SEER: LEARNING SCIENTIFIC BELIEF EVOLUTION THROUGH FALSIFICATION-DRIVEN EXPLORATION

Anonymous authors

ABSTRACT

Scientific discovery requires explanations that evolve with evidence, whereas most AI-for-Science systems optimize fixed predictive objectives. We introduce Falsification-driven Scientific Belief Learning (FSBL), a paradigm that frames discovery as the evolution of structured, falsifiable scientific beliefs. We instantiate FSBL as SEER (Scientific Exploration through Evidence and Reasoning), a neuro-symbolic system combining neural perception, knowledge-graph-grounded hypothesis generation, Bayesian soft-evidence updating, lifecycle-managed revision, and falsification-driven experiment selection. Each theory specifies a mechanism, quantitative predictions, an executable falsification condition, and a posterior belief. An expected-scientific-value planner selects experiments that discriminate among competing theories and challenge well-supported ones. We evaluate SEER across five closed domains with verifiable ground truth and one exploratory superconductivity problem. Passive coverage fails to distinguish SEER from random search, motivating active metrics for rule recovery, held-out agreement, and interventional validity. SEER recovers all 20 hidden rules in the synthetic world and achieves theory-grounded recovery across real-data domains. Its theories replicate on held-out data, whereas post-hoc rules derived from optimization baselines replicate poorly. Ablations show that removing falsification probes leaves spurious theories supported, while removing the belief system eliminates theory production. On the open problem, SEER sustains multi-generation refinement and identifies descriptor patterns consistent with known physics, with all open-domain claims explicitly scoped to surrogate-level evidence.

1 INTRODUCTION

Deep learning now delivers state-of-the-art performance across proteins, materials, and interatomic potentials (Jumper et al., 2021; Merchant et al., 2023; Zeni et al., 2025; Batatia et al., 2022), accelerating screening in domains once limited by experiment throughput (Butler et al., 2018). In parallel, LLM-based agents orchestrate tools, plan experiments, and draft research artifacts (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2024; Gottweis et al., 2025; Schmidgall et al., 2026; Bai et al., 2026). These systems excel at what will happen or which condition performs best: they learn predictive mappings, optimize objectives, or extract statistical dependencies. This success is real but bounded by a framing fixed in advance: a dataset, a target quantity, a score to maximize.

Scientific discovery asks a different question: why does a system behave as it does, and when would that explanation fail. A hypothesis is not merely a predictive model but a structured belief carrying assumptions, mechanisms, predictions, and refutation conditions. Progress in science is driven as much by eliminating incorrect beliefs as by accumulating support, a process existing computational approaches rarely model. Optimization and active learning maintain no explanatory structures (Frazier, 2018; Holland, 1975; Settles, 2009; Ren et al., 2021). Causal discovery recovers relationships without mechanisms or quantitative predictions, and its independence tests falsify graph structure rather than theory content (Spirtes et al., 2000; Shimizu et al., 2006). The strongest LLM agents inherit their hypothesis space from pretrained text rather than from an evolving, evidence-grounded belief system. Recent falsification-centred proposals acknowledge this gap without closing it (Liu et al., 2024; Takahara & Mizoguchi, 2026). The challenge is to construct, evaluate, and revise such beliefs autonomously, in an iterative process resembling scientific reasoning.

We introduce *Falsification-driven Scientific Belief Learning* (FSBL), a paradigm that formulates discovery as the evolution of structured, falsifiable beliefs. Progress then emerges from identifying

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.

when explanations fail, actively searching for discriminative observations, and refining accordingly. We implement FSBL as *SEER* (Scientific Exploration through Evidence and Reasoning), a neuro-symbolic system whose theories carry a mechanism, quantitative predictions, an executable falsification condition, a support region, and a Bayesian posterior. A lightweight neural encoder supplies perception only. Hypothesis formation, belief evolution, experiment planning, and theory revision remain symbolic and interpretable.

Our evaluation uses a positive-control methodology in two tiers. On five *closed* domains with verifiable ground truth, the standard passive metric fails a diagnostic control, because a random explorer attains identical coverage. Discovery is therefore scored with faithful *active* metrics. On an *open* problem (superconductor critical temperature, no hidden rules) we ask whether the same loop behaves scientifically. SEER recovers all 20 hidden rules of the synthetic world, emits predictions that replicate on held-out data, and survives *do*-style intervention tests. Rules extracted post-hoc from coverage-matched optimization baselines replicate poorly. Ablations localize the mechanism: without falsification probes, spurious theories become supported, and without the belief system, theory production vanishes. Autonomous discovery thus requires a paradigm beyond prediction and optimization.

Our contributions are threefold. **(1) A discovery paradigm.** FSBL formalizes autonomous scientific discovery as falsification-driven evolution of structured scientific beliefs, replacing the static hypothesis space of optimization with a grammar-bounded, evolving population of theories that accumulates evidence, faces executed falsification attempts, and undergoes lifecycle-managed revision (§3). **(2) A neuro-symbolic instantiation.** SEER integrates a causal knowledge graph, Bayesian soft-evidence updating, an ESV experiment planner whose falsification term grows with belief maturity, and auditable theory lineages that emit machine-checkable theories rather than bare predictions (§3). **(3) A faithful-metric evaluation.** We show that passive coverage cannot separate SEER from random search and that rule recovery cannot separate the full system from its no-falsification ablation. Only the belief dynamics show genuine self-correction. Benchmarks span five closed domains and one open problem, with open-domain claims scoped to surrogate-level evidence (§4).

2 Related Work

From robot scientists to LLM discovery agents. Closed-loop discovery began with the Robot Scientists, which coupled logical inference with laboratory automation (King et al., 2004; 2009), followed by mobile chemists and the autonomous A-Lab (Burger et al., 2020; Szymanski et al., 2023). LLM-based agents now plan and execute experiments natively (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2024; Gottweis et al., 2025; Swanson et al., 2025; Qu et al., 2024), and the 2026 wave adds scientific foundation models, co-scientist deployments, long-horizon research agents, and self-improving hypothesis loops (Schmidgall et al., 2026; Bai et al., 2026; Zhao et al., 2026; Smith, 2026; Qian et al., 2026). Across generations, proposals are hand-encoded or inherited from pretrained text, and none maintains hypotheses as evidence-grounded beliefs with posteriors, executed falsification conditions, and lifecycle revision. FSBL addresses this gap. Our empirical comparison covers search-based baselines, while LLM agents are discussed qualitatively (§4, §5). Closest in spirit, AIGS centres falsification (Liu et al., 2024), and the Hypothesis Evolution Protocol makes hypothesis evolution auditable (Takahara & Mizoguchi, 2026). Neither maintains a competing theory population under do-style validation with revision lineages. Virtual-cell mechanistic reasoning (Jang et al., 2026) lacks theories under experimental challenge.

AI for Science and acquisition functions. Deep learning delivers state-of-the-art prediction and design (Jumper et al., 2021; Merchant et al., 2023; Zeni et al., 2025; Batatia et al., 2022; Dunn et al., 2020; Butler et al., 2018), with acquisition functions (Bayesian optimization, active learning, evolutionary search) guiding sampling (Frazier, 2018; Settles, 2009; Ren et al., 2021; Holland, 1975). They improve a predictor under a fixed hypothesis space, answering what performs best but not why. FSBL instead selects experiments that discriminate and eliminate explanations.

Causal discovery, symbolic regression, and falsification. Causal structure learning (Spirtes et al., 2000; Shimizu et al., 2006) and symbolic regression (Brunton et al., 2016; Schmidt & Lipson, 2009; Udrescu & Tegmark, 2020; Petersen et al., 2021; Cranmer, 2023; Romera-Paredes et al., 2024; Cory-Wright et al., 2024) recover relations, and neuro-symbolic systems combine neural perception with symbolic inference (Mao et al., 2019; d'Avila Garcez & Lamb, 2023; Trinh et al., 2024; Pan et al., 2023). Equations and graphs lack mechanisms and theory-internal rejection conditions: symbolic expressions do yield testable predictions, but carry no executable criterion for revising or

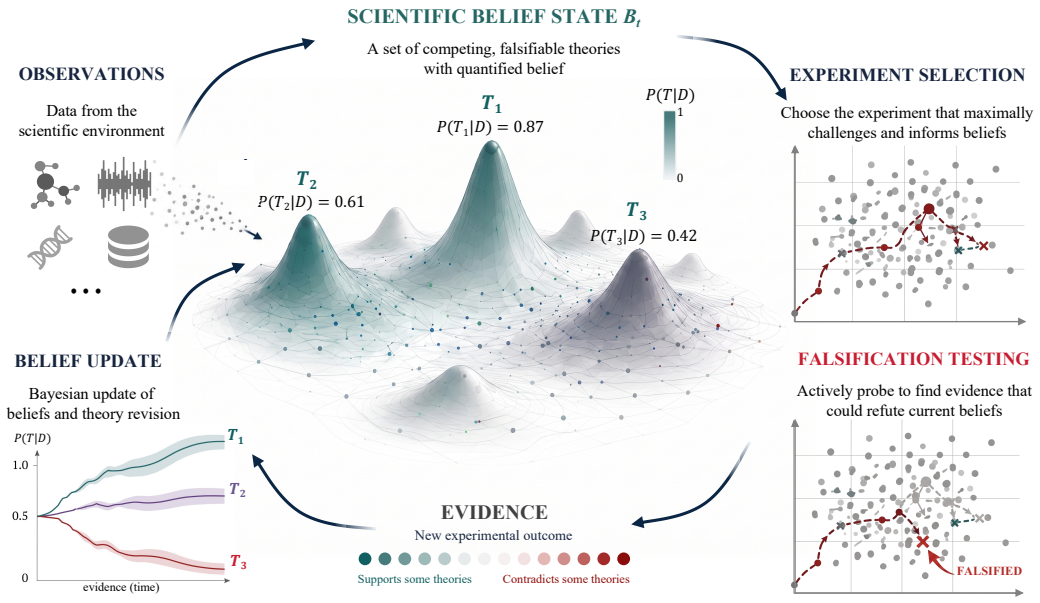

Figure 1: The FSBL loop: observations update a belief state of competing, falsifiable theories; experiment selection maximally challenges and informs them; falsification testing seeks refuting evidence; Bayesian evidence accumulation and lifecycle revision reshape the population. Inset: illustrative posteriors.

rejecting the expression itself. Falsification in these lines (Popper, 1959; Lakatos, 1978; Box, 1980) remains an after-the-fact evaluation. FSBL embeds it inside the loop, where counterexamples drive exploration and revision.

3 Method: Falsification-Driven Scientific Belief Learning

3.1 Problem Formulation

We consider a scientific environment over an experimental space $\mathcal{X}$: an intervention $x \in \mathcal{X}$ yields an observation $y = f^*(x) + \varepsilon$, where f^* is the unknown phenomenon and ε observational noise. At step t the agent maintains a *belief state* $B_t = (\mathcal{T}_t, G_t, E_t)$: a population of candidate scientific theories $\mathcal{T}_t = \{\tau_1, \ldots, \tau_n\}$, a belief graph G_t encoding support and contradiction among them, and an evidence memory E_t. The state evolves as $B_{t+1} = \mathcal{U}(B_t, x_t, y_t)$ through evidence accumulation, theory evaluation, and revision. The learning target is not a predictive function but the explanatory quality of the final B_T. Each theory is a structured, falsifiable object

$$\tau = \big(\mathcal{C}_\tau,\ M_\tau,\ \Pi_\tau,\ F_\tau,\ \mathcal{R}_\tau,\ p_\tau\big), \tag{1}$$

comprising scientific concepts $\mathcal{C}_\tau$, a mechanism M_τ, quantitative predictions Π_τ, an explicit falsification condition F_τ, a support region $\mathcal{R}_\tau$, and a posterior confidence $p_\tau = P(\tau \mid D_t)$.[1] Unlike a predictive model $y = f(x)$, a theory states why and when. Crucially, F_τ is executable by a mechanical refutation probe, and Π_τ is quantitative, so held-out data can confirm or refute it (§4). A synthetic-domain theory illustrates the object end-to-end: mechanism "Interval: $\dim_1 \in [0.69, 1.13]$ causally influences system performance"; prediction "performance rises by $\Delta \approx 0.29$"; falsification condition "sample outside $[0.69, 1.13]$: performance > 0.83 refutes this theory"; posterior $P = 0.85$ (supported). F_τ is generated by template from the mechanism type and fitted support interval, so every theory is born with an executable refutation query (Appendix C).

3.2 System Overview

Figure 1 lays out the Falsification-driven Scientific Belief Learning (FSBL) loop that SEER instantiates. Observations update a belief state B_t, a population of competing, falsifiable theories carrying quantified confidence. The loop chooses the experiment that maximally challenges and informs that state, probes for refuting evidence, and revises beliefs by Bayesian accumulation

[1] Bayesian-*inspired*: p_τ is an auditable Beta-form evidence score (§3.4), not a likelihood-based posterior.

and theory revision. SEER separates perception from reasoning: a lightweight neural encoder compresses observations, while hypothesis formation, planning, and revision are explicit symbolic operations over the belief state. Learning is thus two-level: the encoder learns perception, while the belief state is the learned model, updated by the evidence-facing rules below. What is learned is which candidate theories the evidence sustains. Appendix A (Figure 5) details the component architecture.

3.3 Scientific Belief Construction

Representation learning. A two-layer MLP encoder ($8 \to 64 \to 64 \to 16$, $z = \phi_\theta(x)$, $\sim 5{,}800$ parameters) is trained with a property-aware objective (similar outcomes map to nearby representations). It never predicts directly: it is a trained-online perceptual layer that re-weights the discriminative statistics used by theory proposal.

Symbolic abstraction and the knowledge graph. SEER analyzes relations among variables, latent concepts, and observed properties, covering proportional, inverse, threshold, interval, and interaction patterns, and organizes them into a causal knowledge graph $G = (V, E)$. The graph is an evolving *hypothesis substrate*, not a conclusion: it supplies candidate mechanisms.

Hypothesis generation and grounding. Hypothesis generation proceeds through observation mining, relation extraction, and scientific interpretation, converting statistically consistent patterns into domain-grounded theories of the form equation 1. Semantic grounding draws on a small, explicitly enumerated domain vocabulary of $\approx$ 280 injected terms, none of which encodes ground-truth answers (enumeration and zero-knowledge ablation in Appendix K).

3.4 Bayesian Belief Evolution and the Theory Lifecycle

Each theory maintains a Beta posterior $p_\tau = \alpha_\tau/(\alpha_\tau + \beta_\tau)$ over supporting and contradicting evidence. When an observation (x, y) arrives, every applicable theory receives a *soft* evidence update with consistency score $c_\tau \in [0, 1]$:

$$p_{\text{sup}} = \sigma\big(\kappa\,(c_\tau - \frac{1}{2})\big), \qquad \kappa = 5, \tag{2}$$

after which the evidence counts grow as $\alpha_\tau \leftarrow \alpha_\tau + p_{\text{sup}}$ and $\beta_\tau \leftarrow \beta_\tau + (1 - p_{\text{sup}})$. Partial evidence gradually shifts confidence, essential under noisy measurement. Updating alone, however, only rescales existing beliefs. Discovery requires modifying theory structure. This is Bayesian-*inspired* soft-evidence accumulation in Beta form, a scoring rule chosen for auditability rather than a full likelihood model.

Theories move from Proposed through Testing to Supported and eventually Accepted (Fig. 2). With evidence ratio $r_\tau = e_\tau^+/(e_\tau^+ + e_\tau^-)$ (the supporting share of accumulated evidence) and total evidence E_τ, a Supported theory is challenged when $r_\tau < 0.92$ ($E_\tau > 10$) and a Testing theory when $r_\tau < 0.65$ ($E_\tau \geq 12$). A challenged theory is then resolved each cycle by a deterministic rule,

$$\textsc{Resolve}(\tau) = \begin{cases} \textsc{refute} & r_\tau < 0.35,\ E_\tau \geq 12, \\ \textsc{reinstate} & r_\tau > 0.8,\ E_\tau > 15, \\ \textsc{refine} & \text{otherwise (mixed evidence)}, \end{cases} \tag{3}$$

where refinement re-centers the theory on its surviving support and narrows the region, spawning a child that enters the lifecycle in the parent's place. SEER applies five revision operations (confirm, challenge, refine, generalize, and reject) and records every revision, producing auditable lineages whose multi-generation chains are analysed in §4. The belief graph G_t tracks support and contradiction among competing theories, separating temporary uncertainty from genuine failure.

3.5 Falsification-Driven Planning and the Discovery Loop

SEER selects each experiment by *expected scientific value* (ESV) over the candidate set $\mathcal{C}_t$,

$$x_{t+1} = \arg\max_{x \in \mathcal{C}_t} \text{ESV}(x), \qquad \text{ESV}(x) = \alpha\,\text{KG}(x) + \beta\,\text{BU}(x) + \gamma\,\text{FV}(x), \tag{4}$$

with weights $\alpha = 0.25$, $\beta = 0.45$, $\gamma = 0.30$, fixed constants of the default planning mode rather than learned parameters. The mode schedule below adapts the planner to belief maturity, and Appendix G

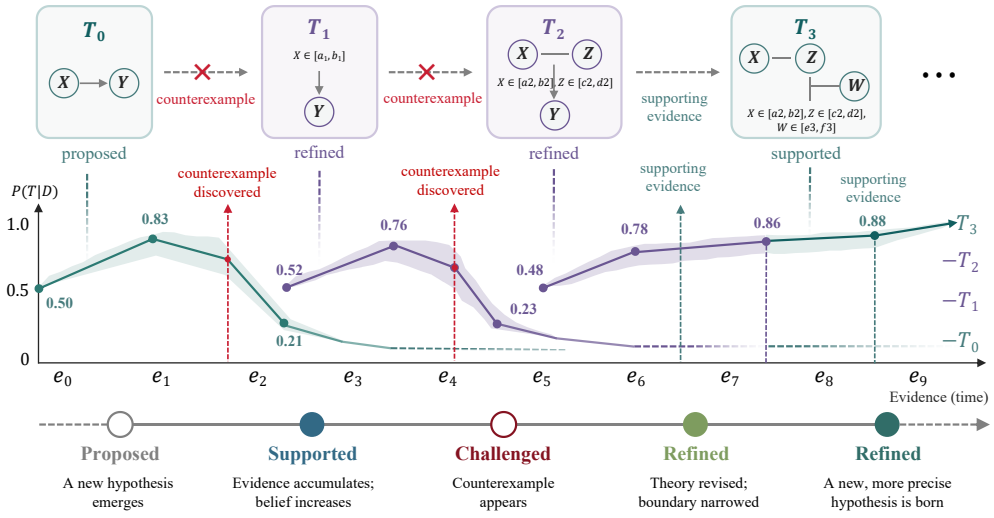

Figure 2: Theory lifecycle along the evidence axis. Top: each counterevidence event refines the theory (conditions are added or narrowed, a child replaces the parent) until supporting evidence consolidates the lineage (SUPPORTED). Bottom: posteriors are non-monotone across the PROPOSED → SUPPORTED → CHALLENGED → REFINED stages.

shows that results are insensitive to their values. The three components quantify complementary scientific value:

$$\mathrm{KG}(x) = \exp\big(-n(x)/5\big), \qquad n(x) = \big|\{x' \in D_t : \|x' - x\| < 1\}\big|, \tag{5}$$

$$\mathrm{BU}(x) = \frac{1}{|\mathcal{T}_t(x)|} \sum_{\tau \in \mathcal{T}_t(x)} u_\tau \Big(1 - \frac{d(\tau, x)}{\mathcal{R}_\tau}\Big), \tag{6}$$

$$\mathrm{FV}(x) = \frac{1}{|\mathcal{T}_t(x)|} \sum_{\tau \in \mathcal{T}_t(x)} v_\tau \Big(1 - \frac{d(\tau, x)}{\mathcal{R}_\tau}\Big), \tag{7}$$

where $\mathcal{T}_t(x) \subseteq \mathcal{T}_t$ is the set of theories whose support region contains x and $d(\tau, x)$ the distance to the theory's support center. *Knowledge gain* decays with local sampling density, preventing collapse onto explained areas. *Belief update* rewards conditions expected to move posteriors, weighted by each theory's uncertainty u_τ and proximity. *Falsification value*, the defining FSBL component, rewards experiments inside the support regions of well-supported theories, where their predictions are strongest, weighted by v_τ, how much a refutation would matter. Selected probes perturb within and around the region and execute the falsification condition mechanically. The planner follows the belief system's maturity. With fewer than three grounded theories (EXPLORE), it falls back to geometric exploration over uncertainty, novelty, and knowledge gain. With three to six (TEST), it uses Eq. equation 4. Beyond six (REFINE), it halves the knowledge-gain weight and adds $0.2\times$ a discriminative term separating competing explanations. Algorithm 1 shows the resulting loop.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL AND FAITHFUL METRICS

Domains and protocol. We evaluate on five closed domains with verifiable ground truth: (i) *controlled synthetic rule worlds* (20 dimensions, 20 hidden composition rules); (ii) *ABC_3 perovskite materials* (Materials Project (Jain et al., 2013)); (iii) *1D topological photonic nanobeams* (51 band structures); (iv) *macrophage activation* (scRNA-seq, GEO GSE103958/GSE161125); and (v) *DC Perturb-seq* (GSE90063), with literature-curated ground-truth rule sets validated on held-out data. All experiments use five seeds, 32 candidates per step, and a 500-step budget unless stated otherwise. Closed-domain tables report mean±std over seeds and hypothesis counts as per-run means (per-seed values in Appendix L). The baselines of §4.3 use a uniform 300-step query budget; LLM discovery agents need a natural-language interface these simulators lack (§5).

Algorithm 1 SEER discovery loop (one exploration run)

Require: task with bounds; budget T; candidate size k
initialise $B_0 \leftarrow (\emptyset, \emptyset, \emptyset)$; optionally pretrain ϕ_θ
for $t = 1, \ldots, T$ **do**
  propose/refresh theories $\mathcal{T}_t$ from knowledge-graph patterns and discriminative statistics
  select $x_t \leftarrow \arg\max_{x \in \mathcal{C}_t} \mathrm{ESV}(x)$ with mode-dependent weights
  observe $y_t = f^*(x_t) + \varepsilon$; $\mathcal{C}_{t+1} \leftarrow \mathcal{C}_t \cup \{(x_t, y_t)\}$
  for all theories τ applicable to x_t **do**
    soft-evidence update $(\alpha_\tau, \beta_\tau)$
  end for
  fire lifecycle transitions and revisions (refine / generalize / reject) on counterevidence
  update belief graph G_t and evidence memory E_t
end for
return final belief state B_T

Table 1: Multi-domain discovery under faithful metrics (500 steps, 5 seeds, mean±std; hypotheses = per-run mean). Coverage is passive: SEER and Random are statistically indistinguishable, motivating the active metrics.

Domain	Coverage (SEER/Random)	aRR ↑	$\mathrm{HA}_{\mathrm{strict}}$ ↑	MVR ↑	#Hyp.
Synthetic	0.85 / 0.85	**0.99±0.02**	0.49±0.11	0.55±0.14	73
Materials	0.88 / 0.88	0.38±0.08	0.67±0.31	0.69±0.32	52
Photonics	1.00 / 1.00	0.40±0.12	**0.99±0.02**	**0.99±0.02**	24
Macrophage	1.00 / 1.00	0.53±0.18	0.98±0.03	**1.00±0.00**	20
Perturb-seq	0.88 / 0.88	0.65±0.15	0.56±0.32	0.61±0.37	34

Why faithful metrics are needed. The standard passive evaluation, coverage of ground-truth regions, collapses under a diagnostic control: a random explorer attains the same coverage as SEER on every domain (Table 1), because marking a region is not inferring a rule. We therefore score discovery with three active metrics over eligible theories (alive, supported or $P > 0.6$):

$$\mathrm{aRR} = \big|\hat{\mathcal{R}}\big| \,/\, \big|\mathcal{R}^*\big|, \tag{8}$$

$$\mathrm{HA}_{\mathrm{strict}} = \frac{1}{|\mathcal{H}|} \sum_{\tau \in \mathcal{H}} \mathbf{1}\big[\bar{y}_\tau^{\mathrm{in}} > \bar{y}_\tau^{\mathrm{out}},\ p_\tau^{\mathrm{Welch}} < 0.05\big], \tag{9}$$

$$\mathrm{MVR} = \frac{1}{|\mathcal{M}|} \sum_{\tau \in \mathcal{M}} \mathbf{1}\big[\mathrm{dir.}_\tau,\ p_\tau^{\mathrm{Welch}} < 0.05\big], \tag{10}$$

where $\hat{\mathcal{R}}$ is the set of ground-truth rules structurally matched by a theory (dimension overlap plus mechanism-type or predicted-interval agreement), and $\mathcal{H}$, $\mathcal{M}$ the theories with quantitative predictions and executable mechanisms. HA and MVR probe each theory with n draws inside its predicted interval and n outside, other dimensions held at the support center (n=25 per arm for $\mathrm{HA}_{\mathrm{strict}}$, 30 for MVR, one-sided Welch test, correct direction, Cohen's d reported). Belief-health metrics (BRR, counterexample response, revision gain RG′, lineage depth TED, calibration ECE) quantify the dynamics. Exact protocols and constants: Appendix B.

Findings. (1) SEER recovers 19–20 of the 20 hidden synthetic rules (aRR 0.99 ± 0.02). (2) On real data, aRR 0.38–0.65 reflects partial but *theory-grounded* recovery. (3) $\mathrm{HA}_{\mathrm{strict}}$ separates domains by signal quality: photonic and macrophage predictions replicate almost perfectly, while materials and Perturb-seq carry larger seed variance under noisy labels (Appendix F, Fig. 9(b)). (4) MVR ≈ 1.00 on photonics and macrophage means the mechanisms are interventionally confirmed. Coverage never separates SEER from Random. The distinguishing property is not where the agent goes but what it infers on the way.

4.2 Falsification-Driven Belief Evolution

Falsification, not confirmation, drives belief correction (Fig. 3).

Refinement trades breadth for precision (Fig. 3(b)). Refinement adapts a theory's content rather than inflating confidence: four counterevidence-driven revisions shrink the support region by 88% with a non-monotone posterior. Lineages reach four generations in closed domains and five in

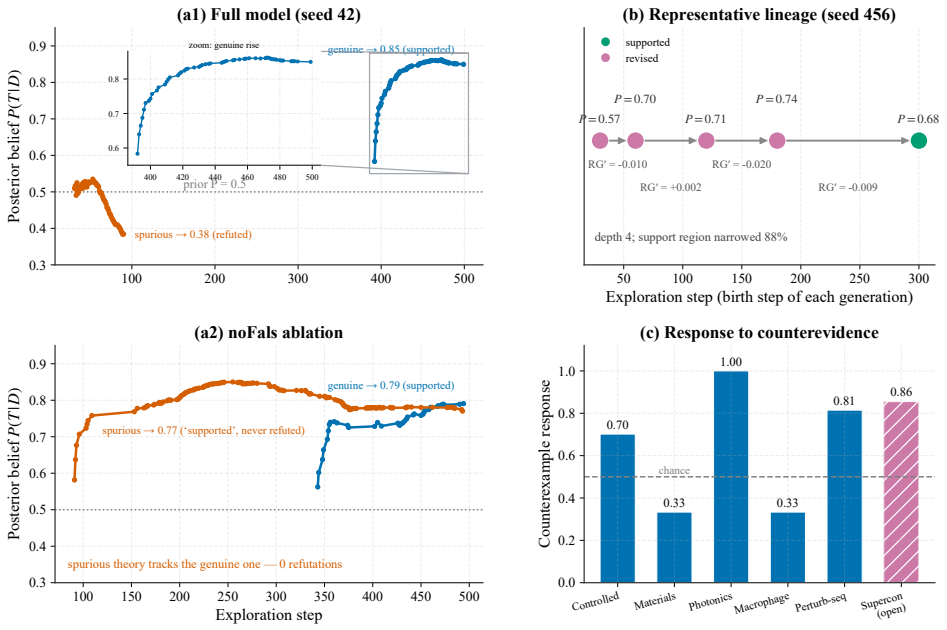

Figure 3: Falsification-driven belief evolution (synthetic domain, 500-step runs). **(a1)** With probes: the genuine theory reaches 0.85 (supported); a representative spurious theory collapses to 0.384 and is refuted (6 refutations). **(a2)** Without probes: a spurious theory ends supported at $P = 0.77$. **(b)** Lineage (seed 456): four refinements narrow the support region by 88%; posteriors non-monotone ($0.57 \to 0.71 \to 0.71 \to 0.74 \to 0.68$). **(c)** Counterexample response by domain (closed: 0.33–1.00; open superconducting domain hatched: 0.75–0.95).

Table 2: Paradigm comparison (five closed domains, uniform 300-step protocol, 5 seeds; mean±std over domain×seed runs). Optimization baselines are given a generic post-hoc rule extractor over their own query histories and scored by the same audited metrics (Appendix E.3); LiNGAM/PC emit edge strings, not rules, and are scored on causal-graph recovery (Appendix E).

Method	Coverage	aRR ↑	HA_{strict} ↑	MVR ↑	#Rules/Hyp.
SEER	0.92±0.07	**0.63±0.25**	**0.71±0.30**	0.74±0.30	**25±12**
Random	0.92±0.07	0.33±0.32	0.23±0.16	0.52±0.36	9±5
GA	0.92±0.07	0.24±0.24	0.36±0.29	0.73±0.26	8±6
BO	0.93±0.06	0.28±0.28	0.39±0.21	0.74±0.31	7±3
LiNGAM	—	0	0	0	edge strings*
PC	—	0	0	0	edge strings*

*Up to 5 hypothesis strings read off top edges; never matched to ground truth (Appendix E).

the open. Revision gains are positive in the four real-data domains and at parity in the synthetic (Appendix D).

What falsification changes at the population level (Appendix F, Table 10). Bayesian evidence accumulation alone corrects nothing: without probes, the spurious theory accumulates support in parallel with the truth and ends the run supported (Fig. 3(a2)). What changes is the composition of the belief set (Table 10). On the controlled world, probes are the only source of refutation events, reducing the spurious-supported rate from 0.28 to 0.19 and improving Perturb-seq replication from 0.42 to 0.56. On macrophage the passive lifecycle refutes more than probes would, because interaction-type ground truth is already exposed by passive evidence. The falsification contribution is thus diagnostic, not universal: it matters where passive evidence does not expose a theory's failures.

4.3 Beyond Optimization: SEER as a Discovery Paradigm

If discovery were optimization, a strong optimizer with a good coverage policy would be an adequate scientist. Table 2 shows it is not (Appendix E, Fig. 8).

Table 3: Component ablation (synthetic domain, 500 steps, 5 seeds). NOBELIEF zeroes theory production in all 5 domains $\times$ 5 seeds; full per-domain tables in Appendix F.

Configuration	aRR ↑	HA_{strict} ↑	MVR ↑	#Hyp.
full	**0.99±0.02**	**0.49±0.11**	**0.55±0.14**	**73**
NOFALS	1.00±0.00	0.53±0.10	0.57±0.09	71
NOENC	0.36±0.44	0.19±0.24	0.22±0.27	16
NOBELIEF	**0.00±0.00**	0.00±0.00	0.00±0.00	**0**

Optimization baselines match SEER's coverage (0.92–0.93) and best score (0.85–0.90 vs. 0.88 ± 0.11). Given the same post-hoc rule extractor over their query histories (Appendix E.3), they emit partial rule sets. Recovery is substantial where the ground truth is interval-type and absent where it is interaction-type. The asymmetry is by construction of the univariate extractor. A multivariate extractor over the same histories recovers interaction-type rules (Appendix E.3, Table 7: aRR 0.47–0.96 where the univariate extractor scores 0.00–0.23), matching or exceeding SEER's breadth on some domains. Replication, however, remains weak everywhere (tree HA_{strict} 0.28–0.48 vs. SEER's 0.42–0.99 under the same protocol), and knowledge production (aRR $\times$ #theories, e.g. 0.63×25 for SEER vs. 0.28×7 for BO) stays 5–8$\times$ below SEER's. In absolute recovery this is ≈ 13 of 20 rules versus ≈ 6 for BO, and the volume term in KP measures discovery output. Results are insensitive to the extractor cutoff (10–30%; Appendix E.3). Single-rule interventional validity is method-agnostic (MVR 0.73–0.74 for GA/BO). The separation lies in recovery breadth and replication, exactly why the faithful-metric hierarchy is needed. Optimization solves the bandit subproblem but does not do discovery.

4.4 Ablation, Robustness, and Scalability

Ablations (Table 3). (1) NOBELIEF zeroes theory production everywhere: no theory is proposed in any of the 5 domains $\times$ 5 seeds, so aRR is 0 and no HA_{strict} or MVR scores exist. The belief system is the only component that turns observations into theories. (2) NOFALS is the subtle failure mode. aRR is unchanged (1.00) and on real-data domains even matches or exceeds full (Appendix F, Table 8), but Fig. 3(a2) shows refutation dropping to zero and a spurious theory ending the run supported. Recovery is volume-driven and insensitive. The damage is to which theories survive: falsification buys belief-set composition and replication, not recovery volume. (3) NOENC degrades the synthetic domain (aRR $1.00 \rightarrow 0.36 \pm 0.44$, three of five seeds collapsing entirely), while real-data recovery is equal or better without the encoder. Perception is load-bearing for raw synthetic streams, whereas tabular real-data descriptors are already informative, consistent with SEER as a primarily symbolic system.

Hyperparameter sensitivity. One-at-a-time perturbation of the ESV weights (± 0.10), $\kappa \in \{2, 10\}$, and the lifecycle thresholds leaves aRR within 0.97–1.00 and HA/MVR within ± 0.07 of the defaults (Appendix G).

Noise robustness and seed stability (Appendix F, Fig. 9(a,b)). Rule recovery is stable under noise for both update rules, while MVR collapses for both: intervention significance, not rule structure, is what noise destroys.

Scalability (Appendix F, Fig. 9(c)). Across $d = 5 \rightarrow 100$ (per-seed values in Appendix H), aRR degrades gracefully to d=50 ($1.00 \rightarrow 0.77 \pm 0.06$) with stable coverage and near-constant $\sim$0.04 s per step. At d=100, recovery drops to 0.48 ± 0.15, the honest boundary of the interval-only grammar.

Held-out worlds (Appendix I). On held-out random worlds (10–30 rules, unseen structures), SEER recovers 0.95–0.97, with extractor-equipped baselines at 0.43–0.67 and replication ≤ 0.20. Three-rule composites mark the grammar boundary.

4.5 Exploratory Analysis on an Open Scientific Problem

Closed domains verify methodology against known answers. An open domain tests whether the loop behaves scientifically without ground truth. We run SEER on the UCI superconductivity dataset (Hamidieh, 2018) (21,263 compositions, 15 descriptors, target T_c, no hidden rule set) in 800-step runs across three seeds (per-seed statistics in Appendix J).

Belief dynamics without ground truth. As in the closed domains, 121–169 theories are proposed across seeds, the belief revision rate is 0.835–0.868, and refinement chains reach 3–5 generations

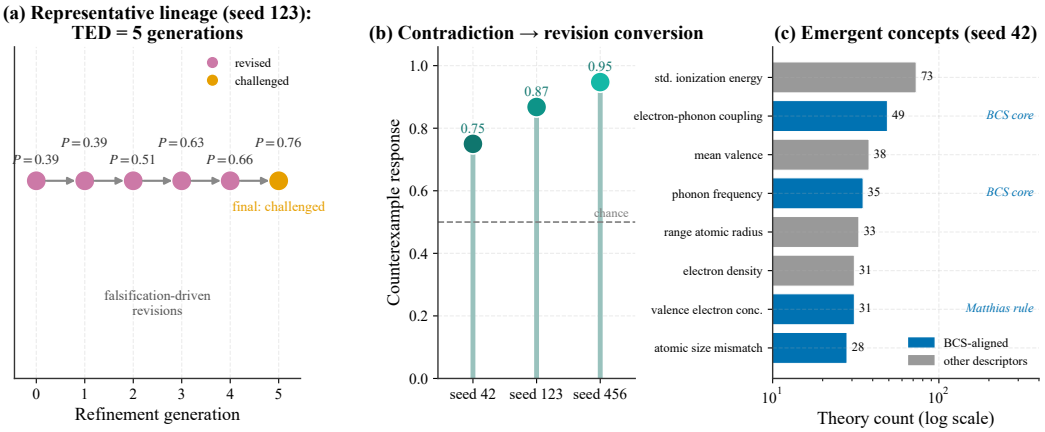

Figure 4: Open-domain belief dynamics (superconductivity, 800-step runs, 3 seeds). **(a)** Five-generation lineage (seed 123), final status challenged. **(b)** Contradiction → revision conversion across seeds (0.75–0.95). **(c)** Top-8 identified concepts (seed 42; physics-aligned set in blue). T_c is a k-NN proxy throughout; no experimental claims are made.

(Fig. 4(a,b); Table 14). Mechanism validation is harder without ground truth: MVR is 0.14–0.44 (mean 0.26) with direction accuracy 0.30–0.75, reported in full without selection.

Concept identification (Fig. 4(c)). Among the top-8 concepts by theory count, the physics-aligned set (*electron–phonon coupling*, *phonon frequency*, *valence electron concentration*, *atomic size mismatch*) appears alongside generic statistical descriptors. These terms belong to the grounding vocabulary enumerated in advance. Which of them the evidence selects is what is learned: concept identification, not invention.

Proxy-level cross-validation. A k-NN surrogate over known compositions evaluates the highest-scored candidates: every top-10 candidate's nearest neighbours lie in the Hg-1223 cuprate family ($T_c = 130$ K; mean nearest-neighbour T_c 130 K vs. 15.5 K for random compositions, an $8.4\times$ contrast). This is an indirect, surrogate-level validation. Proximity of top candidates to known high-T_c compositions is expected by construction, since exploration optimizes the k-NN surrogate; the check is not independent physical evidence. The best proxy score (T_c 122–131 K across seeds, mean 127.4 K) is likewise surrogate-evaluated.

5 Limitations and Future Directions

SEER has three main boundaries. The system is knowledge-grounded: closed-domain conclusions survive without the vocabulary, but open-domain refinement depends on it, and open-domain evidence remains surrogate-level until the loop is coupled to an experimental platform. The zero-knowledge control quantifies what is lost (Appendix K). The mechanism grammar covers intervals and pairwise interactions and degrades in high dimension, so richer, composable grammars are needed. Finally, symbolic-regression extractors over baseline histories remain untested, and the LLM-agent comparison remains qualitative. Each boundary suggests a direction: closing the loop with physical experiments, scaling the grammar (LLM-proposed hypothesis spaces complement enumerated vocabularies), comparing head-to-head with LLM agents, and accumulating beliefs across domains.

6 Conclusion

We presented FSBL, which formulates autonomous scientific discovery as the evolution of structured, falsifiable beliefs, and SEER, its neuro-symbolic instantiation. Passive coverage does not distinguish SEER from random search, and rule recovery does not separate it from the no-falsification ablation. Two observations do. First, the belief dynamics show genuine self-correction. Second, its theories replicate on held-out data better than rules extracted from optimization-based baselines. On the open problem, the same loop sustains multi-generation refinement under explicit surrogate-level scoping. These results suggest a design principle for discovery systems: make hypotheses falsifiable, and spend exploration where current beliefs are most likely to fail.

REPRODUCIBILITY STATEMENT

All results come from a single frozen code path with fixed closed-domain seeds {42, 123, 456, 789, 1024}; runs are deterministic and every table regenerates from the released per-seed records. The audit protocol (metric definitions, probe RNG separation, lifecycle constants) is specified in Appendix B, and ground-truth rule-set synthesis for all domains in Appendix B.5. The complete frozen code path and all per-seed records are provided as supplementary material.

ETHICS STATEMENT

This work targets closed benchmark domains and one public dataset; the open-domain analysis is scoped to surrogate-level evidence and makes no experimental-scientific claims. We highlight interpretability and auditable revision lineages as design commitments for autonomous discovery agents. The research involves no human subjects, no personal data, and no deployment of autonomous systems in real-world settings; in our judgement the work raises no ethical concerns.

AI USE STATEMENT

LLM-based AI tools were used under human direction for language polishing, grammar checking, and minor editorial details. All experiments, data analysis, scientific claims, and figures are the work of the authors.

REFERENCES

Lei Bai, Jiaqi Cao, Chiyu Chen, Guanzhou Chen, Kai Chen, Guangran Cheng, Erfei Cui, Xuanlang Dai, Shengyuan Ding, Shangheng Du, et al. Intern-S2-Preview: Scientific agentic foundation model. *arXiv preprint arXiv:2608.13505*, 2026.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35, 2022.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

George E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A*, 143(4):383–430, 1980.

Andres M. Bran, Sam Cox, Olivier Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Benjamin Burger, Paul M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobao Li, Ben M. Alston, Buyao Li, Rob Clowes, Rehan Sirwardana, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583:237–241, 2020.

Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559:547–555, 2018.

Ryan Cory-Wright, Cristina Cornelio, Sanjeeb Dash, Bachir El Khadir, and Lior Horesh. Evolving scientific discovery by unifying data and background knowledge with AI Hilbert. *Nature Communications*, 15:5922, 2024.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56:12387–12406, 2023.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and automatized benchmarking tools. *npj Computational Materials*, 6:138, 2020.

Peter I. Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with Co-Scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–364, 2018.

John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Yunhui Jang, Lu Zhu, Jake Fawkes, Toby Denton, Dominique Beaini, and Emmanuel Noutahi. Towards autonomous mechanistic reasoning in virtual cells. *arXiv preprint arXiv:2604.11661*, 2026.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Teresa Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Ross D. King, Kenneth E. Whelan, Fergal M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1*. Cambridge University Press, 1978.

Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. AIGS: Generating science from AI-powered automated falsification. *arXiv preprint arXiv:2411.11910*, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019.

Amil Merchant, Simon Batzner, Sam S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Brendan K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021.

Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Ziqing Qian, Jiaying Lei, Yifang Wang, and Nan Cao. Hypoforge: A self-improving multi-agent framework for automated hypothesis generation and testing via scientific skill learning. *arXiv preprint arXiv:2608.25770*, 2026.

Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C. Cousins, William A. Johnson, Xiaotong Wang, Mihir Shah, Russ B. Altman, Denny Zhou, Mengdi Wang, and Le Cong. CRISPR-GPT for agentic automation of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Samuel Schmidgall, Xiaokai Zhu, Marian Shaw, Lin Yang, Valentin Li'evin, Jingyun Yang, Yuchen Zhuang, Tim Strother, Alex Bijamov, Min Woo Sun, et al. Accelerating scientific research with Gemini in the real-world. *arXiv preprint arXiv:2608.26701*, 2026.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin–Madison*, 2009.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Travis Smith. The little scientist: LLM agent-driven discovery via the scientific method. *arXiv preprint arXiv:2608.16951*, 2026.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, et al. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Dan Guevarra, Matthew J. McDermott, Matthew C. Harden, Aaron D. Epstein, Aaron Yang, Ekin Dogus Cubuk, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624:86–91, 2023.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable AI scientists: A hypothesis evolution protocol for LLM agents. *arXiv preprint arXiv:2607.09195*, 2026.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625:476–482, 2024.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: a physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabb'e, Lixin Sun, Jake Smith, et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025.

Mingming Zhao, Jiqian Dong, Kangping Xu, Zadid Hasan, Chengrui Fan, Shan Jiang, Shuai Mao, Yating Ling, Linyi Zou, Tailin Zhou, et al. Scienceflow: A long-horizon agent for ML research, scientific discovery and beyond. *arXiv preprint arXiv:2608.14354*, 2026.

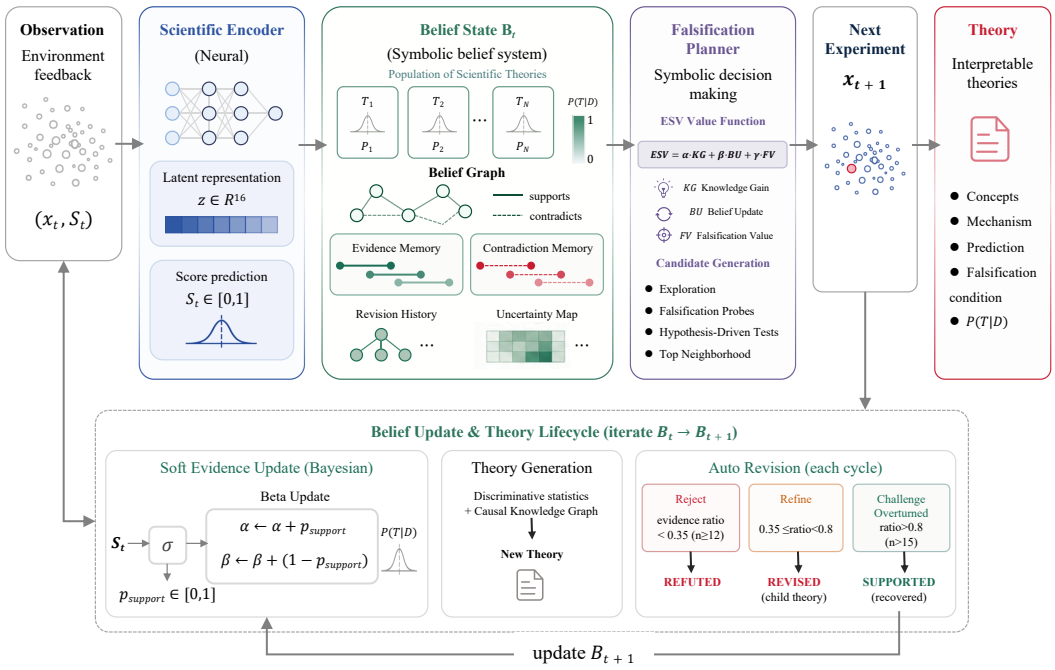

Figure 5: Detailed SEER architecture. Observation → scientific encoder (latent representation z, score prediction) → symbolic belief state B_t (population of theories with posterior, belief graph, evidence / contradiction memory, revision history, uncertainty map) → theory generation (discriminative statistics + causal knowledge graph) → falsification planner (ESV $= \alpha\,\mathrm{KG} + \beta\,\mathrm{BU} + \gamma\,\mathrm{FV}$ and candidate generation) → next experiment, with soft-evidence Bayesian updates and lifecycle auto-revision producing the next state B_{t+1}. The consistency score c_τ that drives the soft-evidence update (§3.4) is rendered p_{support} in Figure 5. Neural components (encoder) and symbolic decision making are separated to keep the emitted theories interpretable.

A Detailed System Architecture

Figure 5 expands the FSBL loop of Figure 1 into the implemented modules that instantiate it. Observations are encoded by a lightweight neural *scientific encoder* into a latent representation $z \in \mathbb{R}^{16}$ with a score prediction, feeding a *symbolic belief state* B_t that holds the population of scientific theories (each with concepts, mechanism, prediction, falsification condition, and posterior $P(\tau \mid D_t)$), a belief graph (supports/contradicts), evidence and contradiction memory, a revision history, and an uncertainty map. *Theory generation* uses discriminative statistics and a causal knowledge graph to propose new theories; the *falsification planner* scores candidate experiments by expected scientific value $\mathrm{ESV} = \alpha\,\mathrm{KG} + \beta\,\mathrm{BU} + \gamma\,\mathrm{FV}$, sampling from exploration, falsification probes, hypothesis-driven tests, and the top neighbourhood; and *soft-evidence Bayesian updates* maintain the Beta posteriors of every theory. An *auto-revision* module resolves challenged theories each cycle (reject / refine / overturn) by the evidence-ratio thresholds of §3.4.

B Metric Definitions and Protocols

All metrics are computed by a single audited metrics module on the frozen code. The constants below are the implemented values. Probe draws use a dedicated random generator (seed 42), independent of the exploration RNG, and all probing runs with the task's rule auto-marking suppressed so that metric computation cannot feed back into exploration credit.

B.1 Active Rule Recovery (aRR)

A theory is eligible if it is alive and either supported/accepted or $P > 0.6$. A ground-truth rule is *actively recovered* if some eligible theory satisfies the structural test of Algorithm 2. The interval overlap test is *relative to the smaller span*, $\mathrm{overlap}(A, B) = |A \cap B| \,/\, \min(|A|, |B|) \geq 0.05$ (not an IoU). Point-threshold rules map to the interval $(\theta, \theta + 10^3)$ and center/target rules to $(c - 2,\ c +$

2). Each rule is counted at most once (first matching theory). $\mathrm{aRR} = \frac{|\text{rules matched}|}{|\text{ground truth}|}$. Matching is deliberately structural: there is no synonym-expansion or template credit, so aRR cannot be gamed by vocabulary overlap alone.

Matcher details (Algorithm 2). The structural test credits a rule when dimension overlap holds and either the mechanism type is compatible or the predicted interval overlaps the rule constraint by at least 0.05 of the smaller span. Three implementation facts matter for reproduction. First, degenerate single-point intervals ($hi = lo$) never satisfy the numeric test ($hi \leq lo$ fails), so recovery of such theories proceeds through the dimension and mechanism-type path; the overlap denominator additionally carries a 10^{-9} span floor. Second, the mechanism-compatibility map is interval→{interval}, interaction→{interaction, product}, ratio→{ratio, ratio-range}, threshold→{threshold}, quadratic→{quadratic}, with two extensions: an interaction theory sharing at least two dimensions with a multi-dim rule is type-compatible, and any theory making a structured interval claim on the dimension of a single-dim rule is type-compatible. Third, the type-OR-interval logic is deliberate: interaction and composite rules carry no interval semantics, and replication-side guarding is done by $\mathrm{HA}_{\text{strict}}$ and MVR.

Algorithm 2 Structural theory–rule match (aRR building block)

Require: theory τ (alive, $P > 0.6$ or supported/accepted), rule ρ
1: dim-hit $\leftarrow \tau$.feature_dims $\cap$ ρ.dims $\neq \emptyset$
2: **if** $\neg$ dim-hit **then return false**
3: **end if**
4: $m_\tau \leftarrow$ InferMechanismType(τ.mechanism + τ.prediction + τ.conditions) ▷ priority: interaction > ratio > threshold > quadratic > interval > direct
5: mech-hit $\leftarrow$ m_τ compatible with ρ.type ▷ compatibility table; interaction theories sharing ≥ 2 dims with multi-dim rules also pass
6: $I_\tau \leftarrow$ structured $(\text{dim}, \text{lo}, \text{hi})$ predictions, else text-parsed intervals
7: int-hit $\leftarrow \exists\, I \in I_\tau :\ \mathrm{overlap}(I, \rho.\text{interval}) \geq 0.05$
8: **return** dim-hit $\wedge$ (mech-hit $\vee$ int-hit)

B.2 Strict Hypothesis Accuracy ($\mathrm{HA}_{\text{strict}}$)

For each eligible theory with an executable falsification condition, $n = 25$ probe points are drawn *inside* the theory's claimed interval and $n = 25$ *outside* it (outside draws fall in $[\text{low}, \text{lo}) \cup (\text{hi}, \text{high}]$ per claimed dimension, split 50/50), with all *non-claimed* dimensions fixed at the theory's support center (ceteris paribus). The prediction is verified when $\bar{y}_{\text{in}} - \bar{y}_{\text{out}} > 0$ and the one-sided Welch t-test gives $p < 0.05$. $\mathrm{HA}_{\text{strict}}$ is the verified fraction. A legacy fixed-margin variant (margin > 0.05) and a fully-random baseline variant are also recorded; the headline metric is the statistical version.

Direction convention. Interval theories encode a favorable region: the prediction asserts a score lift inside the interval, and the direction of the underlying relation only shifts where that region lies. The inside arm is therefore the predicted-favorable side by construction, and the positive-margin convention is lossless.

B.3 Mechanism Validation Rate (MVR)

MVR uses the same ceteris-paribus intervention protocol with $n = 30$ draws per arm: fix the non-claimed dimensions at the theory's support center, draw the claimed dimensions inside vs. outside the predicted interval, and score both arms. A mechanism is *validated* when the direction matches ($\bar{y}_{\text{in}} - \bar{y}_{\text{out}} > 0$) and the one-sided Welch $p < 0.05$. Effect size is reported as Cohen's $d = (\bar{y}_{\text{in}} - \bar{y}_{\text{out}})/s_{\text{pooled}}$. Because the control arm is an *outside-interval* draw under identical background conditions rather than a uniform-random arm, 0.5 is not a built-in chance level. Direction accuracy (the fraction of tested mechanisms with a positive margin) is reported alongside MVR as the chance-adjacent diagnostic.

B.4 Coverage (passive) and Belief-Quality Metrics

Coverage credits a rule when any sampled point satisfies it (*task-side* auto-marking, identical for every method): coverage $= \frac{|\text{rules marked}|}{|\text{ground truth}|}$. An adaptive grid-binning spatial coverage and a Monte-Carlo nearest-neighbour coverage are recorded as secondary diagnostics.

- **Calibration error (ECE)**: mean over theories with ≥ 5 accumulated evidence of $|P(T|D) - r|$, where $r = \frac{\text{evidence}_+}{\text{evidence}_+ + \text{evidence}_-}$ is the empirical support ratio and $P = \alpha/(\alpha + \beta)$ is the Beta posterior mean.
- **Belief revision rate (BRR)**: fraction of challenged theories that are revised (refined) or rejected.
- **Counterexample response**: over theories with substantive counterevidence (support ratio < 0.8, total evidence ≥ 12), the fraction that responded: status refuted, revised/has children, or a recorded challenge overturn.
- **Revision gain (RG′)**: for each refined child, $n = 25$ Gaussian-jittered probes around parent and child support centers. The per-child delta is mean_score(child region) − mean_score(parent region); RG′ aggregates mean, std, and positive fraction.
- **Theory evolution depth (TED)**: maximal parent→child lineage depth in the theory population.
- **Falsifiability**: fraction of alive theories carrying a prediction, an explicit falsification condition, and a support center. The *executable* subset additionally requires the falsification text to specify a concrete sampling/contrast action.

ECE here measures self-consistency between posterior and accumulated evidence rather than out-of-sample predictive calibration; the latter would require held-out labels, which the closed-domain audit does not consume.

B.5 Ground-Truth Rule Sets

Each closed domain ships an explicit, literature-curated ground-truth rule set that defines the aRR and coverage targets. The synthetic domain's 20 rules are generated by the world generator as compositions of interval, ratio, threshold, and interaction primitives. The real-data rule sets are curated from established domain knowledge and defined verbatim in the domain evaluators: perovskite stability uses the Goldschmidt tolerance-factor window (0.8–1.1) and the octahedral-factor requirement ($\mu > 0.41$) together with composition–property regularities (e.g., ionic-radius/bandgap trends); photonic rules encode band-structure targets of the 1D nanobeam families; macrophage and Perturb-seq rules encode marker-based activation programs (interaction-type regularities over cytokine and transcription-factor markers). No rule set is fit to exploration data; each is fixed before any run and shared identically by all methods. The released records include the verbatim rule lists.

B.6 Evidence Update and Planner Internals

This subsection documents the exact formulations behind §3.4 and §3.5 as implemented in the frozen module, so that the belief update and experiment planning loops can be reproduced independently.

Consistency score (c_τ). The observation score is normalized by a running min–max over the most recent 100 evaluations, $c_\tau = (s - s_{\min})/(s_{\max} - s_{\min})$; $c_\tau = 0.5$ before 10 evaluations or under a degenerate window. The same c_τ is delivered to every alive theory whose support region contains the observation, and §3.4 maps it to the soft-evidence split.

Theory uncertainty (u_τ). $u_\tau = 1/(\alpha_\tau + \beta_\tau)$, the inverse total soft-evidence mass; u_τ falls as evidence accumulates.

Refutation weight (v_τ). A status-based weight: REFUTED 0, ACCEPTED 0.1, SUPPORTED 0.5, PROPOSED/TESTING 0.6, CHALLENGED 0.9. Probing effort concentrates on theories whose standing is in question.

Discriminative term (REFINE mode). For the set of contradicting theory pairs, the term is the fraction of pairs for which x lies inside at least one of the two support regions.

Encoder objective. The two-layer encoder is pretrained online every 50 steps on the last 300 (x, y) pairs with $L = L_{\text{pull}} + 0.3\, L_{\text{push}} + 0.5\, L_{\text{reg}}$, where L_{pull} and L_{push} contrastively pull together (push apart) embeddings of similar-score (dissimilar-score) pairs and L_{reg} is the mean-squared error of the score-prediction head (100 epochs).

C Representative Theories per Domain

Table 4 lists the three highest-posterior alive theories per domain (seed 42, 500-step runs), verbatim from the run records. Every theory carries the full six-tuple of §3.1 (concepts, mechanism, quantitative prediction, falsification condition, posterior P, lifecycle status) plus its accumulated evidence and support-region radius. The theory texts are generated symbolically from the discovered

structure (feature dims, intervals, interaction terms); the same text is what the falsification planner parses to execute the refutation probe, so prediction and falsification conditions are machine-checkable, not decorative.

Two formatting details are worth noting. First, several interaction/interval theories show a degenerate single-point interval such as $[1.00, 1.00]$ or $[0.00, 0.00]$. These arise because the raw descriptors are normalized and the theory's support region collapsed onto a single observation (or a saturated feature bound) during refinement. Such theories still satisfy the structural match and remain eligible for the active metrics of §4.1, but the narrowed interval reflects an over-specialized fit rather than a broadly supported law. Second, a minority of theories (e.g., the "jointly tuned" entries in Materials and Perturb-seq) carry an interaction mechanism with a qualitative prediction and a non-executable falsification condition. They are retained verbatim as the highest-posterior representatives. Because they lack quantitative predictions or an executable probe, they are excluded from the $\text{HA}_{\text{strict}}$ and MVR counts of §4.1.

Table 4: Representative discovered theories: the three highest-posterior alive theories per domain (seed 42, 500-step runs; verbatim from the released run records). Ev. = accumulated supporting/contradicting evidence weight; r = support-region radius (standardised units); rev = revisions applied. Duplicate texts with distinct ids are separate proposals that converged on the same region.

Theory	**Six-tuple (concepts, mechanism, prediction, falsification, support region, posterior)**
Synthetic T74	$P = 0.85$ (supported); evidence 40/5; region radius 0.279; 3 revisions
Concepts	causal_influence, stability
Mechanism	Interval: dim_1 in [0.69, 1.13] causally influences system performance
Prediction	Expected system performance rises by $\Delta\approx$0.29 when dim_1 $\in$ [0.69, 1.13]
Falsification	Sample outside [0.69, 1.13]: system performance $>$ 0.83 refutes this theory
T75	$P = 0.85$ (supported); evidence 40/5; region radius 0.279; 3 revisions
Concepts	causal_influence, stability
Mechanism	Interval: dim_1 in [0.69, 1.13] causally influences system performance
Prediction	Expected system performance rises by $\Delta\approx$0.29 when dim_1 $\in$ [0.69, 1.13]
Falsification	Sample outside [0.69, 1.13]: system performance $>$ 0.83 refutes this theory
T42	$P = 0.85$ (supported); evidence 122/18; region radius 1.907; 7 revisions
Concepts	dim_0, stability, performance, high_value
Mechanism	Interval: dim_0 in [1.82, 4.83] promotes high system performance
Prediction	Expected system performance rises by $\Delta\approx$0.26 when dim_0 $\in$ [1.82, 4.83]
Falsification	Sample outside [1.82, 4.83]: system performance $>$ 0.82 refutes this theory
Materials T19	$P = 0.81$ (supported); evidence 194/21; region radius 0.001; 9 revisions
Concepts	causal_influence, lattice_stability, formation_energy
Mechanism	Interval: ionic_radius_A in [1.00, 1.00] causally influences structural stability
Prediction	Expected structural stability rises by $\Delta\approx$0.69 when ionic_radius_A $\in$ [1.00, 1.00]
Falsification	Sample outside [1.00, 1.00]: structural stability $>$ 0.55 refutes this theory
T20	$P = 0.81$ (supported); evidence 194/21; region radius 0.001; 9 revisions
Concepts	causal_influence, lattice_stability, formation_energy
Mechanism	Interval: ionic_radius_A in [1.00, 1.00] causally influences structural stability
Prediction	Expected structural stability rises by $\Delta\approx$0.69 when ionic_radius_A $\in$ [1.00, 1.00]
Falsification	Sample outside [1.00, 1.00]: structural stability $>$ 0.55 refutes this theory
T21	$P = 0.81$ (supported); evidence 194/21; region radius 0.070; 9 revisions
Concepts	ionic_radius_B, ionic_radius_A, high_value, structural_stability
Mechanism	Interaction: ionic_radius_A$\in$[1.00, 1.00] AND ionic_radius_B$\in$[0.00, 0.09] jointly promotes high structural stability
Prediction	Expected structural stability rises by $\Delta\approx$0.69 when ionic_radius_A$\in$[1.00, 1.00] AND ionic_radius_B$\in$[0.00, 0.09]
Falsification	Sample outside ionic_radius_A$\in$[1.00, 1.00] AND ionic_radius_B$\in$[0.00, 0.09]: structural stability $>$ 0.55 refutes this theory
Photonics T62	$P = 0.89$ (challenged); evidence 52/5; region radius 0.069; 3 revisions
Concepts	filling_ratio, asymmetry, high_value, optical_performance

Table 4 continued

Theory	Six-tuple
Mechanism	Interaction: asymmetry∈[0.11, 0.19] AND filling_ratio∈[0.43, 0.54] jointly promotes high optical performance
Prediction	Expected optical performance rises by $\Delta\approx0.68$ when asymmetry∈[0.11, 0.19] AND filling_ratio∈[0.43, 0.54]
Falsification	Sample outside asymmetry∈[0.11, 0.19] AND filling_ratio∈[0.43, 0.54]: optical performance > 0.56 refutes this theory
T61	$P = 0.88$ (challenged); evidence 52/5; region radius 0.009; 3 revisions
Concepts	causal_influence, mode_coupling, bandgap
Mechanism	Interval: thickness in [0.28, 0.29] causally influences optical performance
Prediction	Expected optical performance rises by $\Delta\approx0.68$ when thickness $\in$ [0.28, 0.29]
Falsification	Sample outside [0.28, 0.29]: optical performance > 0.56 refutes this theory
T50	$P = 0.87$ (challenged); evidence 54/6; region radius 0.010; 3 revisions
Concepts	causal_influence, mode_coupling, bandgap
Mechanism	Interval: thickness in [0.28, 0.29] causally influences optical performance
Prediction	Expected optical performance rises by $\Delta\approx0.68$ when thickness $\in$ [0.28, 0.29]
Falsification	Sample outside [0.28, 0.29]: optical performance > 0.56 refutes this theory
Macrophage T10	$P = 0.78$ (supported); evidence 64/5; region radius 0.011; 1 revision
Concepts	IFNγ, treatment_time, high_value, immune_activation
Mechanism	Interaction: treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70] jointly promotes high immune activation
Prediction	Expected immune activation rises by $\Delta\approx0.60$ when treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]
Falsification	Sample outside treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]: immune activation > 0.64 refutes this theory
T14	$P = 0.72$ (supported); evidence 36/3; region radius 0.008; 1 revision
Concepts	IFNγ, treatment_time, high_value, immune_activation
Mechanism	Interaction: treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70] jointly promotes high immune activation
Prediction	Expected immune activation rises by $\Delta\approx0.60$ when treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]
Falsification	Sample outside treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]: immune activation > 0.64 refutes this theory
T19	$P = 0.59$ (testing); evidence 8/1; region radius 0.005; 0 revisions
Concepts	IFNγ, treatment_time, high_value, immune_activation
Mechanism	Interaction: treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70] jointly promotes high immune activation
Prediction	Expected immune activation rises by $\Delta\approx0.60$ when treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]
Falsification	Sample outside treatment_time∈[1.00, 1.00] AND IFNγ∈[0.68, 0.70]: immune activation > 0.64 refutes this theory
Perturb-seq T19	$P = 0.91$ (supported); evidence 100/8; region radius 0.033; 1 revision
Concepts	NFKB1, ATF2, high_value, gene_regulatory_response
Mechanism	Interaction: ATF2∈[0.48, 0.53] AND NFKB1∈[0.00, 0.00] jointly promotes high gene regulatory response
Prediction	Expected gene regulatory response rises by $\Delta\approx0.37$ when ATF2∈[0.48, 0.53] AND NFKB1∈[0.00, 0.00]
Falsification	Sample outside ATF2∈[0.48, 0.53] AND NFKB1∈[0.00, 0.00]: gene regulatory response > 0.45 refutes this theory
T15	$P = 0.90$ (supported); evidence 64/6; region radius 0.027; 10 revisions
Concepts	NFKB1, STAT1, high_value, gene_regulatory_response
Mechanism	Interaction: STAT1∈[0.07, 0.11] AND NFKB1∈[0.00, 0.00] jointly promotes high gene regulatory response
Prediction	Expected gene regulatory response rises by $\Delta\approx0.37$ when STAT1∈[0.07, 0.11] AND NFKB1∈[0.00, 0.00]

Table 4 continued

Theory	Six-tuple
Falsification	Sample outside STAT1∈[0.07, 0.11] AND NFKB1∈[0.00, 0.00]: gene regulatory response $>$ 0.45 refutes this theory
T9	$P = 0.89$ (supported); evidence 180/17; region radius 0.184; 12 revisions
Concepts	causal_influence, transcription, pathway
Mechanism	CEBPB $\times$ JUN interaction modulates gene regulatory response
Prediction	Optimal gene regulatory response when CEBPB and JUN are jointly tuned
Falsification	Find region where CEBPB varies but gene regulatory response does not

D BELIEF-EVOLUTION SUPPLEMENTARY RECORDS

Table 5: Per-domain belief-health summary (500-step protocol, 5 seeds; TED = maximal lineage depth over seeds; ECE from the 300-step records, 3 seeds). Response, RG′ and TED are reproduced from the unified, encoder-integrated audit records (the same per-seed source as Table 1) under the frozen code; ECE is taken from the 300-step records released with the code.

Domain	Response	RG′	$\text{TED}_{\max}$	Hypotheses/run	ECE
Synthetic	0.76	−0.001	4	73	0.070
Materials	0.20	+0.003	4	52	0.055
Photonics	1.00	+0.006	4	24	0.108
Macrophage	0.60	+0.001	5	20	0.115
Perturb-seq	0.81	+0.002	4	34	0.052

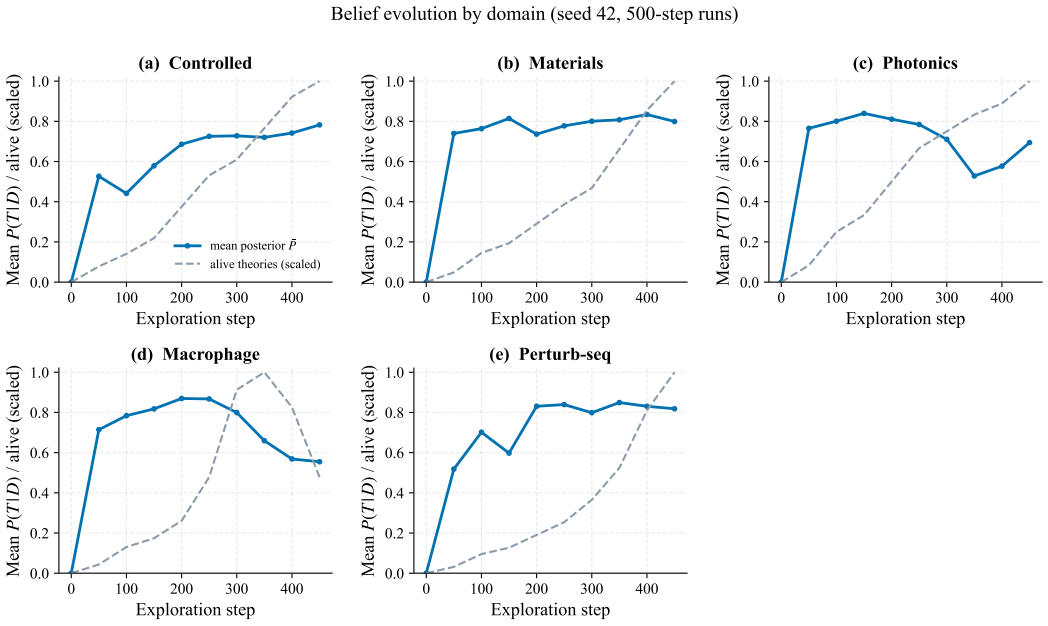

Figure 6: Belief evolution by domain (seed 42, 500-step runs): mean posterior $P(T|D)$ over alive theories (blue, left axis) and alive-theory count (grey dashed, right axis). The synthetic domain converges to a small confident population. The real-data domains sustain larger, mid-confidence populations throughout. Source: the released per-run records.

Extended standalone figures (posterior/lifecycle trajectories, ESV component decomposition, the calibration curve P vs. observed support ratio, counterfactual validity with confidence separation) accompany the release. Calibration errors are small (0.05–0.12) given that beliefs are Beta posteriors over noisy, adversarially probed evidence.

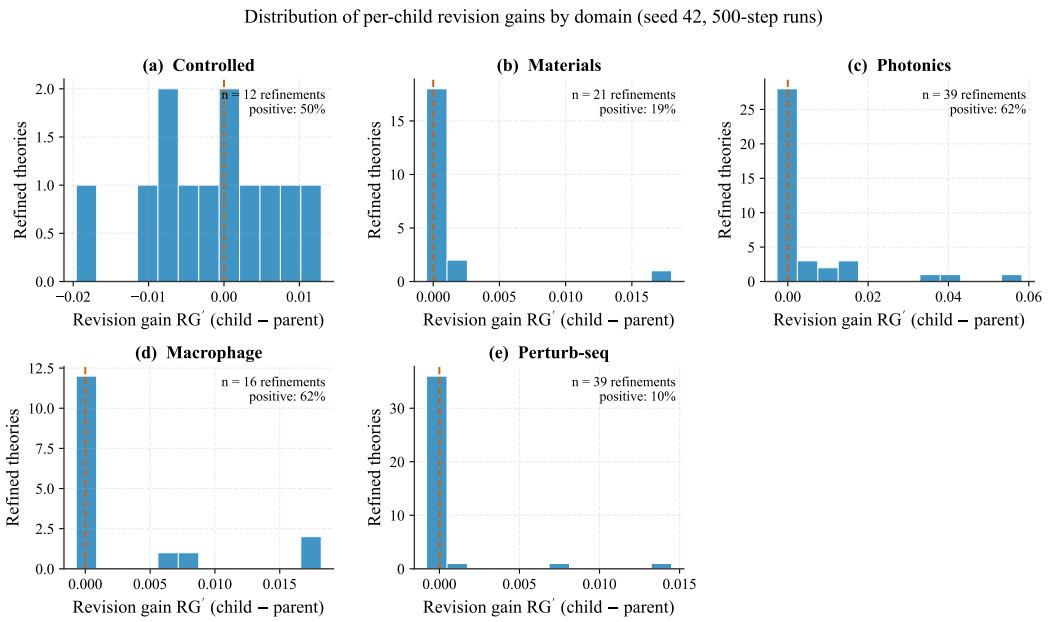

Figure 7: Distribution of per-child revision gains RG′ by domain (seed 42). Refinement gains are positive for the majority of children in the real-data domains. In the synthetic domain, gains cluster at parity (≈ 0), as expected where rule recovery is already at ceiling. Red dashed line: zero gain. Source: per-child records released with the code.

E Baseline Configuration and Causal-Discovery Details

E.1 Optimization baselines

All baselines consume one evaluation per step (SEER evaluates 32 candidates per step and selects one). The comparison protocol gives every method the same 300-step budget.

- **Random**: uniform sampling over the task bounds.
- **GA**: population 20, tournament selection (size 3), elitism 1, blend crossover BLX-α=0.5 with rate 0.9, Gaussian mutation (σ=0.5, per-gene rate 0.2), clipped to bounds (standard generational GA).
- **BO**: sklearn GP surrogate $c \times \mathrm{RBF}(1.0) + \mathrm{White}(10^{-4})$ with normalized targets and 5 optimizer restarts; Expected Improvement ($\xi = 0.01$) maximized by L-BFGS-B with 10 restarts; 10 initial random design points (standard sklearn GP).

E.2 Causal-discovery baselines

Both are implemented in the released baseline-comparison module and consume the same uniform sample budget as the optimization baselines (Table 6 reports their outcomes).

- **LiNGAM**: `DirectLiNGAM`, single deterministic fit (no bootstrap); edges kept at |coefficient| > 0.1.
- **PC**: a deliberately conservative PC-style skeleton: edges are kept at $|\mathrm{corr}(i,j)| > 0.25$ and removed when any conditioning variable explains the pair ($|\partial\mathrm{corr}(i,j|k)| < 0.3\,|\mathrm{corr}(i,j)|$; partial correlations via linear-residual regression). This is not a Fisher-Z implementation. It prunes aggressively and, on these datasets, retains no edges in any run. A standard PC skeleton with Fisher-Z conditional-independence tests ($\alpha = 0.05$, level ≤ 3) run post hoc under the same uniform budget behaves the same way: causal F1 $= 0.0$ on all five domains (Appendix E.3), so the null result is not an artifact of the conservative variant.
- **Ground-truth graph**: for each multi-dimensional ground-truth rule, a complete bidirectional subgraph over the rule's dimensions. **SHD** is the elementwise L_1 difference (a correctly paired edge with wrong direction counts 2); **causal F1** counts a predicted edge as a true positive if either direction matches.

Table 6: Causal-discovery baselines vs. SEER on graph and theory output (300-step protocol, 5 domains × 3 seeds; ranges over runs). PC emitted zero edges in every run: its pruning rule removes all candidate edges on these datasets, so its SHD reduces to the ground-truth graph size (where that is 0, SHD = 0 trivially). Source: per-run records released with the code.

Method	SHD	Causal F1	Edges	Theory output
LiNGAM	3.0 – 60.0 (median 4)	0.000 – 0.250	0 – 8	edge strings only
PC	0.0 – 58.0 (median 4)	0.000	0	edge strings only
SEER	—	—	—	structured theories (Table 1)

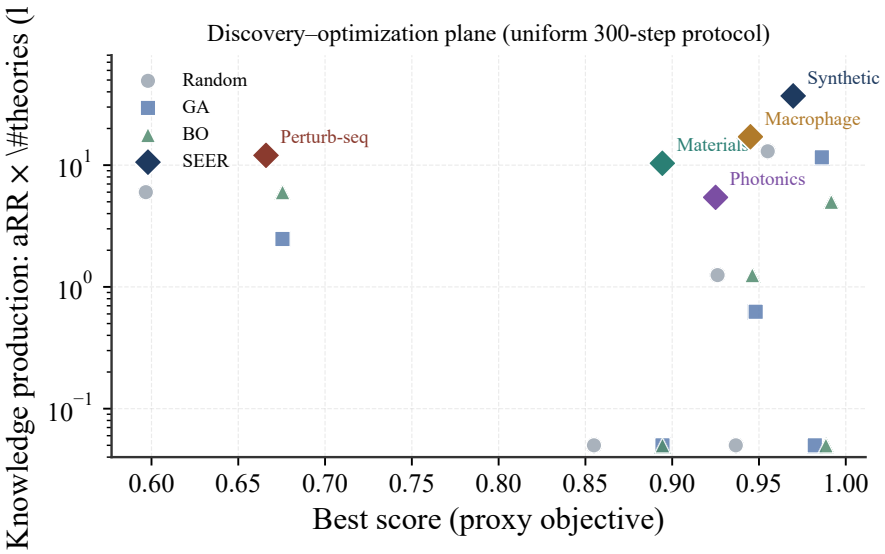

Figure 8: The discovery–optimization plane (uniform 300-step protocol, 5-seed means). Knowledge production = aRR × #theories (log scale); the volume term is deliberate, measuring discovery output rather than recovery alone. SEER (diamonds, one per domain) occupies the high-KP region (KP 5–37); Random/GA/BO reach equal or better best scores at 5–8× lower knowledge production; LiNGAM/PC are omitted (edge strings, no interval rules). Per-method numbers in Table 2. Source: released per-run records.

E.3 Rule-emitting baselines

Random, GA and BO maintain no belief system and emit no theories natively. To make the paradigm comparison non-trivial, each baseline is given a generic post-hoc rule extractor over its own query history: the top-20% highest-scored samples define, per dimension, one candidate interval rule (the 10th–90th percentile band of the positive samples on that dimension). Rules are ranked by in-sample lift, and at most one per dimension is kept, so the rule budget is capped by dimensionality (≤ 20 rules, far below SEER's typical hypothesis counts). Extracted rules carry the same structured fields as SEER theories and are scored by the same audited metrics module (structural match for aRR; held-out replication for $\text{HA}_{\text{strict}}$; do-style intervention for MVR); their coverage uses the identical task-side marking as SEER's. Extraction recovers partial structure where ground truth is interval-type (synthetic aRR up to 0.65; Perturb-seq 0.45–0.75) and fails where it is interaction-type (materials and photonics aRR = 0). Replication is weak throughout ($\text{HA}_{\text{strict}} \leq 0.39$ vs. SEER's 0.71). The cutoff is not load-bearing: re-extracting rules from the same query histories at 10% and 30% positive fractions leaves per-domain aRR essentially flat (median range 0.00, maximum 0.22 for GA on Perturb-seq), never changes the interaction-type zeros, and leaves replication weak at every cutoff. The 20% column reproduces Table 2 exactly (domain-mean aRR 0.33/0.24/0.28).

Multivariate extractor (decision tree). The univariate extractor above caps recovery at one interval rule per dimension, so its interaction-type zeros are structural. To test whether richer post-hoc extraction closes the gap, we fit a depth-4 regression tree to each baseline's query history and convert every leaf into a multi-dimension interval-conjunction rule (rule budget ≤ 16, no pre-filtering), scored by the same audited metrics under the same 300-step protocol, 5 seeds, and identical histories for both extractors. Table 7 reports per-domain means over Random/GA/BO. The

multivariate extractor recovers interaction-type rules wherever the univariate one scores 0 (aRR 0.47 on Materials, 0.51 on Photonics, 0.96 on Macrophage) and even exceeds SEER's breadth on some domains. Replication does not follow: tree rules replicate at HA_{strict} 0.28–0.48 across domains, against SEER's 0.42–0.99 under the same protocol, and only SEER's theories carry executed falsification conditions, posteriors, and revision lineages. Breadth, in other words, is available post hoc; validated, auditable theory structure is what the belief system adds. Released records include all extracted rules.

Table 7: Univariate vs. multivariate post-hoc extraction over the same baseline query histories (300-step protocol, 5 seeds, means over Random/GA/BO), with SEER under the same protocol for reference (audited metrics; released per-run records).

Domain	SEER		1D	Tree	
	aRR	HA	aRR	aRR	HA
Synthetic	0.94	0.42	0.55	0.68	0.36
Materials	0.35	0.68	0.00	0.47	0.41
Photonics	0.40	0.97	0.00	0.51	0.30
Macrophage	0.72	0.99	0.23	0.96	0.28
Perturb-seq	0.72	0.51	0.65	0.61	0.48

LiNGAM/PC emit edge strings rather than rules and are evaluated on causal-graph recovery instead.

Why no LLM-agent baselines. The closed-domain simulators expose a numeric query interface but no natural-language action space. Coupling an LLM discovery agent would require an environment wrapper whose design choices could confound the comparison, so LLM agents are discussed qualitatively (§5); community discovery environments built for end-to-end scientific reasoning are a complementary future testbed.

F Full Ablation and Noise Robustness

This appendix reports per-domain ablation results (Table 8), population-level falsification statistics (Table 10), and update-rule robustness under observation noise (Table 9, Fig. 9). It also includes the noKG control that isolates the knowledge-gain term of the planner (§3.5).

ESV knowledge-gain ablation (closed domains). The separate noKG variant removes the knowledge-gain term $\alpha\,\mathrm{KG}$ from ESV (not the vocabulary). On the synthetic domain (300-step records) this halves refinement activity (BRR $0.73 \to 0.43$, refined theories $38 \to 20$) while leaving rule coverage intact. The KG term steers where the explorer looks, not whether rules can be found.

Table 8: Ablation across all five closed domains (500 steps, 5 seeds, aRR mean±std; noEnc from a dedicated encoder-ablation suite at the same protocol). noBelief is zero in every domain and seed. Note that noFals matches or exceeds the full system's aRR on three of the five domains: recovery is volume-driven, so rule matching is insensitive to the damage, which appears in the composition of the belief set (Table 10) rather than in aRR. This is why the falsification claim is carried by the belief-dynamics evidence, not by rule matching. Source: released per-seed records.

Domain	full	noFals	noEnc	noBelief
Synthetic	0.99±0.02	1.00±0.00	0.36±0.44	0.00±0.00
Materials	0.38±0.08	0.42±0.17	0.50±0.16	0.00±0.00
Photonics	0.40±0.12	0.30±0.17	0.40±0.20	0.00±0.00
Macrophage	0.53±0.18	0.72±0.20	0.65±0.29	0.00±0.00
Perturb-seq	0.65±0.15	0.78±0.09	0.80±0.06	0.00±0.00

G Hyperparameter Sensitivity

One-at-a-time perturbation of the three hand-set constant families, on the synthetic domain (500 steps, 5 seeds, mean±std over seeds). The default row of Table 3 (aRR 0.99 ± 0.02, HA_{strict} 0.49 ± 0.11, MVR 0.55 ± 0.14) is the reference: every perturbation stays within 0.97–1.00 aRR and within ± 0.07 on HA and MVR, except that challenging earlier (threshold 0.85) increases the theory budget (107 vs. 73 hypotheses) without changing hypothesis quality. The soft-evidence slope κ has no measurable effect on end-state metrics. The pipeline's conclusions are therefore insensitive to these constants. Table 11 reports all perturbations.

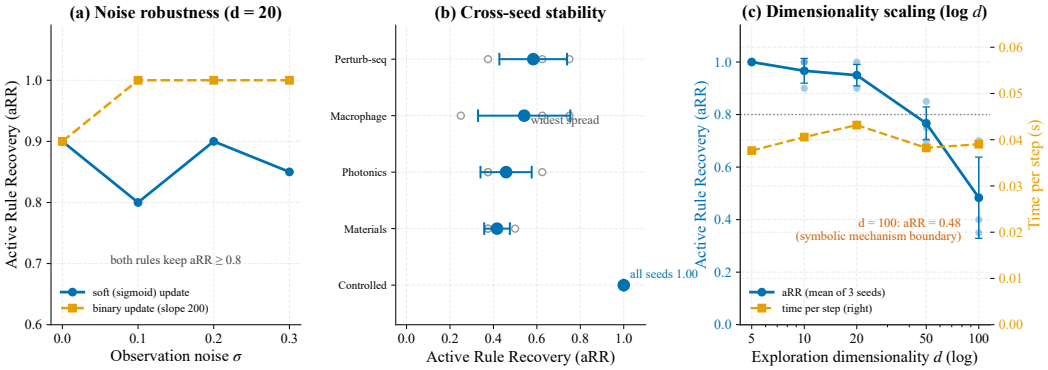

Figure 9: Robustness, stability, and scalability. **(a)** aRR under observation noise for soft vs. near-binary updates (both ≥ 0.8); MVR collapses with noise for both. **(b)** Cross-seed aRR (dots = seeds, bars = mean $\pm$ std). **(c)** Scaling: aRR degrades gracefully to d=50 and reaches 0.48 ± 0.15 at d=100; coverage and per-step time stay stable.

Table 9: Noise robustness of the update rule (synthetic domain, $d = 20$, 500 steps, seed 42). *Soft* is the shipped sigmoid evidence update (slope 5); *binary* pins the slope to 200 at every evidence update (near-hard assignments). Both rules keep aRR at 0.8–1.0 up to $\sigma = 0.3$; MVR collapses under noise for both, as intervention significance, not rule structure, is what observation noise destroys. We found no meaningful robustness advantage for the soft rule on aRR. The single "Coverage (soft)" row reports coverage at the four noise levels $\sigma \in \{0.0, 0.1, 0.2, 0.3\}$ of the aRR block (coverage is MVR-agnostic, so one row suffices). Source: released audit records.

	aRR				MVR			
σ	0.0	0.1	0.2	0.3	0.0	0.1	0.2	0.3
Soft	0.90	0.80	0.90	0.85	0.56	0.07	0.03	0.12
Binary	0.90	1.00	1.00	1.00	0.52	0.15	0.05	0.06
Coverage (soft)	0.90	0.90	0.90	0.90				

Table 10: Population-level falsification statistics (500-step runs, 5 seeds; totals over seeds). Spur. rate = fraction of supported theories that match no ground-truth rule. Probes are the sole refutation source on the synthetic world and cut its spurious-supported rate by a third; recovery itself is volume-driven and insensitive (Table 3). On macrophage the passive lifecycle refutes more without probes and the composition comparison favours NOFALS. Both observations are reported without selection.

	Refut. events		Spur. rate		HA_{strict}	
Domain	full	NOFALS	full	NOFALS	full	NOFALS
Synthetic	6	0	**0.19**	0.28	0.49	0.53
Materials	0	0	0.84	0.89	0.67	0.65
Photonics	8	9	0.69	0.69	0.99	0.97
Macrophage	39	68	0.12	0.04	0.98	0.98
Perturb-seq	7	3	0.40	0.27	**0.56**	0.42

Table 11: Hyperparameter sensitivity (synthetic, 500 steps, 5 seeds, mean$\pm$std). Defaults: α=0.25, β=0.45, γ=0.30, κ=5, challenge 0.92, refute 0.35, reinstate 0.8.

Perturbation	aRR ↑	HA_{strict} ↑	MVR ↑
$\alpha = 0.15$ / $\alpha = 0.35$	0.97±0.04 / 1.00±0.00	0.52 / 0.47	0.54 / 0.52
$\beta = 0.35$ / $\beta = 0.55$	0.97±0.06 / 1.00±0.00	0.48 / 0.51	0.49 / 0.56
$\gamma = 0.20$ / $\gamma = 0.40$	1.00±0.00 / 0.99±0.02	0.47 / 0.53	0.49 / 0.56
$\kappa = 2$ / $\kappa = 10$	0.99±0.02 / 0.99±0.02	0.49 / 0.49	0.55 / 0.55
challenge 0.85 / 0.95	1.00±0.00 / 0.99±0.02	0.46 / 0.49	0.50 / 0.55
refute 0.25 / 0.45	0.99±0.02 / 0.97±0.04	0.49 / 0.49	0.54 / 0.53
reinstate 0.70 / 0.90	0.99±0.02 / 0.99±0.02	0.55 / 0.48	0.59 / 0.55

H SCALABILITY

Table 12: Scaling with exploration dimensionality (synthetic composite worlds, 300 steps, 3 seeds per d, mean±std over seeds; per-seed dots in Fig. 9(c) and detail curves in Fig. 10). aRR degrades gracefully to d=50; at d=100 rule recovery drops to 0.48 ± 0.15, the honest boundary of the current single-theory interval mechanism in very high dimension, while coverage stays at 0.8–0.9 and wall-clock per step is near-constant. Source: released audit records.

d	5	10	20	50	100
aRR	1.00±0.00	0.97±0.05	0.95±0.04	0.77±0.06	0.48±0.15
Coverage	0.80	0.83	0.85	0.87	0.82
MVR	0.67	0.56	0.62	0.37	0.24
Hypotheses ($P > 0.6$)	43	34	40	42	45
Time/step (s)	0.038	0.041	0.043	0.038	0.039

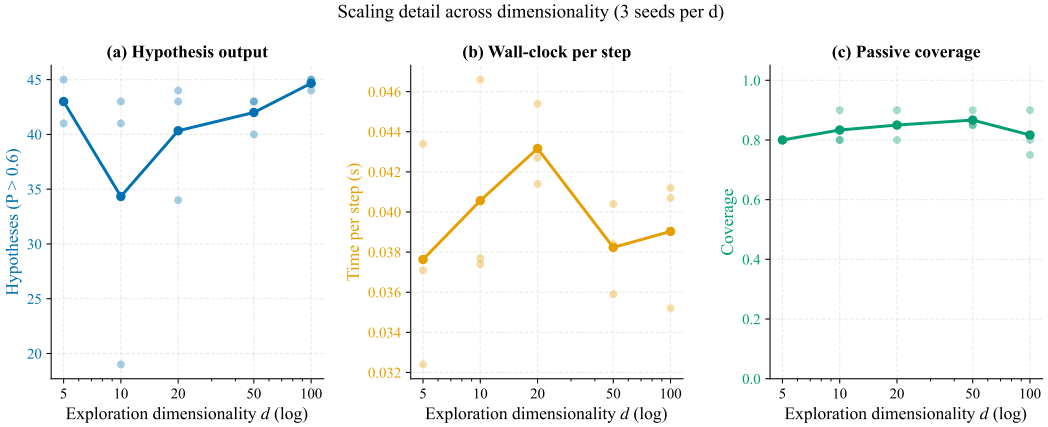

Figure 10: Scaling detail across dimensionality (3 seeds per d, dots = seeds, lines = means): (a) hypothesis output, (b) wall-clock per step, (c) passive coverage. Source: released per-seed audit records.

I HELD-OUT RANDOM WORLDS AND THE GRAMMAR BOUNDARY

To test discovery beyond a single known world, we draw held-out random instances from the same generator family as the synthetic domain (20 dimensions, complexity-1 rules), with the rule count ($n_{rules} \in \{10, 20, 30\}$) and rule structure determined by the seed. The explorer never sees the ground truth, and no configuration is tuned per world. Each instance runs the uniform 300-step protocol, and the baselines carry the same generic top-20% rule extractor as Table 2. Table 13 reports means over three seeds (noise-free block: nine runs per method).

Table 13: Held-out random worlds (20 dimensions, uniform 300-step protocol, 3 seeds per configuration; noise-free block aggregates three rule counts). Baselines carry the generic post-hoc rule extractor of Table 2. Coverage is statistically indistinguishable across methods, as in Table 1. Source: released records.

	Noise-free ($n_{rules} = 10/20/30$)				$\sigma = 0.1$ ($n_{rules} = 20$)		
Method	aRR ↑	HA_{strict} ↑	MVR ↑	Cov.	aRR ↑	HA_{strict} ↑	MVR ↑
SEER	**0.96±0.04**	**0.44±0.13**	0.46±0.18	0.86	**0.80±0.04**	0.07±0.03	0.07±0.03
Random	0.67±0.02	0.04±0.02	0.23±0.07	0.85	0.65±0.00	0.05±0.04	0.07±0.02
GA	0.59±0.04	0.14±0.06	**0.65±0.13**	0.85	0.63±0.02	0.03±0.05	0.09±0.02
BO	0.43±0.09	0.12±0.07	0.41±0.19	0.88	0.62±0.05	0.05±0.00	0.08±0.05

Recovery is rule-count-agnostic (aRR 0.95–0.97 for 10–30 hidden rules the system never saw), held-out replication separates the methods (HA_{strict} 0.34–0.52 for SEER per configuration vs. ≤ 0.20 for every baseline), and the noisy configuration reproduces the noise pattern of Fig. 9(a): rule structure

degrades gracefully while intervention significance collapses for every method. GA's intervention validity on noise-free worlds (0.47–0.75) again confirms that single-rule validity is method-agnostic. The separation lies in breadth and replication.

Grammar boundary. Composite worlds whose hidden rules are two- (C2) or three-way (C3) compositions of ratio, threshold, product, and quadratic sub-rules quantify the boundary stated in §5. SEER's interval/interaction grammar still covers 0.60–0.70 of C2 rules (0.70–0.80 under a relaxed composite-aware matcher reported here only; the frozen audited matcher scores composites 0 by construction because they carry no flat interval), but C3 coverage drops to 0.20–0.40 and the hypothesis population contracts (4–32 theories; two of three seeds below 10). Three-way composition is the honest edge of the current mechanism grammar.

J Superconductivity Open-Discovery Details

Data and surrogate. UCI Superconductivity (Hamidieh, 2018): 21,263 compositions, 15 compositional descriptors, target T_c. The exploration landscape is a k-NN surrogate ($k = 5$, standardized features, evaluation noise 0.02); exploration runs 800 steps $\times$ 3 seeds with 32 candidates/step and ESV mode switching (released exploration scripts).

Two companion record families. For the same 800-step protocol, the audit rerun stores per-seed aggregates and full lineage trees. The open-discovery records store the per-theory list, concept counts, and the proxy T_c curve. Table 14 and Fig. 4(a,b) use the former; best T_c and Fig. 4(c) use the latter.

Table 14: Open-domain per-seed statistics (800-step runs, 3 seeds). BRR = belief revision rate; TED = maximal lineage depth; response = counterexample response rate. #Theories, BRR, TED and Response come from the *audit-rerun* family; best T_c is a k-NN surrogate evaluation from the separate *open-discovery* family (whose per-seed theory counts differ, e.g. seed 42 = 129 here versus 121 in the audit rerun), and is not an experimental measurement. Source: released audit and open-discovery records.

Seed	#Theories	BRR ↑	$\text{TED}_{\max}$	Response ↑	Best T_c (proxy, K)
42	121	0.835	3	0.75	122.1
123	169	0.858	5	0.87	129.4
456	129	0.868	5	0.95	130.8
Mean	**140**	**0.854**	**4.3**	**0.86**	**127.4**

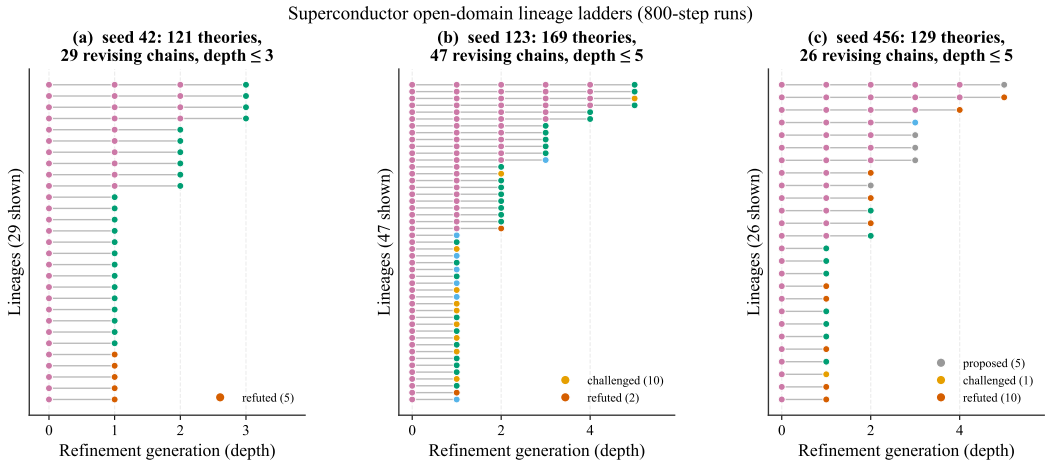

Figure 11: All revision lineages in the open superconductivity runs, one panel per seed: each row is a single-child refinement chain, node colour is theory status. Refinement is the dominant response to counterevidence (29/47/26 revising chains across seeds), with depth reaching 5 generations. Source: theory-tree dumps from the audit rerun.

Concepts and BCS alignment. The open-discovery explorer grounds theories in domain concepts. Table 15 lists the top-8 emergent concepts (seed 42, 129 theories, open-discovery family) with their standard superconductivity correspondences:

Table 15: Emergent concepts and their superconductivity correspondences (counts = theories invoking the concept; seed 42 of the open-discovery record family, 129 theories).

Concept	Correspondence	Count	Annotation
std. first ionization energy	electronic-structure descriptor	73	—
electron–phonon coupling	BCS λ	49	BCS core
mean valence	valence-electron count	38	—
phonon frequency	Debye frequency ω_D	35	BCS core
range atomic radius	radius dispersion (size disorder)	33	—
electron density	carrier density n_e	31	—
valence electron concentration	VEC (electron-concentration criterion)	31	Matthias rule
atomic size mismatch	size-disorder parameter δ	28	—

Proxy-level cross-validation. The 10 highest-scored candidates' nearest neighbours in descriptor space are all Hg-1223 cuprates ($Hg_{0.97}Re_{0.03}Ba_2Ca_2Cu_3O_8$, $T_c = 130\,\mathrm{K}$); mean neighbour T_c $130\,\mathrm{K}$ vs. $15.5\,\mathrm{K}$ for random compositions (random p95 $80.5\,\mathrm{K}$; released cross-validation record). This is surrogate-level, indirect validation of *directed search*, not a materials claim. All T_c figures in this paper are proxy values.

Lineage statistics. Across the three rerun seeds, 121/169/129 theories are proposed with maximal lineage depth 3/5/5 (ladders in Fig. 11); the representative depth-5 chain in Fig. 4(a) carries posteriors $0.39 \to 0.39 \to 0.51 \to 0.63 \to 0.66 \to 0.76$ through five falsification-driven revisions. In the open-discovery records (seed 42), 39 lineages contain at least one revision (84 revision events over 129 theories). The complete per-theory lists (mechanism, prediction, falsification condition, posterior, status; 129/45/45 theories) are stored verbatim in the open-discovery records.

K Background-Knowledge Injection and the Zero-Knowledge Control

What is injected. SEER carries a small, explicitly enumerated domain vocabulary, aggregated and reported by a dedicated accounting module. The complete enumeration below is generated programmatically from the frozen source constants, so the counts cannot drift from the code; none of the entries encodes a ground-truth rule, interval, or answer.

Domain keywords (150 terms). *macrophage* (38): inflammatory response, repair program, metabolic adaptation, M1 polarization, M2 polarization, cytokine expression, gene activation, phenotypic state, cellular response, optimal stimulus range, phenotypic transition zone, dose-response curve, activation threshold, inflammatory window, repair window, stimulus balance, signal ratio, co-stimulation ratio, polarization balance, cytokine ratio, synergistic co-stimulation, combined stimulus effect, non-linear interaction, co-activation, LPS-IFNγ synergy, IL4-IL13 synergy, cross-talk, signaling interaction, macrophage, inflammatory, repair, metabolic, polarization, cytokine, stimulus, activation, phenotype, gene expression.
materials (27): stability, bandgap, formation energy, phase transition, tolerance factor, coordination number, stable range, optimal composition, phase boundary, formation window, tolerance range, ionic radius ratio, electronegativity difference, valence balance, tolerance factor, structural stability, electronic structure, ionic bonding, crystal lattice, perovskite, stability, bandgap, formation, ionic, lattice, crystal, materials.
photonics (26): optical response, mode confinement, quality factor, photonic bandgap, Purcell effect, resonance range, optimal structure, photonic band, confinement window, Q-factor peak, index contrast, fill fraction, defect ratio, photonic crystal, mode coupling, waveguide interaction, optical confinement, index modulation, photonics, optical, mode, confinement, crystal, waveguide, defect, resonance.
synthetic (16): output, function value, target value, score, optimal range, solution region, target zone, variable ratio, parameter balance, variable interaction, joint effect, combined effect, variable, function, parameter, output.
superconductivity (43): critical temperature, superconducting transition, Tc threshold, cooper pair formation, energy gap, condensation energy, phonon frequency cutoff, optimal atomic size range, favorable electron density, Tc enhancement window, structural stability range, valence electron concentration, phonon mode range, atomic radius ratio, electronegativity balance, electron-phonon coupling ratio, mass ratio, valence electron ratio, density of states ratio, electron-phonon interaction, lattice-electron coupling, cooper pair interaction, structural-electronic synergy, size-valence interaction, mass-density effect, phonon-electron synergy, superconductivity, Tc, critical temperature, cooper pair, phonon, electron, lattice, valence electron,

atomic radius, atomic mass, electronegativity, ionization energy, electron affinity, density, fusion heat, thermal conductivity, superconductor.

Property terms (5): structural stability; optical performance; immune activation; gene regulatory response; system performance.

Concept terms (14):

- materials: causal_influence, lattice_stability, formation_energy
- photonics: causal_influence, mode_coupling, bandgap
- macrophage: causal_influence, immune_response, polarization
- perturb_seq: causal_influence, transcription, pathway
- fallback: causal_influence, stability

Knowledge-graph context phrases (16):

macrophage lps: correlates with inflammatory response
macrophage ifn: correlates with pro-inflammatory activation
macrophage il4: correlates with repair program activation
macrophage il13: correlates with alternative polarization
macrophage treatment_time: correlates with temporal response dynamics
materials ionic_radius: correlates with structural stability
materials electroneg: correlates with chemical bonding
materials valence: correlates with oxidation state
materials mass: correlates with atomic weight effects
materials coordination: correlates with crystal structure
photonics period: correlates with photonic bandgap
photonics radius: correlates with mode confinement
photonics thickness: correlates with resonance wavelength
photonics epsilon: correlates with dielectric response
photonics symmetry: correlates with photonic crystal symmetry
photonics filling: correlates with photonic density of states

Synonym map (32 keys, 106 strings):

- irf7 → irf7, interferon regulatory factor 7, ifn regulator
- rela → rela, p65, nfkb subunit, rel a
- stat1 → stat1, signal transducer 1, stat-1
- nfkb1 → nfkb1, nf-kb1, nf kappa b1, p50
- jun → jun, ap1, c-jun, ap-1
- atf2 → atf2, atf-2, activating transcription factor 2
- irf3 → irf3, interferon regulatory factor 3
- cebpb → cebpb, cebpb, ccaat enhancer binding
- lps → lps, lipopolysaccharide, endotoxin
- ifng → ifng, ifn-gamma, interferon gamma, ifnγ
- il4 → il4, il-4, interleukin-4, interleukin 4
- il13 → il13, il-13, interleukin-13, interleukin 13
- ionic_radius_a → ra, ionic radius a, a-site radius, radius_a
- ionic_radius_b → rb, ionic radius b, b-site radius, radius_b
- ionic_radius_x → rx, ionic radius x, x-site radius, radius_x
- electroneg_a → ena, electroneg a, electronegativity a
- electroneg_b → enb, electroneg b, electronegativity b
- valence_a → vala, valence a, oxidation a
- valence_b → valb, valence b, oxidation b
- period → period, lattice constant, a-lattice
- radius → radius, hole radius, rod radius
- thickness → thickness, slab thickness, t
- epsilon → epsilon, dielectric, permittivity, ε
- filling_fraction → filling, fill fraction, ff
- symmetry_break → symmetry, symmetry break, asymmetry

- defect_param → defect, defect param, cavity defect
- interferon → interferon, ifn, antiviral cytokine
- inflammatory → inflammatory, inflammation, nfkb, nf-kb
- antiviral → antiviral, isg, interferon stimulated
- antigen_presentation → antigen, presentation, mhc
- repair → repair, m2, alternative activation
- metabolic → metabolic, metabolism, glycolysis

Aggregator total. The background-knowledge aggregator reports 279 injected terms (145 domain keywords + 28 KG-context tokens + 106 synonym strings); its concept/KG counts are regex overcounts of quoted source tokens: the curated sets enumerated above contain 291 human-readable entries, none encoding ground-truth answers.

Zero-knowledge control (open domain). A wholesale removal of the vocabulary (raw feature names as concepts, generic proposal language) re-runs the superconductivity open-domain protocol. Without domain language the system still searches: it proposes 45 theories and its proxy score is comparable (best normalized score 0.742, best proxy $T_c \approx 137\,\mathrm{K}$). Belief refinement, however, collapses. Against the with-vocabulary open-discovery run at the same protocol (129 theories), the zero-knowledge run reports BRR $0.65 \to 0.00$, falsification-driven revision rate $0.93 \to 0.00$, refinement lineage depth $3 \to 0$, and theory count $129 \to 45$ (released run records). Interpretation: discovery-level exploration persists, but theories become unrefinable one-off proposals because revision operations are expressed through the concept vocabulary. The load-bearing role of background knowledge is therefore refinement and lineage growth, not search quality.

Wording guidance. SEER is not "knowledge-free": its claims should be scoped to *falsification-driven refinement on top of a small, explicitly enumerated domain vocabulary (≈ 280 reported injected terms; no ground-truth content)*, with the zero-knowledge control quantifying exactly what the vocabulary contributes.

L PER-SEED CLOSED-DOMAIN RESULTS

Table 16 lists every closed-domain run behind Table 1 (500-step audit protocol, five seeds). The zero-knowledge control of §5 reproduces the synthetic and materials rows of this table bit-for-bit, which is why no separate table is given for it.

Table 16: Per-seed closed-domain results (500-step audit protocol). The zero-knowledge variant reproduces the synthetic and materials rows bit-for-bit.

Domain	Seed	aRR	HA_{strict}	MVR	Coverage	#Hyp.
Synthetic	42	1.00	0.52	0.61	0.90	69
Synthetic	123	1.00	0.68	0.77	0.90	75
Synthetic	456	1.00	0.47	0.58	0.80	78
Synthetic	789	1.00	0.40	0.42	0.75	73
Synthetic	1024	0.95	0.38	0.36	0.90	72
Materials	42	0.50	0.80	0.86	0.88	50
Materials	123	0.38	0.08	0.08	0.88	25
Materials	456	0.38	0.92	0.92	0.88	65
Materials	789	0.25	0.87	0.89	0.88	54
Materials	1024	0.38	0.66	0.67	0.88	67
Photonics	42	0.38	1.00	1.00	1.00	26
Photonics	123	0.62	1.00	1.00	1.00	17
Photonics	456	0.38	1.00	1.00	1.00	11
Photonics	789	0.25	1.00	1.00	1.00	45
Photonics	1024	0.38	0.96	0.96	1.00	23
Macrophage	42	0.25	1.00	1.00	1.00	2
Macrophage	123	0.62	1.00	1.00	1.00	8
Macrophage	456	0.75	0.97	1.00	1.00	33
Macrophage	789	0.38	0.93	1.00	1.00	14
Macrophage	1024	0.62	1.00	1.00	1.00	41
Perturb-seq	42	0.62	0.16	0.14	0.88	50
Perturb-seq	123	0.38	0.83	1.00	0.88	6
Perturb-seq	456	0.75	0.75	0.85	0.88	20
Perturb-seq	789	0.75	0.87	0.87	0.88	15
Perturb-seq	1024	0.75	0.18	0.18	0.88	78