$\hat{\mathcal{M}} \Rightarrow z = g_{\hat{\mathcal{M}}}(\mathbf{x}), \tag{4}$$

which can then be compared with the reference probe relation $z = g_{\mathcal{M}}(\mathbf{x})$ symbolically and numerically. Because each probe is defined by the scientific meaning of the queried quantity rather than by a particular intermediate symbol, equivalent formulations of the same mechanism should imply the same probe value even when their derivations differ. In practice, we restrict probes to quantities with clear scientific interpretations and explicit, objectively checkable expressions from the variables available in the task; Appendix A.4 details their construction.

3.3 Mechanistic Indistinguishability

Since eliminating the internal quantities $\mathbf{h}$ from $\mathcal{M}$ yields the phenomenal law $\mathcal{P}$, $\mathcal{M}$ can be viewed algebraically as a decomposition of $\mathcal{P}$. However, not every mathematically valid decomposition constitutes a mechanistic model. The provided scientific context C strongly constrains which entities, relations, and assumptions provide a meaningful explanation of the phenomenon. This constraint makes mechanism discovery well posed despite the mathematical underdetermination.

However, the problem becomes ambiguous when there exists a competing mechanism $\mathcal{M}'$ that is substantively different from the reference mechanism $\mathcal{M}$ but entails the same phenomenal law and remains comparably admissible under the provided scientific context C. As illustrated in Figure 1, the same population-level law may arise either from density-dependent suppression of birth or from density-dependent enhancement of death, while the two mechanisms make different predictions about the underlying birth rate. Such cases are a mechanism-level form of empirical equivalence and underdetermination by evidence (Hoefer & Rosenberg, 1994), which can be referred to as *mechanistic indistinguishability*. In this case, the observable evidence together with the provided scientific context offers no sufficient basis for singling out $\mathcal{M}$ over $\mathcal{M}'$, making recovery of the designated reference mechanism intrinsically ambiguous. Avoiding such ambiguity is also necessary for interpreting the mechanism probes introduced in Section 3.2. A mechanism probe tests whether a proposed mechanism $\hat{\mathcal{M}}$ reproduces the same scientifically meaningful internal consequences as the reference mechanism $\mathcal{M}$, but does not determine which of two competing mechanisms is scientifically more plausible. Probe-based agreement or disagreement therefore provides a valid operational

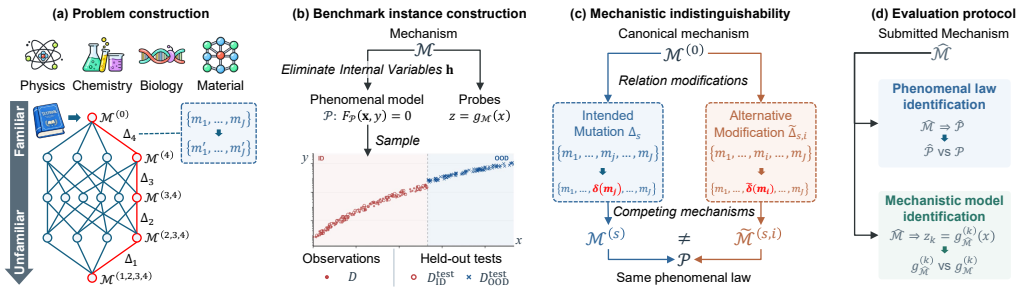


Figure 2: Construction and evaluation of MechBench. (a) Canonical mechanisms are modified and composed to create unfamiliar variants. (b) Each mechanism induces a phenomenal law used to generate observations and held-out ID/OOD tests. (c) Candidate competing mechanisms are screened for mechanistic indistinguishability. (d) Submitted mechanisms are evaluated by both phenomenal-law recovery and mechanism probes.

criterion for mechanism recovery only when no substantively different mechanism that is comparably admissible under C remains compatible with the same phenomenal evidence. We therefore require benchmark instances to avoid such competing mechanisms; the corresponding screening procedure is described in Section 4.2.

4 Benchmark Construction

4.1 Constructing Mechanism Variants

Scientific-discovery benchmarks such as AI Feynman and LLM-SRBench often construct tasks from established scientific laws and equations (Udrescu & Tegmark, 2020; Shojaee et al., 2025b). For LLM-based agents, however, canonical scientific knowledge may already be present in pretraining data, making successful recovery difficult to distinguish from memorization (Shojaee et al., 2025b). To balance scientific grounding with mechanism unfamiliarity, we select canonical scientific mechanisms from physics, chemistry, biology, and materials science and use them as blueprints for constructing unfamiliar mechanism variants (Appendix A.1).

Specifically, a mutation Δ_s applies a scientifically interpretable modification to the canonical mechanism,

$$\mathcal{M}^{(s)} = \Delta_s\left(\mathcal{M}^{(0)}\right). \tag{5}$$

A useful mutation should depart sufficiently from the canonical mechanism to prevent straightforward reproduction of its textbook derivation, while remaining scientifically admissible under the relevant domain knowledge. We therefore favor local modifications with clear scientific semantics over arbitrary symbolic perturbations. Representative mutations include changing an interaction scaling law, introducing a source or leakage term into a balance relation, replacing a linear constitutive response with a nonlinear one, or modifying a geometric constraint; additional examples and details are provided in Appendix A.3.

Compatible mutations can further be composed to construct variants that depart more substantially from the canonical mechanism and are less amenable to direct textbook recall. As illustrated in Figure 2(a), given a set of compatible mutations $\{\Delta_{s_1}, \ldots, \Delta_{s_q}\}$, we construct

$$\mathcal{M}^{(s_1,\ldots,s_q)} = (\Delta_{s_q} \circ \cdots \circ \Delta_{s_1})\left(\mathcal{M}^{(0)}\right). \tag{6}$$

The five mutations in each family are selected and, where needed, adjusted to remain compatible under composition. The resulting task family retains a shared scientific scaffold while varying which and how many elements of the canonical mechanism are modified, from the original through all mutation subsets.

4.2 Constructing Benchmark Instances

For each constructed mechanism $\mathcal{M}$, we derive its corresponding phenomenal law $\mathcal{P}$ by eliminating the internal quantities $\mathbf{h}$ and use this law to generate observable data. As illustrated in Figure 2(b), let $\mathcal{X}$ denote the scientifically meaningful range of the observable inputs. We partition this range into two disjoint regions, $\mathcal{X}_{\mathrm{ID}}$ and $\mathcal{X}_{\mathrm{OOD}}$, corresponding respectively to interpolation and extrapolation regimes. We sample observations D from $\mathcal{X}_{\mathrm{ID}}$ to provide to the agent, and we construct held-out test sets

$$D_{\mathrm{ID}}^{\mathrm{test}} = \{(\mathbf{x}_i^{\mathrm{ID}}, y_i^{\mathrm{ID}})\}_{i=1}^{N_{\mathrm{ID}}}, \quad \mathbf{x}_i^{\mathrm{ID}} \in \mathcal{X}_{\mathrm{ID}}, \qquad D_{\mathrm{OOD}}^{\mathrm{test}} = \{(\mathbf{x}_i^{\mathrm{OOD}}, y_i^{\mathrm{OOD}})\}_{i=1}^{N_{\mathrm{OOD}}}, \quad \mathbf{x}_i^{\mathrm{OOD}} \in \mathcal{X}_{\mathrm{OOD}}, \tag{7}$$

to evaluate in-domain and out-of-domain numerical consistency of the recovered phenomenal law. Appendix A.5 describes the sampling ranges, public task view, and hidden evaluation data.

To avoid constructing mechanistically indistinguishable instances, we explicitly search for competing mechanisms during instance construction. As illustrated in Figure 2(c), consider a canonical mechanism $\mathcal{M}^{(0)} = \{m_1, \ldots, m_J\}$ and a variant $\mathcal{M}^{(s)} = \Delta_s(\mathcal{M}^{(0)})$ in which the intended mutation replaces relation m_j with $\delta(m_j)$, so that $\mathcal{M}^{(s)} = \{m_1, \ldots, \delta(m_j), \ldots, m_J\}$. We derive candidate alternative modifications $\tilde{\delta}(m_i)$ of other relations m_i ($i \neq j$) that can reproduce the same induced phenomenal law, creating

$$\widetilde{\mathcal{M}}^{(s,i)} = \{m_1, \ldots, \tilde{\delta}(m_i), \ldots, m_J\}. \tag{8}$$

If such an alternative requires comparable or lower structural complexity and no greater departure from the relevant scientific background knowledge than the intended modification $\delta(m_j)$, this provides strong evidence that the resulting task is mechanistically indistinguishable, and the corresponding variant is revised or excluded.

4.3 Benchmark Evaluation Protocol

As illustrated in Figure 2(d), for a submitted mechanism $\hat{\mathcal{M}}$, our evaluation separates two complementary signals: whether $\hat{\mathcal{M}}$ recovers the phenomenal law and whether it reproduces the additional mechanistic consequences encoded by the mechanism probes. To evaluate phenomenal-law recovery, we first derive the phenomenal law implied by the submission,

$$\hat{\mathcal{M}} \Longrightarrow \hat{\mathcal{P}}, \tag{9}$$

and compare $\hat{\mathcal{P}}$ with the reference $\mathcal{P}$ using symbolic and numerical equivalence. This determines whether the proposed mechanism accounts for the observed phenomenal relationship.

To evaluate mechanism recovery, we freeze the submitted mechanism and present the scientific definition of each mechanism probe z_k to the agent independently, asking it to derive the relation between the defined quantity and the observable variables based on its submitted $\hat{\mathcal{M}}$,

$$\hat{\mathcal{M}} \Longrightarrow z_k = g_{\hat{\mathcal{M}}}^{(k)}(\mathbf{x}). \tag{10}$$

The predicted relation $g_{\hat{\mathcal{M}}}^{(k)}$ is compared with the reference relation $g_{\mathcal{M}}^{(k)}$ using the same symbolic and numerical equivalence machinery. Probes are queried independently so that information revealed by one probe cannot alter the mechanism used to answer another.

5 Experiments

5.1 Experimental Setup

Benchmark and evaluated systems. We evaluate on two MechBench subsets constructed in Section 4 and detailed in Appendix A. The *Full-set* contains 512 instances from 16 task families spanning physics, chemistry, biology, and materials science, while the *Core-set* is a fixed subset of 80 instances from eight families. We use the Core-set to compare a broader range of model–agent configurations and the Full-set to evaluate performance across more task families. We evaluate seven configurations spanning Codex, Claude Code, and DeepSeek Harness with GPT-5.6-sol, GLM-5.3-flash, DeepSeek-v4-flash-0731, and DeepSeek-v4-pro-0813. All seven configurations are evaluated

on the Core-set, while Codex with GPT-5.6-sol and GLM-5.3-flash is additionally evaluated on the Full-set. We also include PySR as a non-LLM equation-discovery baseline and PySR + Direct-Ask, where DeepSeek-v4-flash-0731 answers mechanism probes from the equation discovered by PySR.

Gold-P control. To isolate mechanism reasoning from phenomenal-law discovery, we additionally evaluate *Gold-P Direct-Ask*, where the reference phenomenal law P is provided directly and the model answers each mechanism probe independently in a tool-free turn. We evaluate this control with GPT-5.6-sol, GLM-5.3-flash, and DeepSeek-v4-flash-0731 on both the Core-set and Full-set.

Protocol and metrics. In the standard discovery setting, agents receive only the scientific context C and training observations D; the reference mechanism, mechanism probes, and held-out observations remain inaccessible during search. We report phenomenal-law accuracy $\mathrm{SA}(P)$ and task-level mechanism accuracy $\mathrm{SA}(M)$, where a task is counted as mechanistically correct only if all of its probes are symbolically correct. Detailed configurations, numerical ID/OOD metrics, scoring rules, and repeated-run aggregation are provided in Appendix B.

5.2 The Phenomenal–Mechanism Recovery Gap

Table 1 reports phenomenal-law and mechanism recovery in the standard discovery setting. Across all seven agent configurations, mechanism accuracy is consistently lower than phenomenal-law accuracy, revealing a systematic gap between recovering observable behavior and reconstructing the mechanism that generates it. Taking Codex with GPT-5.6-sol as an example, $\mathrm{SA}(P)$ reaches 35.00% on the Core-set, whereas $\mathrm{SA}(M)$ is only 13.75%; on the Full-set, the corresponding scores are 16.02% and 7.81%. The same pattern holds across the remaining configurations despite their substantially lower overall recovery rates. (See Appendix C.2 for variant-only results.)

To separate mechanism failures from failures to recover the observable law, we additionally report the conditional failure rate $1 - (M \mid P)$, which is the probability of failing mechanism recovery among instances with successful phenomenal-law recovery (i.e., $\Pr(S_M = 0 \mid S_P = 1)$). For Codex with GPT-5.6-sol, this rate is 64.29% on the Core-set and 57.32% on the Full-set, suggesting that even when the phenomenal law is recovered correctly, mechanism recovery fails in more than half of the cases. The other agent configurations exhibit the same pattern, with Core-set conditional failure rates ranging from 49.07% to 67.22%. These results suggest that mechanism recovery is not simply a stricter version of equation recovery: agents that successfully identify observable relations still frequently fail to reconstruct the internal scientific structure that produces them.

Moreover, Figure 3 examines this rate as a function of mutation count for Codex with GPT-5.6-sol. Although phenomenal-law and mechanism accuracy decline together as mutations accumulate, mechanism accuracy deteriorates more sharply, causing the conditional mechanism failure rate to rise. Specifically, the rate is approximately 33% on the Core-set and 17% on the Full-set for Original instances, and increases rapidly with additional mutations, reaching 100% at three mutations on the Core-set and four mutations on the Full-set. This widening separation suggests that increasing mechanism unfamiliarity disproportionately impairs recovery of internal mechanistic structure relative to recovery of the observable law, indicating that mechanism discovery poses an additional and qualitatively more difficult challenge than phenomenal-law discovery.

5.3 Oracle Control: Knowing the Law Is Not Enough

We use Gold-P Direct-Ask to test whether models can reliably answer mechanism probes when the correct phenomenal law is provided. According to Table 2, agents achieve higher mechanism accuracy under Gold-P Direct-Ask than under standard discovery across both datasets for every model evaluated in both settings. For GPT-5.6-sol, accuracy reaches 45.00% on the Core-set and 49.02% on the Full-set, compared with 13.75% and 7.81%, respectively, under standard discovery. Nevertheless, mechanism accuracy remains below 50% on both datasets, while the two flash models remain substantially less successful. Thus, access to the correct observable relationship does not make mechanism recovery reliable. This is partly because the phenomenal law itself does not encode the internal mechanistic information required by the probes; on the other hand, it also reflects models' limited ability to reconstruct such missing mechanistic structure even when the observable law is given, underscoring the need for dedicated evaluation of mechanism discovery.

Table 1: Symbolic recovery results on the Core-set and Full-set. $\mathrm{SA}(P)$ and $\mathrm{SA}(M)$ denote phenomenal-law and mechanism recovery, respectively.

Base Model	Agent	Core-set (n=80)			Full-set (n=512)		
		SA($\mathcal{P}$)↑	**SA($\mathcal{M}$)↑**	$1-(\mathcal{M}\mid\mathcal{P})$	**SA($\mathcal{P}$)↑**	**SA($\mathcal{M}$)↑**	$1-(\mathcal{M}\mid\mathcal{P})$
Deepseek-v4-flash-0731	PySR+Direct-Ask	**0.00%**	**1.25%**	/	**0.20%**	**0.20%**	/
GPT-5.6-sol	Codex	**35.00%**	**13.75%**	64.29%	**16.02%**	**7.81%**	57.32%
GLM-5.3-flash	Codex	6.25%	2.92%	67.22%	4.30%	1.76%	63.64%
GLM-5.3-flash	Claude Code	8.75%	3.75%	57.14%	/	/	/
Deepseek-v4-flash-0731	Codex	7.92%	3.75%	49.07%	/	/	/
Deepseek-v4-flash-0731	Claude Code	6.25%	3.75%	60.00%	/	/	/
Deepseek-v4-flash-0731	Deepseek Harness	6.25%	2.50%	60.00%	/	/	/
Deepseek-v4-pro-0813	Deepseek Harness	10.00%	3.75%	62.50%	/	/	/

Table 2: Mechanism recovery under standard discovery and Gold-P Direct-Ask on the Core-set and Full-set.

Base Model	Core-set (n=80) · SA(M)↑		Full-set (n=512) · SA(M)↑	
	Standard discovery	Gold-P Direct-Ask	Standard discovery	Gold-P Direct-Ask
GPT-5.6-sol	**13.75%**	**45.00%**	**7.81%**	**49.02%**
GLM-5.3-flash	2.92%	6.25%	1.76%	3.91%
Deepseek-v4-flash-0731	3.75%	10.00%	/	4.10%

Figure 4 further shows that mechanism recovery under Gold-P deteriorates rapidly as mutation count increases. For GPT-5.6-sol, mechanism accuracy is 87.50% on Original instances in both sets, but falls to 12.50% on the Core-set and 22.50% on the Full-set with four mutations; the two flash models remain near zero across most multi-mutation groups. The sharp degradation from Original to increasingly mutated mechanisms suggests a substantial generalization gap in mechanistic reasoning: strong performance on familiar scientific mechanisms does not readily transfer to novel mechanistic structures, highlighting the value of controlled mutations for evaluating mechanism discovery.

5.4 Comparisons Across Models, Harnesses, and Task Families

We further compare recovery performance across base models, agent harnesses, and scientific task families. Table 3 reports domain-level Core-set recovery; Appendix C.1 provides Full-set and numerical results, Appendix C.3 provides statistical comparisons, and Table 13 reports family-level results.

With Codex fixed, GPT-5.6-sol achieves substantially higher phenomenal-law and mechanism recovery than the two flash models (Table 1), while no significant difference is detected between GLM-

Table 3: Complete Core-set symbolic-recovery results by scientific domain. A dash denotes a metric that is not defined for the corresponding control or baseline.

Base model	Agent / condition	Physics		Chemistry		Biology		Materials	
		$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$
No LLM	PySR	0.00	–	0.00	–	0.00	–	0.00	–
DeepSeek-v4-flash-0731	PySR + Direct-Ask	–	5.00	–	0.00	–	0.00	–	0.00
GPT-5.6-sol	Codex	40.00	20.00	40.00	25.00	40.00	10.00	20.00	0.00
GLM-5.3-flash	Codex (3-run avg.)	11.67	11.67	5.00	0.00	3.33	0.00	5.00	0.00
GLM-5.3-flash	Claude Code	25.00	15.00	5.00	0.00	0.00	0.00	5.00	0.00
DeepSeek-v4-flash-0731	Codex (3-run avg.)	15.00	10.00	8.33	5.00	3.33	0.00	5.00	0.00
DeepSeek-v4-flash-0731	Claude Code	5.00	10.00	15.00	5.00	0.00	0.00	5.00	0.00
DeepSeek-v4-flash-0731	DeepSeek Harness	15.00	10.00	0.00	0.00	5.00	0.00	5.00	0.00
DeepSeek-v4-pro-0813	DeepSeek Harness	10.00	5.00	15.00	10.00	5.00	0.00	10.00	0.00

Figure 3: Recovery accuracy and conditional mechanism failure rate across mutation counts for Codex with GPT-5.6-sol.

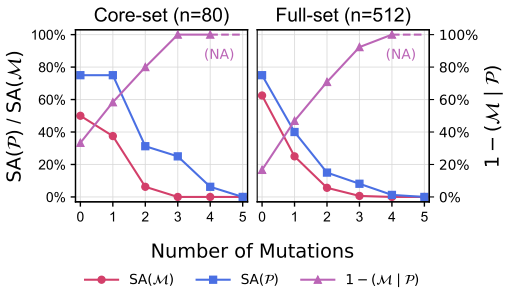
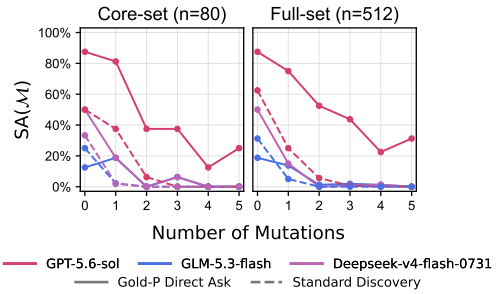


Figure 4: Mechanism recovery by mutation count under standard and Gold-P settings.



5.3-flash and DeepSeek-v4-flash-0731. GPT-5.6-sol also achieves higher accuracy under Gold-P Direct-Ask (Table 2), which may reflect a stronger ability to infer latent mechanistic structure from observable consequences and scientific context.

In contrast, holding the base model fixed, we do not detect significant performance differences among the evaluated agent harnesses. Similarly, no significant domain-level differences are detected among physics, chemistry, biology, and materials science. However, individual task families exhibit substantially different phenomenal-law and mechanism recovery profiles. For example, *Inclined Rolling* has relatively high phenomenal-law recovery but low mechanism recovery, whereas *Competitive Inhibition Pulse Reaction* shows the opposite pattern. Overall, the observed variation is more pronounced across base models and task families than across harnesses or scientific domains, indicating that mechanism-recovery difficulty is highly heterogeneous across specific task families and is not well characterized by broad domain labels alone.

6 Conclusion

We introduced MechBench to evaluate mechanism discovery beyond phenomenal-law recovery. MechBench combines controlled mechanism mutations, mechanism probes, and screening for mechanistic indistinguishability to enable objective evaluation of internal scientific structure beyond observational agreement. Across representative scientific agents, we observe a substantial gap between phenomenal-law and mechanism recovery, with the gap widening as mechanisms depart further from their canonical forms. Even when the correct phenomenal law is provided, mechanism recovery remains difficult, particularly on increasingly mutated mechanisms. Together, these results reveal a substantial generalization gap in mechanistic reasoning and establish mechanism discovery as a distinct challenge beyond recovering observable scientific laws.

A primary limitation of MechBench is that it currently focuses on mechanisms whose observable phenomenal laws can be obtained through tractable symbolic elimination. Many scientific systems do not admit such a closed-form reduction: deriving observable behavior from the underlying mechanism may require solving nonlinear or transcendental equations, integrating differential equations, or performing numerical simulation. As a result, the current benchmark covers only a subset of mechanistic discovery problems for which the mapping $\mathcal{M} \to \mathcal{P}$ can be expressed analytically. A natural extension is to bypass explicit construction of $\mathcal{P}$ and generate observations directly from the mechanism through simulation, $\mathcal{M} \to D$. Such a setting would substantially broaden the scope of mechanism discovery, but would also require new methods for validating submitted mechanisms and evaluating mechanism probes under numerical approximation, as well as revisiting mechanistic indistinguishability when equivalence can no longer be established symbolically.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation and hypothesis refinement, the development and critique of benchmark methodology and experimental design, the construction and refinement of synthetic benchmark instances, mechanism variants, and mechanism probes, and the implementation and debugging of research code. We have not used generative AI for formulating or proving mathematical theorems or for translation; tasks involving externally collected datasets or qualitative data analysis were not applicable to this work. Additionally, we used generative AI tools for literature search and summarization, brainstorming, improving presentation and readability, and suggesting paper structure and terminology. We have reviewed all AI-assisted work. Specifically, we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and verified and tested LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates scientific reasoning in AI systems using controlled benchmark tasks and synthetically generated observations derived from scientific mechanism models. It does not involve human subjects, personal or private data, or decisions affecting individuals. We foresee no ethical concerns.

REPRODUCIBILITY STATEMENT

We provide detailed documentation of benchmark construction and experimental evaluation throughout the paper and appendices. Our code, task instances, and data are publicly available at `https://anonymous.4open.science/r/mdbench-F567/README.md`.

REFERENCES

JE Bauerle. Study of solid electrolyte polarization by a complex admittance method. *Journal of Physics and Chemistry of Solids*, 30(12):2657–2670, 1969.

Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current opinion in genetics & development*, 15(2):116–124, 2005.

Bruno A Boley, Jerome H Weiner, and Ellis H Dill. Theory of thermal stresses, 1961.

Mattia Cerrato, Lennart Baur, Jannis Brugger, Sajjad Shumaly, Nicholas Schmitt, Edward Finkelstein, Selina Jukic, Lars Münzel, Felix Peter Paul, Pascal Pfannes, et al. Science-gym: A simple testbed for ai-driven scientific discovery. *Machine Learning*, 115(1):16, 2026.

William G Cochran. The comparison of percentages in matched samples. *Biometrika*, 37(3/4): 256–266, 1950.

Carl F Craver and David M Kaplan. Are more details better? on the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*, 2020.

Manfred Eigen. Proton transfer, acid-base catalysis, and enzymatic hydrolysis. part i: elementary processes. *Angewandte Chemie International Edition in English*, 3(1):1–19, 1964.

Joseph Fourier. *Theorie analytique de la chaleur, par M. Fourier*. Chez Firmin Didot, père et fils, 1822.

David E Goldman. Potential, impedance, and rectification in membranes. *The Journal of general physiology*, 27(1):37–60, 1943.

Herbert Goldstein. *Classical mechanics*. Pearson Education India, 2011.

Edwin H Hall. On a new action of the magnet on electric currents. *American Journal of Mathematics*, 2(3):287–292, 1879.

Carl Hoefer and Alexander Rosenberg. Empirical equivalence, underdetermination, and systems of the world. *Philosophy of science*, 61(4):592–607, 1994.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Yiming Huang, Ziche Liu, Zhuohang Wu, Yiqian Wang, Junxia Cui, Xinkai Zou, Linjun Mao, Nan Huang, Naicheng Yu, Kaijie Zhu, et al. Can llms discover scientific laws in real and parallel worlds? *arXiv preprint arXiv:2609.01552*, 2026.

David Michael Kaplan and Carl F Craver. The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of science*, 78(4):601–627, 2011.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de França, Ying Jin, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

Irving Langmuir. The mechanism of the catalytic action of platinum in the reactions 2co+ o2= 2co2 and 2h2+ o2= 2h2o. *Transactions of the Faraday Society*, 17(0):621–654, 1922.

Frederick A Lindemann. The radiation theory of chemical action. In *Selected Readings in Chemical Kinetics*, pp. 93–96. Elsevier, 1967.

J Ross Macdonald. Impedance spectroscopy. *Annals of biomedical engineering*, 20(3):289–305, 1992.

Peter Machamer, Lindley Darden, and Carl F Craver. Thinking about mechanisms. *Philosophy of science*, 67(1):1–25, 2000.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. Rethinking symbolic regression datasets and benchmarks for scientific discovery. *arXiv preprint arXiv:2206.10540*, 2022.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Helmut Mehrer. *Diffusion in solids: fundamentals, methods, materials, diffusion-controlled processes*. Springer, 2007.

Jacque Monod, Jeffries Wyman, Jean-Pierre Changeux, et al. On the nature of allosteric transitions: a plausible model. *J mol biol*, 12(1):88–118, 1965.

Jacques Monod. The growth of bacterial cultures. *Selected Papers in Molecular Biology by Jacques Monod*, 139(606), 2012.

Isaac Newton. *Philosophiae naturalis principia mathematica*, volume 1. G. Brookman, 1833.

Aaron Novick and Leo Szilard. Description of the chemostat. *Science*, 112(2920):715–716, 1950.

Irwin H Segel. Enzyme kinetics. *Behavior and analysis of rapid equilibrium and steady-state enzyme systems*, pp. 29–34, 1975.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 16054–16085, 2025a.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025b.

George Gabriel Stokes. On the effect of the internal friction of fluids on the motion of pendulums. *Trans. Camb. Phil. Soc.*, 9:8, 1851.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science advances*, 6(16):eaay2631, 2020.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Shijie Xia, Yuhan Sun, and Pengfei Liu. Sr-scientist: Scientific equation discovery with agentic ai. In *International Conference on Learning Representations*, volume 2026, pp. 75787–75811, 2026.

Jianke Yang, Ohm Venkatachalam, Mohammad Kianezhad, Sharvaree Vadgama, and Rose Yu. Think like a scientist: Physics-guided llm agent for equation discovery. *arXiv preprint arXiv:2602.12259*, 2026.

Tom Zahavy. Position: Llms can't jump. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Wenxiao Zhao, Dong Liu, Kaiyi Xu, Feng Liu, Zhen Zhao, Fei Ben, Shu Wang, Wenhao Li, Yingnian Wu, Fenghua Ling, et al. A-sr: Self-evolving agentic llms for symbolic regression via hierarchical coordination. *arXiv preprint arXiv:2608.04872*, 2026.

Tianshi Zheng, Kiu Wai Tam, Kim Hue Nam Nguyen, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 99018–99088, 2026.

A Task Construction Details

A.1 Task Families and Reference Mechanisms

MechBench contains 16 task families, with four families drawn from each of physics, chemistry, biology, and materials science. Each family is organized around a standard scientific construction. To ensure that every internal quantity and observable has an unambiguous value, we combine canonical relations with explicit regime assumptions and, where needed, synthetic but scientifically interpretable closures. Table 4 summarizes the scientific blueprint of each Original instance.

Physics. The four Original mechanisms instantiate familiar physical models. *Circular Orbit* (Newton, 1833) uses Newtonian inverse-square gravity and circular-motion balance to model the orbital period of a test body. *Hall Response* (Hall, 1879) uses the classical carrier-transport tensor under an open transverse circuit, together with an offset contact geometry, to model a measured Hall voltage. *Inclined Rolling* (Goldstein, 2011) uses the coupled translational and rotational dynamics of a rigid body subject to a no-slip constraint. *Terminal Settling* (Stokes, 1851) uses the steady gravity–buoyancy balance of a sphere in the low-Reynolds-number Stokes regime.

Chemistry. The four Original mechanisms model transient chemical responses using state-population balances. *Acid–Base Buffer Relaxation* (Eigen, 1964) uses proton exchange between the solution and buffer sites to model the pH response to a pulse. *Competitive Inhibition Pulse Reaction* (Segel, 1975) uses transitions among free, substrate-bound, and inhibitor-bound enzyme states to model product formation under competition. *Lindemann Pulse Reaction* (Lindemann, 1967) uses collisional activation, deactivation, and product formation to model the response of a unimolecular reaction. *Surface Pair Catalysis* (Langmuir, 1922) uses adsorption, desorption, paired surface-complex formation, and surface reaction to model catalytic product formation.

Biology. The four Original mechanisms represent distinct forms of biological regulation and transport. *Allosteric Regulation* (Monod et al., 1965) uses ligand occupancies across two protein conformations, with conformation-dependent catalytic turnover, to model activity. *Gene Repression* (Bintu et al., 2005) uses mutually exclusive promoter occupancies for repressor and polymerase to model

Table 4: The 16 task families and the scientific constructions used for their Original reference mechanisms.

Domain	Task family	Description
Physics	Circular Orbit	Models the orbital period of a test body using Newtonian inverse-square attraction, circular radial balance, and the speed–period relation.
	Hall Response	Models the voltage measured across offset contacts using a classical carrier-conductivity tensor and an open transverse circuit.
	Inclined Rolling	Models the downslope acceleration of a rigid body from coupled translational and rotational balances under no-slip rolling.
	Terminal Settling	Models the terminal speed of a sphere from gravity, buoyancy, and linear Stokes drag at low Reynolds number.
Chemistry	Acid–Base Buffer Relaxation	Models transient pH response through acid–base buffer capacities and coupled proton-exchange population balances.
	Competitive Inhibition Pulse Reaction	Models product response through mass-action balances among free enzyme, substrate-bound enzyme, and inhibitor-bound enzyme.
	Lindemann Pulse Reaction	Models product response with a collisionally activated Lindemann scheme containing activation, deactivation, and product formation.
	Surface Pair Catalysis	Models catalytic product response through adsorption, desorption, paired surface-complex formation, and surface reaction.
Biology	Allosteric Regulation	Models catalytic activity using a two-conformation statistical occupancy model with conformation-dependent turnover.
	Gene Repression	Models transcription initiation through mutually exclusive promoter occupancies and productive polymerase binding.
	Membrane Channel Flux	Models membrane current density by coupling channel-state occupancy to constant-field electrodiffusive ion flux.
	Population Resource Balance	Models stationary population density using chemostat resource balances and saturating resource-limited growth.
Materials	Composite Heat Transport	Models heat flux by one-dimensional Fourier transport through two material layers in series.
	Grain-Boundary Electrical Response	Models effective conductivity through series grain-interior and grain-boundary complex admittances in a brick-layer geometry.
	Thermoelastic Heating Response	Models stress from preparation-dependent thermal relaxation, thermal eigenstrain, and elastic bar–frame compatibility.
	Vacancy Diffusion	Models diffusivity from equilibrium vacancy formation and thermally activated jumps in a three-dimensional random walk.

transcription initiation. *Membrane Channel Flux* (Goldman, 1943) couples closed and open channel states to constant-field electrodiffusion to model ionic current density. *Population Resource Balance* (Monod, 2012; Novick & Szilard, 1950) uses steady chemostat balances and saturating resource-limited growth to model population density.

Materials science. The four Original mechanisms describe transport and coupled thermomechanical responses in materials. *Composite Heat Transport* (Fourier, 1822) uses one-dimensional Fourier conduction through two layers in series to model heat flux. *Grain-Boundary Electrical Response* (Bauerle, 1969; Macdonald, 1992) uses grain-interior and grain-boundary admittances in a brick-layer geometry to model effective conductivity. *Thermoelastic Heating Response* (Boley et al., 1961) combines preparation-dependent thermal relaxation, thermal eigenstrain, and bar–frame compatibility to model stress. *Vacancy Diffusion* (Mehrer, 2007) combines equilibrium vacancy formation with thermally activated atomic jumps in a three-dimensional random walk to model diffusivity.

A.2 Full-set and Core-set Membership

For each of the 16 families, the Full-set contains the Original mechanism and all 31 nonempty combinations of five family-specific base mutations, for 32 instances per family and 512 instances in

Mutation type	Scientific interpretation	Representative changes in the benchmark
Constitutive or component response	Change how a component property depends on its local state or an observable control.	Mobile carrier trapping in Hall Response; temperature-difference-dependent conductivity in Composite Heat Transport; pressure-dependent migration barrier in Vacancy Diffusion.
Internal state or storage	Add a scientifically meaningful population, compartment, reservoir, or relaxation mode.	Product-bound enzyme and surface-product intermediates; inactive channel state; slow thermal and anelastic reservoirs.
Additional transport or reaction channel	Open a parallel pathway that contributes a new flux while retaining the original pathway.	Finite-range orbital attraction, conductive heat bypass, fast grain-boundary path, and alternate vacancy-migration channel.
Source, sink, or balance modification	Add a force, production, loss, or exchange term to a governing balance.	Resisting torque in rolling, external lift in settling, direct Lindemann product formation, and feed-dependent mortality.
Geometric or kinematic constraint	Change how geometric quantities or motions are related without directly editing the final observable law.	Non-Euclidean circumferential radius, controlled rolling slip, conducting-area ratio, and pressure-dependent jump distance.

Table 5: Scientific interpretations of the mechanism mutations.

total. The Core-set is a fixed subset containing two families per domain: Circular Orbit and Hall Response (physics); Lindemann Pulse Reaction and Surface Pair Catalysis (chemistry); Allosteric Regulation and Membrane Channel Flux (biology); and Thermoelastic Heating Response and Vacancy Diffusion (materials science). Each selected family contributes ten existing Full-set instances: the Original, two single-mutation, two double-mutation, two triple-mutation, two four-mutation, and one five-mutation instance. Thus the Core-set contains 80 instances and is stratified by mutation count rather than generated as a separate collection.

A.3 Mechanism Mutations and Their Composition

A base mutation changes one or more relations within the reference mechanism. The phenomenal law for the resulting variant is then derived from the modified mechanism rather than edited directly. We chose each mutation to admit a concrete scientific reading and to preserve the dimensional consistency and operating regime of its family. The same mathematical form can have different meanings in different families, and conversely a single scientific change may require several equations. For example, introducing a reversible chemical intermediate requires a new state, forward and reverse fluxes, and corresponding changes to multiple population balances. We group the changes descriptively in Table 5; these categories provide directional guidance only, do not cover every possible mechanism mutation, and may not apply to every task family.

Family-specific base mutations. Every family has five labeled mutations $\Delta_1, \ldots, \Delta_5$. Tables 6, 7, 8, and 9 catalog their scientific content for physics, chemistry, biology, and materials science, respectively.

Table 6: The five base mutations for each physics task family.

Family	Mutation
Circular Orbit	$\triangle_1$: Make relative inertia depend on radius. $\triangle_2$: Add a finite-range central attraction. $\triangle_3$: Modify circumferential geometry together with its radial derivative. $\triangle_4$: Add an external central trapping field. $\triangle_5$: Store tangential momentum in a coupled internal reservoir.
Hall Response	$\triangle_1$: Trap a density-dependent fraction of electrons. $\triangle_2$: Change the electron magnetic-deflection response. $\triangle_3$: Add a minority-hole transport channel. $\triangle_4$: Change the transverse electric-force response. $\triangle_5$: Add a longitudinal bypass carrying no transverse Hall current.
Inclined Rolling	$\triangle_1$: Make the body's rotational inertia angle-dependent. $\triangle_2$: Replace no-slip rolling with a controlled slip ratio. $\triangle_3$: Couple an internal rotor to body rotation. $\triangle_4$: Add an angle-dependent resisting torque. $\triangle_5$: Apply an external upslope tangential force.
Terminal Settling	$\triangle_1$: Add a saturating surface-friction channel. $\triangle_2$: Reduce displaced fluid volume through accessible pores. $\triangle_3$: Change the particle-local fluid viscosity. $\triangle_4$: Apply a radius-dependent external upward force. $\triangle_5$: Add a parallel dissipative internal-flow channel.

Table 7: The five base mutations for each chemistry task family.

Family	Mutation
Acid–Base Buffer Relaxation	$\triangle_1$: Split buffer sites between two acid families. $\triangle_2$: Introduce a slowly exchanging interior buffer region. $\triangle_3$: Add a diffusively coupled solvent pocket. $\triangle_4$: Add proton-catalyzed exchange without changing equilibrium capacity. $\triangle_5$: Insert a proton-storing relay between solution and primary buffer.
Competitive Inhibition Pulse Reaction	$\triangle_1$: Add substrate-assisted catalytic release. $\triangle_2$: Add inhibitor-assisted association to the inhibited state. $\triangle_3$: Insert a product-bound enzyme intermediate. $\triangle_4$: Allow slow reversible locking of inhibitor-bound enzyme. $\triangle_5$: Allow a mixed substrate–inhibitor enzyme complex.
Lindemann Pulse Reaction	$\triangle_1$: Add a two-collision irreversible deactivation channel. $\triangle_2$: Insert a second reactive activated intermediate. $\triangle_3$: Allow reversible nonreactive cage storage before activation. $\triangle_4$: Add a reversible two-collider activation–deactivation channel. $\triangle_5$: Open a direct product-forming channel from unactivated reactant.
Surface Pair Catalysis	$\triangle_1$: Insert a product-bearing surface intermediate. $\triangle_2$: Allow reversible immobilization of singly adsorbed A. $\triangle_3$: Add reversible collision-assisted A exchange beside adsorbed B. $\triangle_4$: Allow gas-A-assisted reversible exchange at the B site. $\triangle_5$: Open a gas–surface reaction channel from singly adsorbed B.

Table 8: The five base mutations for each biology task family.

Family	Mutation
Allosteric Regulation	$\triangle_1$: Add an inactive conformation binding three regulators. $\triangle_2$: Couple inactive-state substrate affinity to substrate exposure. $\triangle_3$: Add another productive active-state complex. $\triangle_4$: Partition regulator into a local compartment. $\triangle_5$: Change inactive-conformation catalytic turnover.
Gene Repression	$\triangle_1$: Add a looped state containing two repressors. $\triangle_2$: Allow a weakly productive repressor–polymerase complex. $\triangle_3$: Change repressor affinity with polymerase exposure. $\triangle_4$: Partition polymerase into a promoter-local compartment. $\triangle_5$: Add a repressor-dependent paused-polymerase state.
Membrane Channel Flux	$\triangle_1$: Make channel opening voltage-dependent. $\triangle_2$: Add an extracellular-concentration-dependent inactivated state. $\triangle_3$: Partition the intracellular ion compartment. $\triangle_4$: Make channel permeability voltage-dependent. $\triangle_5$: Introduce a second monovalent species with proportionally clamped reservoirs.
Population Resource Balance	$\triangle_1$: Add feed-dependent mortality. $\triangle_2$: Partition organisms between active and protected compartments. $\triangle_3$: Add resource-consuming maintenance. $\triangle_4$: Reduce maximum birth frequency at high feed concentration. $\triangle_5$: Partition feed resource into usable and inaccessible pools.

Table 9: The five base mutations for each materials-science task family.

Family	Mutation
Composite Heat Transport	$\triangle_1$: Make the first component's conductivity depend on the imposed temperature difference. $\triangle_2$: Add a composition-dependent thermal interface resistance. $\triangle_3$: Introduce a parallel conductive bypass. $\triangle_4$: Change the second component's conducting area. $\triangle_5$: Apply controlled volumetric heating within the second component.
Grain-Boundary Electrical Response	$\triangle_1$: Add a thin space-charge region beside each boundary. $\triangle_2$: Add delayed boundary polarization. $\triangle_3$: Allow a parallel fast boundary path. $\triangle_4$: Add delayed polarization inside grains. $\triangle_5$: Include an additional boundary-local interface impedance.
Thermoelastic Heating Response	$\triangle_1$: Add a second thermal equilibration mode. $\triangle_2$: Add a recoverable anelastic strain reservoir. $\triangle_3$: Include finite elastic compliance of an adhesive layer. $\triangle_4$: Add a slow thermal-expansion reservoir. $\triangle_5$: Add a second anelastic strain-relaxation time.
Vacancy Diffusion	$\triangle_1$: Trap a temperature-dependent fraction of vacancies. $\triangle_2$: Add a quadratic pressure contribution to the migration barrier. $\triangle_3$: Contract the effective atomic jump distance. $\triangle_4$: Introduce correlated return jumps. $\triangle_5$: Open a parallel activated migration channel.

Composition rule. For a family with mutation index set $S \subseteq \{1, \ldots, 5\}$, we construct one mechanism $\mathcal{M}^{(S)}$. If mutations act on disjoint relations, the modified relations are simply included together. If they share a state or balance equation, we construct the joint equation from the union of their fluxes or components, so that every contribution occurs once and conservation is maintained. Shared constants and unchanged relations are likewise merged, not duplicated. This construction makes the compositions order-independent: the name "Variant 1-3-5," for example, denotes the set $S = \{1, 3, 5\}$. The five base mutations in every family were selected and, where necessary, adjusted so that all subsets are compatible; hence the Full-set retains all $2^5 = 32$ choices of S, including $S = \emptyset$ as Original.

We measure variant complexity by $|S|$, the number of selected base mutations. A single base mutation may modify several related equations, but it still contributes one to $|S|$. We retain a composed mechanism only when its equations admit a unique explicit algebraic solution for every internal quantity and the target as functions of the input and auxiliary variables. After composing $\mathcal{M}^{(S)}$, we eliminate its internal variables to obtain $\mathcal{P}^{(S)}$ and derive the corresponding probe answers. All variants are therefore constructed at the mechanism level before elimination.

Composition example: Vacancy Diffusion. This family provides a compact example. Omitting unchanged zero-valued channels, the Original mechanism contains

$$c_v = \exp\left[-\frac{E_f + p\Omega}{k_B T}\right], \qquad c_m = c_v, \tag{11}$$

$$E_b = E_m, \qquad r = \nu \exp\left[-\frac{E_b}{k_B T}\right], \qquad D = \frac{a^2 c_m r}{6}. \tag{12}$$

Mutation 1 replaces only the mobile-vacancy relation with

$$c_m = \frac{c_v}{1 + 3\exp[-E_t/(k_B T)]}, \tag{13}$$

representing temperature-dependent trapping. Mutation 2 instead replaces the migration-barrier relation with

$$E_b = E_m + \frac{\Omega p^2}{p_{\mathrm{ref}}}, \tag{14}$$

representing a nonlinear pressure contribution. Applying both changes gives the joint mechanism containing Eqs. 13 and 14, and elimination yields

$$\mathcal{P}^{(1,2)}: \quad D = \frac{a^2 \nu}{6} \frac{\exp[-(E_f + p\Omega)/(k_B T)] \exp[-(E_m + \Omega p^2/p_{\mathrm{ref}})/(k_B T)]}{1 + 3\exp[-E_t/(k_B T)]}. \tag{15}$$

Thus the combined phenomenal law is a consequence of two changes made at different mechanistic locations. We probe the mobile vacancy fraction c_m and the migration barrier E_b separately, so matching their combined effect on D is not sufficient to recover the mechanism.

A.4 Mechanism Probe Construction

Selection principles. For each instance, we select scientifically interpretable internal quantities whose values follow from the reference mechanism but are not fixed by the phenomenal law and the provided scientific context alone. A probe is specified by a verbal description of the quantity and a reference expression $z = g_{\mathcal{M}}(\mathbf{x})$. The expression is obtained by solving the mechanism relations and eliminating the remaining internal variables, so it can ultimately be evaluated from the observable inputs and known numerical constants. Although the final expression is a function of observable inputs, deriving it still requires internal constitutive, balance, geometric, or dynamical relations absent from $\mathcal{P}$.

We use probes with clear scientific meanings. Eligible probe quantities must contribute to the derivation of the phenomenal law: if removing the relations involving a quantity leaves the same phenomenal law derivable, that quantity is omitted. A probe quantity must also have the same value across comparably plausible mechanisms consistent with the observable evidence; otherwise, scoring it would reward agreement with the designated reference rather than recovery of a mechanism supported by the available evidence. Mechanistic-indistinguishability screening in Section 3.3 enforces the latter requirement.

Split	Region for each x_j	Samples	Agent access
Training	$[l_j, b_j)$	1,000	Yes
ID test	$[l_j, b_j)$	1,000	No
OOD test	$[b_j, u_j)$	1,000	No

Table 10: Default observation-generation configuration.

Example. In the Vacancy Diffusion composition above, the phenomenal law contains the product of the mobile vacancy fraction and an activated jump rate. The probe for c_m requires the equilibrium vacancy relation and the trapping correction in Eq. 13, whereas the probe for E_b requires the pressure-dependent barrier relation in Eq. 14. Their reference answers are

$$c_m = \frac{\exp[-(E_f + p\Omega)/(k_B T)]}{1 + 3\exp[-E_t/(k_B T)]}, \qquad E_b = E_m + \frac{\Omega p^2}{p_{\mathrm{ref}}}. \tag{16}$$

Both are functions of observable inputs and known constants after the mechanism has been applied, but neither factorization is determined by the single expression for D alone. Querying them separately therefore distinguishes recovery of the trapping and pressure-barrier relations from recovery of their aggregate observable effect.

Variation across instances. Probe sets are adapted to the active mechanism relations rather than held fixed across a family. The Original instance uses a representative internal quantity from its reference mechanism. A single mutation may change that quantity or replace it with a more diagnostic internal consequence; a composition may require several probes to cover changes at distinct mechanistic locations. We do not require one probe per mutation because one internal quantity can reflect several coupled changes, while some mutations introduce more than one scientifically relevant consequence. Across the 512 Full-set instances, K ranges from one to five: 94 instances have one probe, 314 have two, 68 have three, 30 have four, and six have five.

A.5 Observation Generation and Task Specification

Reference expressions and valid domains. Before generating observations, we solve the mechanism equations for every internal variable and the target as explicit functions of the input and auxiliary variables. For each observable source variable x_j, we choose an instance-specific lower limit l_j, extrapolation boundary b_j, upper limit u_j, and either a uniform or log-uniform sampling distribution according to its scientific meaning, characteristic scale, and physically meaningful operating regime. The resulting ranges also avoid known singularities or invalid regimes.

ID and OOD sampling. We independently sample all source variables using their specified distributions. Training observations and the held-out ID test set use $x_j \in [l_j, b_j)$ for every j, whereas the OOD test set uses $x_j \in [b_j, u_j)$ for every j. Thus ID testing uses new draws from the same region as training, while OOD testing evaluates joint high-side extrapolation across the source variables. The default configuration, used throughout the benchmark, is summarized in Table 10.

The target and all internal quantities are evaluated from the solved reference mechanism for each sampled input. Samples producing a non-real or non-finite value for any derived quantity are discarded and resampled. Generation uses a task-local pseudorandom generator with default seed 0. No observational noise is added.

Information exposed during discovery. The agent receives a short scientific task description; the name, description, unit, and role of each observable input and target; the row order of the data array; and the training observations. The array has shape $(n_{\mathrm{variables}}, 1000)$, with the target in the first row followed by the input and auxiliary variables. The task description is identical across all variants of a family, and neither the instance identifier nor the active mutations are disclosed. Sampling limits and distributions, internal variables, the reference $\mathcal{M}$ and $\mathcal{P}$, mechanism-probe definitions and answers, and the two held-out test sets remain inaccessible during discovery. For example, the discovery view for the Vacancy Diffusion composition in Eqs. 13–15 contains the metadata shown below together with a four-row training array in the listed variable order.

```
{
  "task_description": "Determine atomic diffusivity from temperature,
hydrostatic pressure, and lattice spacing.",
  "variables": [
    {"name": "diffusivity", "description": "Atomic tracer
diffusivity.",
      "unit": "m^2 s^-1", "role": "target"},
    {"name": "temp", "description": "Temperature.",
      "unit": "K", "role": "input"},
    {"name": "pressure", "description": "Hydrostatic pressure.",
      "unit": "Pa", "role": "input"},
    {"name": "spacing", "description": "Lattice spacing.",
      "unit": "m", "role": "input"}
  ],
  "data_columns": ["diffusivity", "temp", "pressure", "spacing"],
  "data_layout": "variables_by_samples"
}
```

The agent must infer the observable law and its internal explanation from the shared scientific context and observations; variant-specific metadata is not exposed.

Reference feedback script. During the search process, the agent may repeatedly run a feedback script on a candidate set of equations together with the public metadata and training observations. The script is provided as a reference implementation that helps the agent understand the required mechanism-submission format, how the submitted equations are algebraically eliminated to obtain the observable target, and how the resulting expression is evaluated on the training set. For a valid candidate, it returns the explicit solution implied by the submitted equations and numerical fit metrics computed on the training observations; malformed, inconsistent, or unsolved candidates instead return an error. It reveals no information beyond the already available task description and training data, and does not access or report the reference mechanism, the reference phenomenal law, mechanism probes, or held-out test sets.

Post-search probe queries. After the search process ends, the submitted mechanism is locked, and each mechanism probe is presented in an independent query. The query provides the scientific description of the requested probe quantity, but not the reference probe answer or any held-out observations. Separating this later query from the discovery interface prevents the probe itself from revealing the intended internal structure while the agent is constructing its mechanism.

B Experimental Details

B.1 Evaluated Configurations

Table 11 summarizes the model–agent configurations and control conditions used in our experiments. All seven standard-discovery configurations are evaluated on the Core-set. For the Full-set, we evaluate Codex with GPT-5.6-sol and GLM-5.3-flash. PySR, PySR + Direct-Ask, and the Gold-P Direct-Ask controls are evaluated on both sets.

For standard discovery, Core-set results for Codex with GLM-5.3-flash and DeepSeek-v4-flash-0731 are averaged over three independent runs. All other reported standard-discovery configurations use a single run. For each repeated configuration, metrics are first computed separately using the fixed task denominator and are then averaged across runs rather than pooling predictions across runs.

B.2 Standard Discovery and Baselines

Standard discovery. All agents interact with the same benchmark interface described in Appendix A.5. For each task, the agent receives the scientific task description, observable-variable metadata, and training observations D. During search, it may evaluate candidate mechanisms using the reference feedback script, which solves the submitted equations when possible and reports objective fit information computed only from the training observations. The script does not access the reference mechanism, the reference phenomenal law, mechanism probes, or held-out ID/OOD

Setting	Base Model	Agent / Pipeline	Core	Full
Standard discovery	GPT-5.6-sol	Codex	✓	✓
Standard discovery	GLM-5.3-flash	Codex	✓	✓
Standard discovery	DeepSeek-v4-flash-0731	Codex	✓	
Standard discovery	GLM-5.3-flash	Claude Code	✓	
Standard discovery	DeepSeek-v4-flash-0731	Claude Code	✓	
Standard discovery	DeepSeek-v4-flash-0731	DeepSeek Harness	✓	
Standard discovery	DeepSeek-v4-pro-0813	DeepSeek Harness	✓	
Equation baseline	–	PySR	✓	✓
Pipeline baseline	DeepSeek-v4-flash-0731	PySR + Direct-Ask	✓	✓
Gold-P	GPT-5.6-sol	Direct-Ask	✓	✓
Gold-P	GLM-5.3-flash	Direct-Ask	✓	✓
Gold-P	DeepSeek-v4-flash-0731	Direct-Ask	✓	✓

Table 11: Model–agent configurations, baselines, and controls used in the experiments.

observations. After search terminates, the final mechanism submission is locked before any mechanism probe is revealed.

PySR. We use PySR as a non-LLM baseline for phenomenal-law discovery. PySR receives the same observable training data and attempts to recover an equation relating the observable inputs to the target. Its recovered equation is evaluated against the reference phenomenal law using the same phenomenal-law evaluation procedure as agent submissions.

PySR + Direct-Ask. To test whether an equation recovered by a conventional symbolic-regression method can support subsequent mechanism reasoning, we additionally evaluate a two-stage PySR + Direct-Ask pipeline. PySR first discovers a phenomenal equation from the training observations. DeepSeek-v4-flash-0731 is then provided with this recovered equation and independently answers the mechanism probes.

Gold-P Direct-Ask. Gold-P Direct-Ask removes phenomenal-law discovery entirely. For each task, the model is provided with the reference phenomenal law P together with the numerical constants required to interpret the requested quantity. Each mechanism probe is asked in an independent, tool-free query. The model is not provided with the reference mechanism or the reference probe expression. Success therefore depends on reconstructing enough internal scientific structure to derive the requested quantity from the observable law and context.

B.3 Evaluation Metrics

Let P_i and $\widehat{P}_i$ denote the reference and predicted phenomenal laws for task i, and let g_{ik} and $\widehat{g}_{ik}$ denote the reference and predicted expressions for its k-th mechanism probe. Writing $\equiv$ for symbolic equivalence and K_i for the number of probes associated with task i, we define

$$\mathrm{SA}(P) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\widehat{P}_i \equiv P_i\right], \tag{17}$$

and

$$\mathrm{SA}(M) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K_i} \mathbf{1}[\widehat{g}_{ik} \equiv g_{ik}]. \tag{18}$$

Thus, $\mathrm{SA}(M)$ is a strict task-level mechanism-recovery metric: a task is counted as correct only when all of its mechanism probes are symbolically correct. It is unconditional with respect to phenomenal-law recovery, so a mechanism score is evaluated regardless of whether $\widehat{P}_i$ is correct.

For the Core-set, we additionally report numerical agreement on the held-out ID and OOD splits. For $S \in \{\mathrm{ID}, \mathrm{OOD}\}$, let y_{ij}^{S} and $\widehat{y}_{ij}^{S}$ denote the reference and predicted target values for held-out

sample j of task i. We define

$$\mathrm{Acc}_{0.1}^{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\max_{j\in S_i}\left|\frac{\widehat{y}_{ij}^{S} - y_{ij}^{S}}{y_{ij}^{S}}\right| \le 0.1\right]. \tag{19}$$

This is a task-level accuracy rather than the fraction of individual samples within tolerance: a task is counted as correct only if its maximum relative error over the entire held-out split does not exceed 10%.

B.4 Mutation Strata and Reported Subsets

We report results both over the complete benchmark subsets and after stratifying instances by mechanism familiarity. The *Original* instance of each task family has mutation count zero, while constructed variants contain between one and five applied base mutations. The Full-set contains one Original and all 31 nonempty mutation combinations for each of its 16 task families. The Core-set is a fixed subset stratified by mutation count, as detailed in Appendix A.2.

For variant-only analyses, we exclude the Original instance from every included task family, yielding 496 Full-set variants and 72 Core-set variants. Results stratified by mutation count and the complete variant-only results are reported in Appendix C.2.

B.5 Implementation and Runtime Settings

Search and wall-clock budgets. Standard discovery uses a 900-second wall-clock limit per task for Codex, Claude Code, DeepSeek Harness, and PySR. Codex, Claude Code, and DeepSeek Harness have no benchmark-imposed maximum number of agent turns and no cap on calls to the observable-fit feedback script; search is limited by wall-clock time instead. In particular, the reported DeepSeek Harness runs set `dsh-feedback-evaluations=0`, selecting the time-based prompt rather than the experimental fixed-evaluation prompt. No hard monetary budget is enabled for any reported agent configuration. Mechanism-probe turns following standard discovery have a 120-second limit each and are executed independently from the same frozen discovery result.

Reasoning settings. Codex uses reasoning effort `xhigh` with GPT-5.6-sol, `high` with GLM-5.3-flash, and `low` with DeepSeek-v4-flash-0731. Claude Code uses `high` with GLM-5.3-flash and `low` with DeepSeek-v4-flash-0731. Both DeepSeek Harness configurations use `low`. These settings control the provider's reasoning mode rather than the number of benchmark feedback evaluations.

Codex invocation. Each discovery task launches `codex exec` with `--json`, `--skip-git-repo-check`, `--color never`, the selected `--model`, and `model_reasoning_effort`; the prompt is supplied on standard input and the final message is written with `-o`. The benchmark additionally sets `approval_policy="never"`, disables web search, apps, plugins, hooks, multi-agent operation, memories, and shell snapshots, and installs a per-task permission policy. That policy grants write access only to the temporary task workspace, denies the repository, answers, credentials, and shell startup files, and allows network access only to the loopback feedback server. Probe turns use `codex exec resume` on an independent copy of the frozen discovery transcript with a read-only sandbox and all shell, code-execution, browser, image, and computer tools disabled.

Claude Code invocation. Discovery uses `claude --print --verbose --output-format stream-json --include-partial-messages --input-format text --setting-sources '' --strict-mcp-config --disable-slash-commands --permission-mode dontAsk`, an empty MCP configuration, and `--tools Bash,Read,Write,Edit,Glob,Grep`, followed by the selected `--model` and `--effort`. A generated settings file disables hooks and memory, denies access outside the isolated task workspace, protects `problem.json` and `train.npy` from writes, and restricts network access to the feedback host. Each probe uses `--resume <frozen-session> --fork-session` with `--tools ''`; consequently probe branches cannot alter either the frozen model or one another. The optional Claude Code `--max-budget-usd` setting is unset in the reported experiments.

DeepSeek Harness invocation. Discovery uses `dsh --profile headless --patch <generated-patch> <task-prompt>`. The generated patch fixes the requested OpenRouter model, disables web search/fetch, subagents, workflows, and related tools, and routes model requests through the benchmark's metered loopback gateway. Probe profiles additionally disable filesystem, shell, job, skill, todo, and goal tools. The two reported models are `deepseek/deepseek-v4-flash-0731` and `deepseek/deepseek-v4-pro-0813`; OpenRouter failures do not fall back to a different provider. The live discovery checkpoint is `working_model.txt`. The gateway records token and monetary usage but enforces no hard dollar limit.

Prompt comparability. The three harnesses use prompts with the same scientific objective, observed data, mechanism requirements, access to the observable-fit feedback script, equation-only submission contract, 900-second budget, and isolation rules. Harness-specific wording and checkpoint conventions remain; DeepSeek Harness, for example, refers to `working_model.txt` rather than `submission.txt`. Each CLI also retains its native system instructions and tool descriptions. After discovery, all three standard agents receive the same benchmark-generated mechanism-probe task prompt, and every probe is evaluated in an independent branch from the frozen submitted model.

Sampling, context, and output limits. The standard Codex and Claude Code launches do not explicitly set temperature, top-p, maximum context length, or maximum output tokens; these remain the selected CLI/provider defaults and no uniform values are available in the run receipts. DeepSeek Harness likewise does not override temperature or top-p, but caps each model response at 16,000 output tokens.

PySR hyperparameters. PySR is run for the same 900-second wall-clock budget with `niterations=1,000,000`, so the timeout rather than the nominal iteration count normally terminates search. We use 31 populations, maximum expression size 30, 64-bit internal precision, multithreaded parallelism, and `model_selection="accuracy"`; population size and other unlisted evolutionary settings retain PySR defaults. Binary operators are $+$, $-$, $\times$, division, and $|x|^y$. Unary operators are square, cube, $\sqrt{|x|}$, $\exp(x)$, $\log|x|$, $\sin(x)$, $\cos(x)$, and $|x|$. The exponent operator uses the PySR constraint `(-1,3)`, and `exp` uses constraint 8. Nested exponentials and nested sine/cosine compositions are disallowed. Each task's recorded batch seed is passed to PySR as its random state; seeds are task-specific rather than fixed to a common value.

B.6 Direct-Ask Prompt and Runtime

Gold-P prompt template. Gold-P replaces the frozen submission in the common mechanism-probe template with the exact reference phenomenal equation after eliminating internal variables. It also discloses named numerical-constant equations and numerical literals labelled by the mechanism quantity whose defining equation contains them. The user-message template is:

```
Frozen submitted phenomenal model:
{reference phenomenal equation}

Known named numerical constants for this task:
{name = numeric value, or (none)}

Additional numerical constant values, labelled by the mechanism
quantity whose defining equation contains them (without disclosing
the equation itself):
{quantity: numeric literals, or (none)}

Using the phenomenal model and known numerical constants above,
derive {probe name} ({probe description}) as a function of:
{variables}.

Return exactly one equation in the form:
{probe name} = <expression>
```

```
Use only the given variables, numeric literals, and supported
operators/functions. Do not add variables, equations, or prose;
do not modify the frozen phenomenal model; and do not use tools.
```

Direct-Ask execution. Each probe is a fresh one-turn, tool-free query; probe answers do not share conversational state. The GPT-5.6-sol control uses Codex with medium reasoning effort and a 300-second timeout. The GLM-5.3-flash control uses direct OpenRouter chat-completions requests with `temperature=0`, minimal reasoning effort, an 8,192-token output cap, a 600-second request timeout, and up to four transport attempts. The reported DeepSeek-v4-flash-0731 Gold-P control uses `temperature=0`, minimal reasoning effort, a 2,048-token output cap, a 180-second request timeout, and up to four transport attempts. The PySR + Direct-Ask pipeline uses the same direct-API system instruction and decoding contract as the latter, also with `temperature=0`, minimal reasoning effort, a 2,048-token output cap, a 180-second request timeout, and up to four attempts.

Direct-Ask system instruction. For direct OpenRouter requests, the system message is: "Answer the user's mechanism probe directly from the supplied frozen equation. Follow its output contract exactly: emit one plain-text equation and no reasoning, prose, Markdown, or additional equations." The Codex version appends the same one-equation contract to the user prompt and disables filesystem, network, shell, web, and other tools while retaining its native system scaffolding.

Agent failures. A run is treated as an agent-side failure when the agent execution itself is available for evaluation but its answer is unusable or scientifically wrong: examples include an empty checkpointed working model, malformed or unparsable equations, a model that cannot jointly solve the required variables, multiple or prose-bearing Direct-Ask answers, and an incorrect numerical or symbolic result. We re-run these failed cases up to five times until they succeed; runs that fail to submit a valid answer after five attempts are recorded as failures.

C Experimental Results

C.1 Complete Recovery and Numerical-Accuracy Results

Tables 3 and 12 give the complete domain-level symbolic-recovery results underlying the aggregate values in Table 1. For the two three-run Codex configurations, the rate is computed separately for each run and then averaged. As in the main text, $\mathrm{SA}(M)$ is the strict task-level measure that requires every mechanism probe for an instance to be symbolically correct.

Codex with GPT-5.6-sol has the highest overall $\mathrm{SA}(P)$ and $\mathrm{SA}(M)$ on both sets and leads mechanism recovery in each Full-set domain. The size of the phenomenal–mechanistic gap nevertheless varies substantially. On the Core-set, for example, this configuration recovers 40.00% of phenomenal laws but only 10.00% of mechanisms in biology, and recovers no complete mechanism in materials despite recovering 20.00% of phenomenal laws.

Table 13 further reports all 16 Full-set task families for the two standard discovery configurations evaluated over the Full-set. A task family that is easier for phenomenal recovery is not necessarily easier for mechanism recovery. For GPT-5.6-sol, *Inclined Rolling* has the highest phenomenal recovery (50.00%) but only 6.25% mechanism recovery, whereas *Competitive Inhibition Pulse Reaction* reaches 31.25% and 25.00%, respectively. Conversely, several families have nonzero phenomenal recovery but no completely recovered mechanism.

Finally, Table 14 reports the complete Core-set numerical results. Numerical agreement is less stringent than exact symbolic recovery: Codex with GPT-5.6-sol attains $\mathrm{Acc}_{0.1}^{\mathrm{ID}} = 55.00\%$ and $\mathrm{Acc}_{0.1}^{\mathrm{OOD}} = 43.75\%$, compared with $\mathrm{SA}(P) = 35.00\%$. The same separation is visible for the smaller models, whose overall symbolic recovery is at most 10.00% while their OOD numerical accuracy ranges from 16.25% to 25.00%. OOD accuracy is lower than ID accuracy for every reported agent configuration.

Table 12: Complete Full-set symbolic-recovery results by scientific domain.

Base model	Agent / condition	Physics		Chemistry		Biology		Materials	
		$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$
No LLM	PySR	0.78	–	0.00	–	0.00	–	0.00	–
DeepSeek-v4-flash-0731	PySR + Direct-Ask	0.78	0.78	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.6-sol	Codex	25.78	9.38	17.19	12.50	13.28	5.47	7.81	3.91
GLM-5.3-flash	Codex	10.16	3.12	0.78	0.00	2.34	2.34	3.91	1.56

Table 13: Full-set symbolic-recovery results by task family for the two standard discovery configurations evaluated on all 512 instances. Each family contains 32 instances.

Domain	Task family	Codex (GPT-5.6-sol)		Codex (GLM-5.3-flash)	
		$\mathcal{P}$	$\mathcal{M}$	$\mathcal{P}$	$\mathcal{M}$
Physics	Circular Orbit	21.88	15.63	9.38	3.12
	Hall Response	25.00	12.50	3.12	6.25
	Inclined Rolling	50.00	6.25	28.12	3.12
	Terminal Settling	6.25	3.12	0.00	0.00
Chemistry	Acid–Base Buffer Relaxation	3.12	3.12	0.00	0.00
	Competitive Inhibition Pulse Reaction	31.25	25.00	3.12	0.00
	Lindemann Pulse Reaction	25.00	15.63	0.00	0.00
	Surface Pair Catalysis	9.38	6.25	0.00	0.00
Biology	Allosteric Regulation	18.75	3.12	0.00	0.00
	Gene Repression	21.88	12.50	9.38	9.38
	Membrane Channel Flux	9.38	3.12	0.00	0.00
	Population Resource Balance	3.12	3.12	0.00	0.00
Materials	Composite Heat Transport	12.50	12.50	12.50	6.25
	Grain-Boundary Electrical Response	3.12	3.12	0.00	0.00
	Thermoelastic Heating Response	0.00	0.00	0.00	0.00
	Vacancy Diffusion	15.63	0.00	3.12	0.00

C.2 Results by Mutation Count and Variant Subset

Table 15 reports aggregate recovery after excluding the Original instance from every included task family. The resulting variant-only subsets contain 72 Core-set and 496 Full-set instances. For Codex with GPT-5.6-sol, removing the Original instances reduces $\mathrm{SA}(\mathcal{P})$ from 35.00% to 30.56% on the Core-set and from 16.02% to 14.11% on the Full-set. The corresponding strict $\mathrm{SA}(\mathcal{M})$ values decrease from 13.75% to 9.72% and from 7.81% to 6.05%, respectively. The other standard discovery configurations achieve at most 4.17% phenomenal recovery and 1.39% mechanism recovery on the Core-set variants; Codex with GLM-5.3-flash reaches 3.23% and 0.81% on the Full-set variants.

Tables 16 and 17 further stratify recovery by the number of applied mutations. For Codex with GPT-5.6-sol, Core-set $\mathrm{SA}(\mathcal{P})$ declines from 75.00% on Original instances to 75.00%, 31.25%, 25.00%, 6.25%, and 0.00% with one through five mutations. Strict $\mathrm{SA}(\mathcal{M})$ declines more sharply, from 50.00% on Original instances to 37.50% with one mutation, 6.25% with two mutations, and zero with three or more. The Full-set shows the same broad pattern: phenomenal recovery falls from 75.00% on Original instances to zero at five mutations, while mechanism recovery falls from 62.50% to zero by four mutations.

For the smaller models, most successful recoveries are concentrated among Original and single-mutation instances. On the Core-set, none of these standard discovery configurations recovers a complete mechanism with three or more mutations, and only isolated phenomenal-law successes remain beyond two mutations. The Full-set GLM-5.3-flash configuration similarly reaches only 0.63% phenomenal recovery at three mutations and no phenomenal or mechanism recovery at four or five mutations.

Table 14: Core-set numerical accuracy by scientific domain. Each entry is the percentage of tasks whose maximum relative error over the indicated split is at most 0.1. A dash denotes an unavailable OOD result.

Base model	Agent	Physics		Chemistry		Biology		Materials	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
No LLM	PySR	35.00	–	0.00	–	20.00	–	0.00	–
GPT-5.6-sol	Codex	55.00	45.00	50.00	50.00	60.00	45.00	55.00	35.00
GLM-5.3-flash	Codex (3-run avg.)	56.67	40.00	18.33	8.33	53.33	10.00	38.33	25.00
GLM-5.3-flash	Claude Code	60.00	50.00	20.00	15.00	35.00	10.00	35.00	25.00
DeepSeek-v4-flash-0731	Codex (3-run avg.)	45.00	36.67	18.33	11.67	46.67	16.67	38.33	23.33
DeepSeek-v4-flash-0731	Claude Code	35.00	20.00	25.00	25.00	30.00	15.00	35.00	25.00
DeepSeek-v4-flash-0731	DeepSeek Harness	60.00	35.00	5.00	0.00	50.00	15.00	40.00	15.00
DeepSeek-v4-pro-0813	DeepSeek Harness	45.00	25.00	10.00	10.00	35.00	15.00	45.00	30.00

Table 15: Variant-only symbolic-recovery results after excluding Original instances. Codex with GLM-5.3-flash and Codex with DeepSeek-v4-flash-0731 use three runs only on the Core-set. A dash denotes a metric or set that is not evaluated for the corresponding configuration.

Base model	Agent	Core-set variants		Full-set variants	
		$\mathrm{SA}(\mathcal{P})$	$\mathrm{SA}(\mathcal{M})$	$\mathrm{SA}(\mathcal{P})$	$\mathrm{SA}(\mathcal{M})$
No LLM	PySR	0.00	–	0.00	–
DeepSeek-v4-flash-0731	PySR + Direct-Ask	–	0.00	–	0.00
GPT-5.6-sol	Codex	30.56	9.72	14.11	6.05
GLM-5.3-flash	Codex (3-run avg. / 1 run)	2.78	0.46	3.23	0.81
GLM-5.3-flash	Claude Code	4.17	1.39	–	–
DeepSeek-v4-flash-0731	Codex (3-run avg.)	3.24	0.46	–	–
DeepSeek-v4-flash-0731	Claude Code	2.78	1.39	–	–
DeepSeek-v4-flash-0731	DeepSeek Harness	1.39	0.00	–	–
DeepSeek-v4-pro-0813	DeepSeek Harness	4.17	0.00	–	–

C.3 Statistical Comparisons

Base-model comparisons. GPT-5.6-sol significantly outperforms both flash models across phenomenal recovery, mechanism recovery, and OOD numerical accuracy. For each comparison, we first average the flash model's binary outcome within each task and then apply a paired Wilcoxon test over the 80 tasks (Wilcoxon, 1945). After Holm correction across the six model–metric comparisons (Holm, 1979), GPT-5.6-sol exceeds GLM-5.3-flash by 28.75 percentage points in $\mathrm{SA}(\mathcal{P})$, 10.83 points in $\mathrm{SA}(\mathcal{M})$, and 22.92 points in OOD $\mathrm{Acc}_{0.1}$, with adjusted $p < .001$, $p = .008$, and $p < .001$, respectively. Its corresponding advantages over DeepSeek-v4-flash-0731 are 27.08, 10.00, and 21.67 points, with the same adjusted significance levels. Thus, the performance advantage of GPT-5.6-sol is present in all three evaluated outcomes.

Base-model and harness comparisons. With Codex fixed, paired Wilcoxon comparisons between DeepSeek-v4-flash-0731 and GLM-5.3-flash detect no difference in phenomenal recovery ($p = .234$), mechanism recovery ($p = .414$), or OOD numerical accuracy ($p = .662$). Exact McNemar comparisons under Claude Code (McNemar, 1947) give the same conclusion ($p = .688$, 1.000, and .607). Holding DeepSeek-v4-flash-0731 fixed and comparing Codex, Claude Code, and DeepSeek Harness on the same 80 tasks, Cochran's Q test (Cochran, 1950) is nonsignificant for all three metrics ($p = .867$, .607, and .264). For GLM-5.3-flash, paired Wilcoxon comparisons between Codex and Claude Code are likewise nonsignificant ($p = .152$, .655, and .075). Accordingly, the present comparisons distinguish GPT-5.6-sol from the flash models, but do not establish a performance difference between the two flash models or among the evaluated harnesses.

Table 16: Core-set symbolic recovery stratified by the number of applied mutations. Mutation count zero denotes Original instances.

Base model	Agent	$\mathrm{SA}(\mathcal{P})$ by mutation count						$\mathrm{SA}(\mathcal{M})$ by mutation count					
		0	1	2	3	4	5	0	1	2	3	4	5
No LLM	PySR	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–	–	–	–
DeepSeek-v4-flash-0731	PySR + Direct-Ask	–	–	–	–	–	–	12.50	0.00	0.00	0.00	0.00	0.00
GPT-5.6-sol	Codex	75.00	75.00	31.25	25.00	6.25	0.00	50.00	37.50	6.25	0.00	0.00	0.00
GLM-5.3-flash	Codex (3-run avg.)	37.50	8.33	0.00	4.17	0.00	0.00	25.00	2.08	0.00	0.00	0.00	0.00
GLM-5.3-flash	Claude Code	50.00	12.50	6.25	0.00	0.00	0.00	25.00	0.00	6.25	0.00	0.00	0.00
DeepSeek-v4-flash-0731	Codex (3-run avg.)	50.00	8.33	6.25	0.00	0.00	0.00	33.33	2.08	0.00	0.00	0.00	0.00
DeepSeek-v4-flash-0731	Claude Code	37.50	12.50	0.00	0.00	0.00	0.00	25.00	6.25	0.00	0.00	0.00	0.00
DeepSeek-v4-flash-0731	DeepSeek Harness	50.00	6.25	0.00	0.00	0.00	0.00	25.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-v4-pro-0813	DeepSeek Harness	62.50	18.75	0.00	0.00	0.00	0.00	37.50	0.00	0.00	0.00	0.00	0.00

Table 17: Full-set symbolic recovery stratified by the number of applied mutations. Mutation count zero denotes Original instances.

Base model	Agent	$\mathrm{SA}(\mathcal{P})$ by mutation count						$\mathrm{SA}(\mathcal{M})$ by mutation count					
		0	1	2	3	4	5	0	1	2	3	4	5
No LLM	PySR	6.25	0.00	0.00	0.00	0.00	0.00	–	–	–	–	–	–
DeepSeek-v4-flash-0731	PySR + Direct-Ask	–	–	–	–	–	–	6.25	0.00	0.00	0.00	0.00	0.00
GPT-5.6-sol	Codex	75.00	40.00	15.00	8.13	1.25	0.00	62.50	25.00	5.63	0.63	0.00	0.00
GLM-5.3-flash	Codex	37.50	13.75	2.50	0.63	0.00	0.00	31.25	5.00	0.00	0.00	0.00	0.00

Domain comparisons. We find no statistically significant performance difference among physics, chemistry, biology, and materials science. We summarize each Core-set task by its mean outcome across the 11 standard-discovery runs and compare the resulting four groups of 20 tasks using the Kruskal–Wallis test (Kruskal & Wallis, 1952). The domain effect is nonsignificant for $\mathrm{SA}(\mathcal{P})$ ($p = .363$), strict $\mathrm{SA}(\mathcal{M})$ ($p = .092$), and OOD $\mathrm{Acc}_{0.1}$ ($p = .149$). Although physics has the highest descriptive mean on all three metrics, the current evidence does not support a general domain-level performance ordering; the family-level differences reported in Section 5.4 instead indicate that difficulty varies substantially within domains.

Table 18 provides the rates and p values for all comparisons discussed above. For the GPT-5.6-sol comparisons, the displayed p values are Holm-adjusted; the other entries are the corresponding two-sided paired or omnibus tests.

Table 18: Core-set statistical comparisons across base models, harnesses, and scientific domains. Rates follow the order of the groups named in each row. Codex rates for the two flash models are task-level averages over three runs except in the three-harness DeepSeek comparison, which uses the first Codex run alongside the single Claude Code and DeepSeek Harness runs.

Factor	Comparison or grouping	Metric	Rates (%)	p
Base model	GPT-5.6-sol vs GLM-5.3-flash (Codex)	$\mathrm{SA}(\mathcal{P})$	35.00 / 6.25	$< .001$
		$\mathrm{SA}(\mathcal{M})$	13.75 / 2.92	.008
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	43.75 / 20.83	$< .001$
	GPT-5.6-sol vs DeepSeek-v4-flash-0731 (Codex)	$\mathrm{SA}(\mathcal{P})$	35.00 / 7.92	$< .001$
		$\mathrm{SA}(\mathcal{M})$	13.75 / 3.75	.008
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	43.75 / 22.08	$< .001$
Base model	DeepSeek vs GLM (Codex)	$\mathrm{SA}(\mathcal{P})$	7.92 / 6.25	.234
		$\mathrm{SA}(\mathcal{M})$	3.75 / 2.92	.414
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	22.08 / 20.83	.662
	DeepSeek vs GLM (Claude Code)	$\mathrm{SA}(\mathcal{P})$	6.25 / 8.75	.688
		$\mathrm{SA}(\mathcal{M})$	3.75 / 3.75	1.000
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	21.25 / 25.00	.607
Harness	Codex / Claude Code / DSH (DeepSeek)	$\mathrm{SA}(\mathcal{P})$	5.00 / 6.25 / 6.25	.867
		$\mathrm{SA}(\mathcal{M})$	3.75 / 3.75 / 2.50	.607
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	23.75 / 21.25 / 16.25	.264
	Codex vs Claude Code (GLM)	$\mathrm{SA}(\mathcal{P})$	6.25 / 8.75	.152
		$\mathrm{SA}(\mathcal{M})$	2.92 / 3.75	.655
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	20.83 / 25.00	.075
Domain	Physics / chemistry / biology / materials	$\mathrm{SA}(\mathcal{P})$	15.9 / 10.5 / 6.4 / 6.8	.363
		$\mathrm{SA}(\mathcal{M})$	11.4 / 5.0 / 0.9 / 0.0	.092
		$\mathrm{Acc}_{0.1}^{\mathrm{OOD}}$	36.8 / 14.5 / 16.4 / 25.0	.149