EVOSCM: SCIENTIFIC BELIEF REVISION THROUGH CAUSAL MODEL EVOLUTION AND EXPERIMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific discovery depends on the ability to form hypotheses, test them through experiments, and revise them when evidence disagrees. Existing LLM agents support this process by improving their reasoning or actions, but their scientific beliefs are often scattered across free-form reasoning and difficult to update coherently. This makes it difficult to identify what failed, what should change, and whether revisions remain consistent with prior evidence. We introduce EvoSCM, which represents scientific beliefs as a population of structural causal model (SCM) hypotheses that can be tested and revised across experiments. EvoSCM formulates scientific discovery as a closed loop in which causal hypotheses guide experimentation and experimental outcomes drive causal model evolution. Competing SCM hypotheses make falsifiable predictions and guide discriminative experiments that separate alternative explanations. When observations contradict these predictions, EvoSCM distills discrepancies into correction rules identifying which aspects of the hypotheses fail to explain the evidence. These rules guide revisions to causal dependencies, latent factors, mechanisms, and parameters. Revised hypotheses are validated against accumulated evidence and carried forward to guide subsequent experiments, allowing scientific beliefs to evolve cumulatively. We evaluate EvoSCM across physics, chemistry and materials, and biology. It consistently outperforms baseline agents and existing evolution methods, yielding more accurate explanations and predictions with more effective use of experimental budgets. The evolved SCMs also transfer across base models, suggesting reusable scientific knowledge beyond any single model's reasoning process.

1 INTRODUCTION

Scientific discovery is inherently iterative: hypotheses are proposed, tested through experiments, and revised as new evidence accumulates. Recent agents can improve how they reason and act through prompts, memories, skills, or workflows (Gao et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026; Yang et al., 2026; Zhang et al., 2025; Nguyen et al., 2026), and can increasingly generate scientific hypotheses, design experiments, and interpret outcomes (Boiko et al., 2023; Lu et al., 2024; Huang et al., 2025; Abhyankar et al., 2026; Kabra et al., 2026; Wiemann et al., 2026). These capabilities improve how agents carry out scientific tasks, but they do not necessarily provide an explicit and revisable representation of what the agent currently believes about the underlying system. Instead, such beliefs often remain distributed across free-form reasoning traces, experimental records, or memory summaries (Takahara & Mizoguchi, 2026). This makes it difficult to connect hypotheses with the evidence that supports, contradicts, or motivates their revision.

This limitation becomes particularly important when new evidence contradicts a prediction. The agent must determine which assumption may be responsible, what part of its current explanation should be revised, and whether the revised explanation remains consistent with prior observations. When scientific beliefs are not represented explicitly, it becomes difficult to trace prediction failures back to specific assumptions and to propagate the resulting revisions coherently across subsequent experiments. Thus, scientific agents can struggle to revise their hypotheses reliably when confronted with contradictory evidence (Ríos-García et al., 2026). This motivates an explicit representation of scientific beliefs that links hypotheses, predictions, and experimental evidence, and can be tested and revised as new evidence arrives.

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

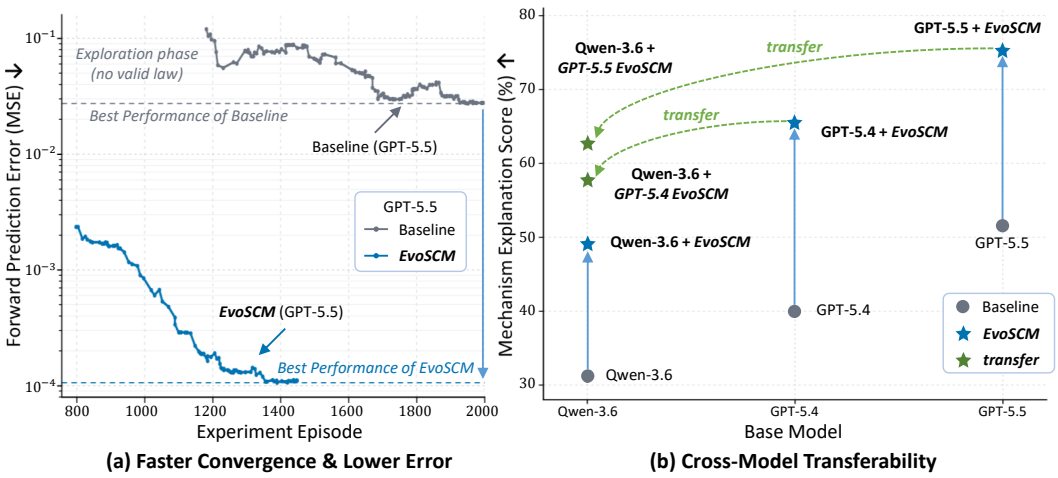

Figure 1: **Discovery performance and transferability of EvoSCM on DiscoverPhysics** (Wiemann et al., 2026). (a) *Faster Convergence & Lower Error*: EvoSCM achieves lower prediction error while requiring fewer experimental episodes than the baseline agent. (b) *Cross-Model Transferability*: evolved SCMs can be transferred directly across different base models.

To this end, we introduce EvoSCM, a framework that represents the agent's evolving scientific beliefs as a population of competing structural causal model (SCM) hypotheses. Each hypothesis provides an explicit causal explanation through its graph structure, latent factors, functional mechanisms, and parameters, making the agent's current understanding directly testable and revisable. EvoSCM couples this representation with experimentation in a closed loop. In each round, the current hypotheses abductively interpret accumulated evidence, guide interventions that discriminate among competing explanations, and make falsifiable predictions. Prediction–observation discrepancies are inductively distilled into correction rules that drive targeted revisions to causal structure, latent variables, mechanisms, and parameters. The revised hypotheses are then deductively evaluated against accumulated evidence and structural consistency before guiding the next round of experimentation. In this way, EvoSCM extends agent evolution from improving how an agent reasons to also revising what the agent believes about the world.

We evaluate EvoSCM across three scientific domains: non-canonical physical law discovery (DiscoverPhysics (Wiemann et al., 2026)), chemistry and materials discovery (LLEMA (Abhyankar et al., 2026)), and biological network inference (ActiveSciBench-GRN (Kabra et al., 2026)). Across all three domains, EvoSCM consistently outperforms baseline agents and existing evolution methods, yielding more accurate explanations and predictions while using experimental budgets more effectively (Figure 1(a)). Moreover, the evolved SCMs transfer across base architectures, suggesting that the learned causal models capture reusable scientific knowledge beyond any particular model's reasoning process (Figure 1(b)).

2 Related Work

Self-Evolving LLM Agents. LLM agents can improve from experience without updating their parameters (Gao et al., 2026), for example by optimizing prompts (Agrawal et al., 2026), distilling reasoning trajectories into reusable memories (Ouyang et al., 2026), acquiring skills (Wang et al., 2023; Yang et al., 2026), searching over workflows (Zhang et al., 2025), or recursively refining agent states (Nguyen et al., 2026). A common thread across these approaches is that experience improves the agent's *procedure*: how it reasons, plans, and acts. EvoSCM targets a complementary axis, using experience to revise the agent's *epistemic state*: its explicit model of how the external world works.

Scientific Agents. Recent systems automate parts of the scientific pipeline (Boiko et al., 2023; Lu et al., 2024; Huang et al., 2025), yet evidence-driven belief revision could be fragile: refutation rarely triggers genuine updates to an agent's hypotheses (Ríos-García et al., 2026). Among hypothesis-driven discovery systems, LLM-AutoSciLab (Kabra et al., 2026) evolves competing hypotheses and selects discriminative experiments, LLEMA (Abhyankar et al., 2026) applies evolutionary search to

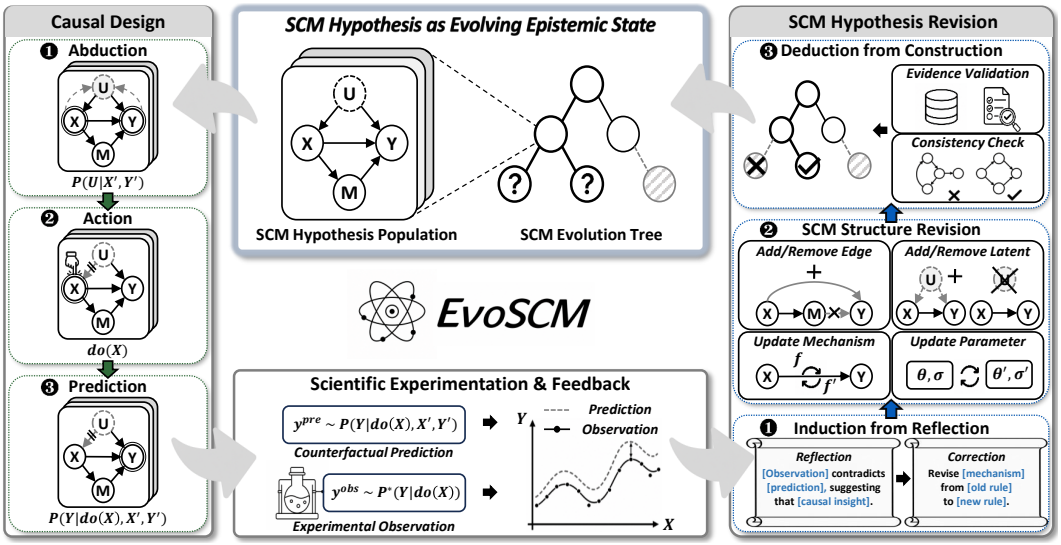

Figure 2: **Overview of EvoSCM.** EvoSCM maintains a population of competing SCM hypotheses and evolves it through a closed discovery loop: (1) *Causal Experiment Design*: the agent abductively interprets accumulated evidence, selects discriminative interventions, and commits to falsifiable predictions tested through experimentation; (2) *SCM Hypothesis Revision*: prediction–observation discrepancies are inductively distilled into correction rules that guide targeted revisions to causal structure and mechanisms, while revised hypotheses are deductively validated against accumulated evidence and structural consistency to form the next-round population.

multi-objective materials discovery, and PiEvo (Pu et al., 2026) evolves natural-language scientific principles via uncertainty minimization. In contrast, EvoSCM combines executable causal models with structural revisability, grounding belief revision in falsifiable, mechanism-level predictions.

3 EvoSCM: Causal Model Evolution

3.1 SCM Hypothesis as Evolving Epistemic State

Scientific Discovery Formulation. Scientific discovery is the process of uncovering the laws and principles that govern observed phenomena through iterative hypothesis formation, experimental testing, and belief revision (Popper, 2005). We formalize agentic scientific discovery as the sequential recovery of an unknown causal model through experimentation. An agent interacts with an environment $\mathcal{E}$ governed by a true structural causal model $\mathcal{M}^* = (\mathbf{V}, \mathbf{U}^*, G^*, \mathbf{F}^*, \mathbf{\Theta}^*)$ (Pearl, 2009), where $\mathbf{V} = \{V_i\}_{i=1}^n$ are observable endogenous variables, $\mathbf{U}^* = \{U_i^*\}_{i=1}^n$ are mutually independent latent exogenous variables, G^* is a directed acyclic graph (DAG) encoding the true causal structure, $\mathbf{F}^* = \{f_i^*\}_{i=1}^n$ are the structural equations (causal mechanisms) with $V_i = f_i^*(\mathrm{Pa}_i^{G^*}, U_i^*; \Theta_i^*)$, and $\mathbf{\Theta}^* = \{\Theta_i^*\}_{i=1}^n$ parameterizes these mechanisms. Discovery proceeds over a budget of T rounds. At each round t, the agent selects an intervention $\mathrm{do}(\mathbf{X}_t = \mathbf{x}_t)$ on a subset $\mathbf{X}_t \subseteq \mathbf{V}$ and observes outcomes $\mathbf{y}_t \sim P^*(\mathbf{Y}_t \mid \mathrm{do}(\mathbf{X}_t = \mathbf{x}_t))$ for the remaining variables $\mathbf{Y}_t = \mathbf{V} \setminus \mathbf{X}_t$. The accumulated evidence through round t is $\mathcal{D}_t = \{(\mathrm{do}(\mathbf{x}_\tau), \mathbf{y}_\tau)\}_{\tau=1}^t$.

The agent models the environment with an SCM hypothesis $\mathcal{H} = (\mathbf{V}, \mathbf{U}, G, \mathbf{F}, \mathbf{\Theta})$ that shares the observable variables $\mathbf{V}$ but posits its own latent variables, causal graph, mechanisms, and parameters. The goal of agentic scientific discovery is to evolve $\mathcal{H}$ over the T rounds so that it recovers the causal structure, mechanisms, and predictive behavior of $\mathcal{M}^*$. An SCM is a natural representation for the epistemic state of a scientific agent: its causal graph G encodes qualitative causal structure, while its structural equations $\mathbf{F}$ with parameters $\mathbf{\Theta}$ encode quantitative mechanisms. Together, $(G, \mathbf{F}, \mathbf{\Theta})$ constitutes a complete, falsifiable scientific hypothesis: the $\mathrm{do}(\cdot)$ operator (Pearl, 2009) allows the agent to derive interventional predictions that can be directly tested against experi-

mental outcomes, and any prediction–observation discrepancy can be traced to specific structural or parametric commitments within the hypothesis.

Population-Based Epistemic State. A single hypothesis, however, may be insufficient for effective scientific discovery. Multiple causal structures can be consistent with the evidence available at any given round, and prematurely committing to one risks overlooking the true mechanism. EvoSCM therefore maintains a population of competing hypotheses $\mathcal{P}_t = \{\mathcal{H}_1^t, \ldots, \mathcal{H}_K^t\}$ as its epistemic state at round t. Each hypothesis $\mathcal{H}_k^t = (\mathbf{V}, \mathbf{U}_k^t, G_k^t, \mathbf{F}_k^t, \mathbf{\Theta}_k^t)$ encodes a distinct candidate causal explanation of $\mathcal{E}$. Maintaining a diverse population serves two purposes: it guards against premature convergence to an incorrect explanation, and the disagreement among competing hypotheses directly guides experiment design by identifying interventions under which current candidates make maximally divergent predictions (Lindley, 1956; Box & Hill, 1967).

Overview of EvoSCM. As shown in Figure 2, EvoSCM maintains a population of competing SCM hypotheses $\mathcal{P}_t$ as the agent's current scientific belief state and evolves this population through a closed discovery loop. Each round consists of two complementary phases: the current hypothesis population first guides experiment design and prediction, and the resulting experimental feedback then drives hypothesis revision to produce the next population $\mathcal{P}_{t+1}$.

(1) *Casual Experiment Design.* Given $\mathcal{P}_t$ and accumulated evidence, the agent abductively interprets the observations under each hypothesis, selects interventions that discriminate among competing explanations, and commits to falsifiable predictions before experimentation. The experimental outcomes are then compared with these predictions to expose where the current hypotheses fail.

(2) *SCM Hypothesis Revision.* Prediction–observation discrepancies are inductively distilled into correction rules that guide revisions to causal structure, latent variables, mechanisms, and parameters. The revised hypotheses are used deductively to derive consequences under accumulated interventions and validated against prior evidence and structural consistency. Well-supported hypotheses form the updated population $\mathcal{P}_{t+1}$, which in turn guides the next round of experimentation.

After T rounds, the evolved SCM population $\mathcal{P}_T$ encodes the agent's accumulated causal knowledge and can be used for downstream reasoning, prediction, and generalization to unseen interventions.

3.2 Casual Experiment Design

Given the current hypothesis population $\mathcal{P}_t$ and accumulated evidence $\mathcal{D}_t$, the agent designs the next experiment by reasoning causally within each hypothesis and comparatively across the population. This phase follows the logic of Pearl's counterfactual framework (Pearl, 2009): abduction (infer latent states), action (select an intervention), and prediction (derive expected outcomes), extended with an experimentation step that produces the empirical feedback driving hypothesis revision.

Abduction. For each hypothesis $\mathcal{H}_k^t \in \mathcal{P}_t$, the agent infers the latent exogenous variables most consistent with the accumulated evidence: $P(\mathbf{U}_k^t \mid \mathcal{D}_t, \mathcal{H}_k^t)$. This abductive step conditions on all past interventions and observations in $\mathcal{D}_t$ to determine, under hypothesis $\mathcal{H}_k^t$, what values of the latent variables $\mathbf{U}_k^t$ would best explain the observed outcomes. Because each hypothesis posits its own latent variables and causal structure, the same evidence may yield different abduced latent states across hypotheses, anchoring each candidate in a distinct causal interpretation of the available data.

Action. The agent selects the next intervention $\mathrm{do}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1})$ to maximally discriminate among the competing hypotheses in $\mathcal{P}_t$. Concretely, the agent identifies interventions under which the hypotheses' predicted outcomes diverge most, so that the experimental result will provide maximal evidence for distinguishing between candidate explanations (Lindley, 1956; Box & Hill, 1967). This population-level disagreement is what makes experiment design *comparative*: rather than testing a single hypothesis in isolation, the agent exploits the diversity of the population to select experiments whose outcomes are most likely to narrow the set of viable candidates (Chaloner & Verdinelli, 1995).

Prediction. Before executing the intervention, the agent commits to a falsifiable, hypothesis-specific prediction for each candidate. Under hypothesis $\mathcal{H}_k^t$ with abduced latent states, the predicted outcome is: $\mathbf{y}_k^{\mathrm{pre}} \sim P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}_{t+1}{=}\mathbf{x}_{t+1}), \mathcal{D}_t, \mathcal{H}_k^t)$. These predictions are recorded prior to experimentation, ensuring that each hypothesis stakes a concrete, testable claim on the experimental outcome. This commitment is essential: it prevents post-hoc rationalization and allows subsequent discrepancies to be unambiguously attributed to specific hypotheses.

Experimentation and Feedback. The agent executes the selected intervention $\mathrm{do}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1})$ in the environment $\mathcal{E}$ and observes the outcome: $\mathbf{y}_{t+1}^{\mathrm{obs}} \sim P^*(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1}))$, drawn from the true data-generating process $\mathcal{M}^*$. The agent then compares each hypothesis's committed prediction $\mathbf{y}_k^{\mathrm{pre}}$ against the observed outcome $\mathbf{y}_{t+1}^{\mathrm{obs}}$, producing a set of prediction–observation discrepancies that identify where and how each candidate's causal account fails to match reality. The accumulated evidence is updated as $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\mathrm{do}(\mathbf{x}_{t+1}), \mathbf{y}_{t+1}^{\mathrm{obs}})\}$, and both the discrepancies and the updated evidence are passed to the hypothesis revision phase.

3.3 SCM Hypothesis Revision

The prediction–observation discrepancies and updated evidence $\mathcal{D}_{t+1}$ produced by the experimentation phase drive the evolution of the hypothesis population. This phase transforms each hypothesis through three steps: induction extracts structured correction rules from empirical discrepancies, revision applies these rules as targeted edits to causal structure and mechanisms, and deduction validates the revised hypotheses against accumulated evidence and structural consistency to produce $\mathcal{P}_{t+1}$.

Induction from Reflection. For each hypothesis $\mathcal{H}_k^t \in \mathcal{P}_t$, the agent analyzes the discrepancies between its committed predictions $\mathbf{y}_k^{\mathrm{pre}}$ and the observed outcomes $\mathbf{y}_{t+1}^{\mathrm{obs}}$ to identify systematic patterns of failure. Rather than treating each discrepancy as an isolated error, the agent examines the accumulated evidence $\mathcal{D}_{t+1}$ to detect recurring regularities that the current hypothesis fails to capture. These regularities are distilled into explicit *correction rules*: concise, natural-language descriptions of the empirical pattern that the hypothesis must accommodate. Correction rules serve as the interface between empirical evidence and structural model editing: they translate quantitative prediction failures into qualitative directives that specify *what* must change in the hypothesis without prescribing *how* that change should be realized at the structural level.

SCM Structure Revision. Guided by the correction rules from the induction step, the agent applies targeted revisions to each hypothesis $\mathcal{H}_k^t$. EvoSCM supports four revision operators, each modifying a different component of an SCM hypothesis: (1) *Add/Remove Edge*: modify the causal graph G_k^t by inserting a new directed edge to represent a previously unmodeled causal dependency, or removing an existing edge that the evidence no longer supports; (2) *Add/Remove Latent*: introduce a new latent variable into $\mathbf{U}_k^t$ to account for unexplained confounding or mediating effects, or remove a latent variable that has become redundant given the revised structure; (3) *Update Mechanism*: replace or modify a structural equation $f_i \in \mathbf{F}_k^t$ to change the functional form of a causal mechanism; (4) *Update Parameter*: adjust the parameters $\mathbf{\Theta}_k^t$ to improve quantitative fit while preserving the current structure and mechanisms. The agent selects the minimal set of operators sufficient to accommodate the correction rules, preferring shallow revisions (parameter and mechanism updates) when possible and resorting to deeper structural edits (edge and latent changes) only when the evidence warrants them. Each revision produces a candidate revised hypothesis $\tilde{\mathcal{H}}_k^{t+1} = (\mathbf{V}, \tilde{\mathbf{U}}_k^{t+1}, \tilde{G}_k^{t+1}, \tilde{\mathbf{F}}_k^{t+1}, \tilde{\mathbf{\Theta}}_k^{t+1})$, which is then passed to the deduction step for validation.

Deduction from Construction. The deduction step validates each candidate revised hypothesis $\tilde{\mathcal{H}}_k^{t+1}$ before it is admitted into the next-round population $\mathcal{P}_{t+1}$. Validation comprises two checks: (1) *Evidence Validation*: the agent derives predictions from $\tilde{\mathcal{H}}_k^{t+1}$ for all interventions in the accumulated evidence $\mathcal{D}_{t+1}$ and verifies that the revised hypothesis accounts for past observations. A hypothesis that resolves the most recent discrepancy but introduces new contradictions with earlier evidence is rejected; (2) *Consistency Check*: the agent verifies the internal coherence of the revised hypothesis: the causal graph $\tilde{G}_k^{t+1}$ must remain a valid DAG, the structural equations must be well-defined for all variables given the revised graph, and the posited mechanisms must be mutually compatible. Hypotheses that pass both checks are retained; those that fail either check are discarded from the population. This selective step instantiates the population-level evolution: individual hypotheses are structurally revised in the preceding step, while deductive validation determines which revised candidates survive into the next round, producing the evolved population $\mathcal{P}_{t+1}$. The retained population, together with the updated evidence $\mathcal{D}_{t+1}$, seeds the next round of experiment design.

3.4 Inference with Evolved SCM

After T rounds of the discovery loop, the evolved hypothesis population $\mathcal{P}_T$ encodes the agent's accumulated causal knowledge as a structured, executable scientific model. Given a novel intervention

Table 1: **Physical law discovery on DiscoverPhysics** (Wiemann et al., 2026). EvoSCM consistently outperforms all baselines in explanation score, prediction error, success rate, and experimental efficiency on both GPT-5.4 (OpenAI, 2026a) and GPT-5.5 (OpenAI, 2026b).

Backbone	Method	⟨Explanation⟩ ↑	norm ⟨MSE⟩ ↓	pass@1 ↑	pass@2 ↑	pass@3 ↑	pass@4 ↑	pass@5 ↑	Episodes ↓
GPT-5.4	Baseline	39.80%	5.25e-1	1.86%	3.88%	5.60%	7.14%	9.09%	1,045
	GEPA	35.64%	5.83e-1	0.00%	0.00%	0.00%	0.00%	0.00%	976
	ReasoningBank	36.55%	1.84e0	5.45%	10.00%	13.64%	16.36%	18.18%	908
	ModelSMC	28.55%	1.28e0	1.82%	3.64%	5.45%	7.27%	9.09%	901
	PiEvo	28.18%	2.25e0	1.82%	3.64%	5.45%	7.27%	9.09%	1,033
	EvoSCM (Ours)	**65.30%**	**2.34e-3**	**38.08%**	**50.16%**	**53.74%**	**54.55%**	**54.55%**	**736**
GPT-5.5	Baseline	51.64%	2.83e-2	7.27%	13.64%	19.09%	23.64%	27.27%	2,206
	GEPA	45.09%	3.55e-2	9.09%	14.55%	17.27%	18.18%	18.18%	2,062
	ReasoningBank	51.27%	2.78e-2	10.91%	19.09%	24.55%	27.27%	27.27%	1,797
	ModelSMC	43.27%	1.83e-2	7.27%	13.64%	19.09%	23.64%	27.27%	1,884
	PiEvo	48.18%	5.01e-2	9.09%	16.36%	21.82%	25.45%	27.27%	1,845
	EvoSCM (Ours)	**75.09%**	**2.77e-4**	**56.36%**	**62.73%**	**63.64%**	**63.64%**	**63.64%**	**1,685**

$\mathrm{do}(\mathbf{X}' = \mathbf{x}')$ not present in $\mathcal{D}_T$, each surviving hypothesis derives a prediction through its structural equations, $\mathbf{y}'_k \sim P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}' = \mathbf{x}'), \mathcal{D}_T, , \mathcal{H}_k^T)$, and when multiple hypotheses survive, predictions are aggregated into a consensus outcome. This generalization capacity follows from representing knowledge as a causal model: the $\mathrm{do}(\cdot)$ operator derives predictions for unseen interventions by simulating causal consequences through the discovered graph and mechanisms, rather than interpolating from stored input–output associations (Pearl, 2009; Bareinboim & Pearl, 2016). The evolved SCM is also decoupled from the experimental trajectory that produced it, constituting a portable, inspectable model that can be transferred to new agents or used to plan future experiments, a property procedural improvements such as better prompts or workflows do not naturally support.

4 Experiment

4.1 Experimental Setup

Benchmark. We evaluate EvoSCM across three scientific domains. (1) *Physics*: DiscoverPhysics (Wiemann et al., 2026) requires agents to uncover the hidden dynamics of 22 noncanonical physical worlds through interactive experimentation. (2) *Chemistry & Materials*: LLEMA (Abhyankar et al., 2026) tasks agents with multi-objective materials discovery across 14 material design challenges. (3) *Biology*: ActiveSciBench-GRN (Kabra et al., 2026) requires agents to recover the causal graph structure of 45 gene regulatory networks from interventional gene expression data through active experimentation. Each task instance is evaluated over 5 independent random seeds.

Base Agents and Evolution Methods. Each benchmark domain provides a baseline agent implementing the domain-specific experimental loop (hypothesis generation, experiment execution, and result interpretation) (Wiemann et al., 2026; Abhyankar et al., 2026; Kabra et al., 2026). We equip each base agent with five evolution methods: GEPA (Agrawal et al., 2026) (reflective prompt evolution), ReasoningBank (Ouyang et al., 2026) (reasoning memory distillation), ModelSMC (Wahl et al., 2026) (probabilistic model discovery), PiEvo (Pu et al., 2026) (principle evolution via uncertainty minimization), and EvoSCM. All methods share the same base agent, LLM backbone, and experimental budget per domain, differing only in how they evolve the agent's state from evidence.

4.2 Evaluation

Physical Law Discovery. Table 1 reports results on DiscoverPhysics (Wiemann et al., 2026). Following its evaluation protocol, we measure mechanism explanation score (Explanation ↑), normalized prediction error (norm MSE ↓), and per-world success rate (pass@k ↑). We additionally report experimental efficiency (Episodes ↓) to measure how quickly each method converges. EvoSCM substantially outperforms all baselines on GPT-5.4 (OpenAI, 2026a) and GPT-5.5 (OpenAI, 2026b), reducing prediction error by over two orders of magnitude while requiring fewer experimental episodes. These results suggest that refining how an agent reasons is insufficient when its underlying scientific beliefs remain implicit, and that EvoSCM's gains stem from maintaining and revising an explicit causal model that grounds every prediction in a falsifiable structural commitment.

Table 2: **Chemistry & Materials discovery on LLEMA** (Abhyankar et al., 2026). EvoSCM consistently outperforms all baselines in hit rate (H.R., %) and stability (Stab., %) based on Qwen3.6-35B-A3B (Qwen Team, 2026).

Method	Wide-Bandgap Semicond.		SAW/BAW Acoustic Substrates		High-k Dielectrics		Solid-State Electrolytes		Piezo Energy Harvesters		Transparent Conductors		Insulating Dielectrics	
	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑
Baseline	3.33	0.83	45.00	0.00	0.00	0.00	5.00	1.67	32.50	0.00	0.00	0.00	0.83	0.00
GEPA	10.00	0.00	38.33	0.00	1.67	0.00	2.50	0.00	35.83	0.00	0.00	0.00	2.50	0.00
ReasoningBank	14.17	6.67	32.50	0.00	0.00	0.00	13.33	6.67	26.67	0.00	0.00	0.00	3.33	0.00
ModelSMC	10.83	5.83	45.83	3.33	0.00	0.00	30.00	0.83	39.17	0.00	0.00	0.00	1.67	0.00
PiEvo	20.83	5.83	40.83	0.00	1.67	0.00	22.50	15.00	54.17	0.00	0.00	0.00	0.83	0.00
EvoSCM (Ours)	**21.67**	**15.00**	**75.83**	**15.00**	**5.00**	**5.00**	**40.83**	**30.00**	**69.17**	**6.67**	**10.83**	**10.83**	**6.67**	**5.00**

Method	Photovoltaics Absorbers		Hard Coating Materials		Hard, Stiff Ceramics		Aerospace Materials		Acousto-optic Hybrids		Low Density Structures		Perovskite Oxides	
	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑
Baseline	0.00	0.00	0.00	0.00	44.17	0.00	40.00	0.00	7.50	0.00	0.00	0.00	1.67	0.00
GEPA	0.00	0.00	2.50	0.00	41.67	0.00	37.50	0.00	8.33	0.00	0.00	0.00	0.00	0.00
ReasoningBank	0.00	0.00	0.00	0.00	50.00	0.00	42.50	0.00	12.50	0.00	0.00	0.00	0.83	0.00
ModelSMC	0.00	0.00	0.00	0.00	25.00	0.00	28.33	0.00	12.50	0.00	0.00	0.00	0.00	0.00
PiEvo	1.67	0.00	0.00	0.00	61.67	0.00	47.50	0.00	9.17	0.00	1.67	1.67	0.00	0.00
EvoSCM (Ours)	**11.67**	**7.50**	**24.17**	**23.33**	**68.33**	**8.33**	**60.00**	**8.33**	**20.83**	**6.67**	**5.00**	**5.00**	**7.50**	**4.17**

Table 3: **Biological network inference on ActiveSciBench-GRN** (Kabra et al., 2026). EvoSCM consistently outperforms baselines in edge F1 (%), exact graph accuracy (%) and sign accuracy (%) on Qwen3.6-35B-A3B (Qwen Team, 2026) and GPT-5.6-Luna (OpenAI, 2026c).

Method	Qwen3.6-35B-A3B			GPT-5.6-Luna		
	Edge F1 ↑	Exact Graph Accuracy ↑	Sign Accuracy ↑	Edge F1 ↑	Exact Graph Accuracy ↑	Sign Accuracy ↑
Baseline	81.71	28.15	95.44	81.49	18.52	98.32
GEPA	84.28	36.30	98.09	85.48	25.19	98.30
ReasoningBank	84.10	35.56	97.05	86.04	34.07	99.15
ModelSMC	76.57	28.89	96.74	82.59	21.48	97.29
PiEvo	77.79	25.93	97.58	77.38	17.41	97.79
EvoSCM (Ours)	**96.12**	**70.37**	**99.85**	**96.94**	**77.04**	**99.82**

Chemistry & Materials Discovery. Table 2 reports results on LLEMA (Abhyankar et al., 2026) across 14 material design tasks based on Qwen3.6-35B-A3B (Qwen Team, 2026). Following its evaluation protocol, we measure hit rate (H.R. ↑), the percentage of discovered materials satisfying target property constraints, and stability (Stab. ↑), the percentage of discovered materials remaining valid under perturbation. EvoSCM achieves the highest hit rate and stability on all 14 tasks. Most baselines achieve zero stability on the majority of tasks, indicating that their discovered materials are brittle and fail under perturbation, while EvoSCM achieves nonzero stability on every task. This suggests that an explicit causal model of material property relationships enables EvoSCM to discover candidates grounded in structural mechanisms rather than surface-level heuristics.

Biological Network Inference. Table 3 reports results on ActiveSciBench-GRN (Kabra et al., 2026), where agents must recover a signed, directed causal graph of gene regulatory networks from budget-limited perturbation experiments. Following its evaluation protocol, we measure edge F1 (↑), exact graph accuracy (↑), and sign accuracy (↑), evaluating the agent's ability to recover correct regulatory edges, complete network topologies, and edge polarities, respectively. EvoSCM substantially outperforms all baselines on every metric, on both Qwen3.6-35B-A3B (Qwen Team, 2026) and GPT-5.6-Luna (OpenAI, 2026c). The most notable gap is in exact graph accuracy, indicating that EvoSCM recovers complete network topologies far more reliably than baselines that improve individual edge predictions without maintaining a coherent graph-level hypothesis.

4.3 Analysis

Cross-Model SCM Transferability. A key advantage of representing scientific knowledge explicitly as an SCM is its portability across model architectures, as shown in Table 4. We use

Table 4: **Cross-model transferability results on DiscoverPhysics** (Wiemann et al., 2026). SCMs evolved by GPT transfer seamlessly to Qwen3.6-35B-A3B, substantially improving its performance.

Method	⟨Explanation⟩ ↑	norm ⟨MSE⟩ ↓	pass@1 ↑	pass@2 ↑	pass@3 ↑	pass@4 ↑	pass@5 ↑
Qwen3.6 Baseline	30.73%	1.90e-1	0.00%	0.00%	0.00%	0.00%	0.00%
Qwen3.6 + Qwen3.6 SCM	48.91%	6.31e-2	14.95%	26.07%	32.92%	36.36%	36.36%
Qwen3.6 + GPT-5.4 SCM	57.64%	2.85e-2	16.40%	23.43%	26.41%	27.27%	27.27%
Qwen3.6 + GPT-5.5 SCM	62.55%	8.69e-3	34.87%	39.52%	41.96%	43.72%	45.45%
Qwen3.6 + GPT-5.6-Sol SCM	74.00%	1.06e-4	39.85%	49.76%	57.14%	61.87%	63.64%

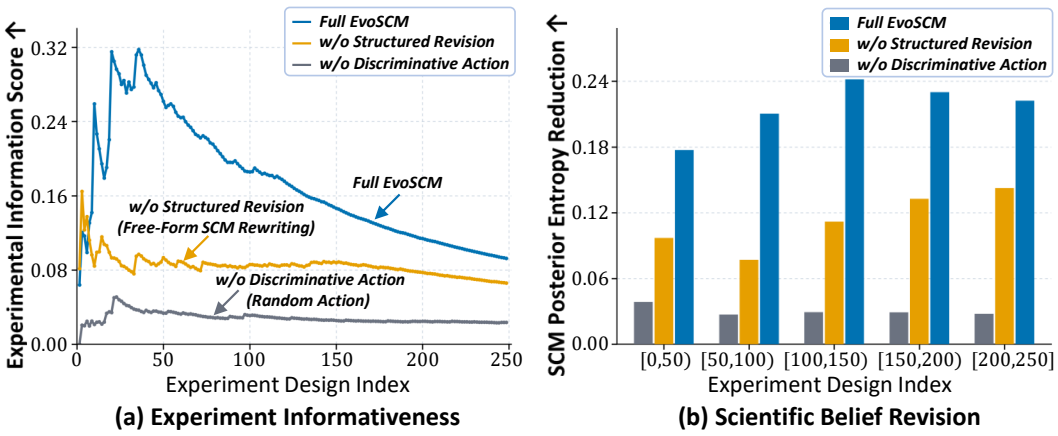

Figure 3: **Experimentation and belief revision analysis on DiscoverPhysics** (Wiemann et al., 2026). (a) *Experiment Informativeness*: EvoSCM designs experiments with higher information scores (Lindley, 1956; MacKay, 1992). The score peaks early when the hypothesis population is most diverse and declines as competing hypotheses converge, leaving less disagreement to exploit. (b) *Scientific Belief Revision*: EvoSCM achieves consistently larger posterior entropy reductions (Chaloner & Verdinelli, 1995; MacKay, 1992), confirming more effective belief revision.

Qwen3.6-35B-A3B (Qwen Team, 2026), which fails outright on DiscoverPhysics, as the transfer target. Equipping Qwen3.6 with an SCM it evolved itself already recovers substantial capability, and injecting SCMs evolved by stronger models (GPT-5.4 (OpenAI, 2026a), GPT-5.5 (OpenAI, 2026b), and GPT-5.6-Sol (OpenAI, 2026c)) pushes explanation accuracy and prediction error further in the same direction, with gains scaling monotonically with the strength of the source model. This confirms that the evolved SCMs capture reusable scientific knowledge beyond any particular model's reasoning process and can be transferred directly across base models.

Experimentation and Belief Revision Analysis. Figure 3 examines the two directions of EvoSCM's discovery loop, comparing EvoSCM against two ablations: w/o discriminative action (random action) and w/o structured revision (free-form SCM rewriting). EvoSCM designs experiments with substantially higher information scores (Lindley, 1956; MacKay, 1992), demonstrating that the SCM population guides the agent toward interventions that maximally discriminate between competing hypotheses. These informative experiments translate into consistently larger SCM posterior entropy reductions (Chaloner & Verdinelli, 1995; MacKay, 1992): the agent's uncertainty over the correct causal model decreases faster per experiment, concentrating posterior weight on hypotheses that best explain the evidence and driving more effective scientific belief revision.

SCM Hypothesis Evolution Analysis. Figure 4 traces the evolution of EvoSCM's SCM hypothesis on the Yukawa world, from an embryonic Coulomb-like hypothesis at Round 9 ($\boldsymbol{a}_t \propto 1/r_t$) to the ground-truth Yukawa law. The intermediate states along this trajectory are not arbitrary but physically meaningful, each resolving a distinct class of prediction failures. At Round 30, the agent introduces a screening factor $S_t = K_1(r_t/\lambda)/(2\pi\lambda)$, a structure coinciding with the Yukawa screening mechanism (Yukawa, 1935), which resolves the systematic overprediction at intermediate distances. At Round 59, a further revision introduces a softened distance $\rho_t = \sqrt{r_t^2 + \varepsilon^2}$, coinciding with the finite core model (Jastrow, 1951), which resolves the residual short-range discrepancies that

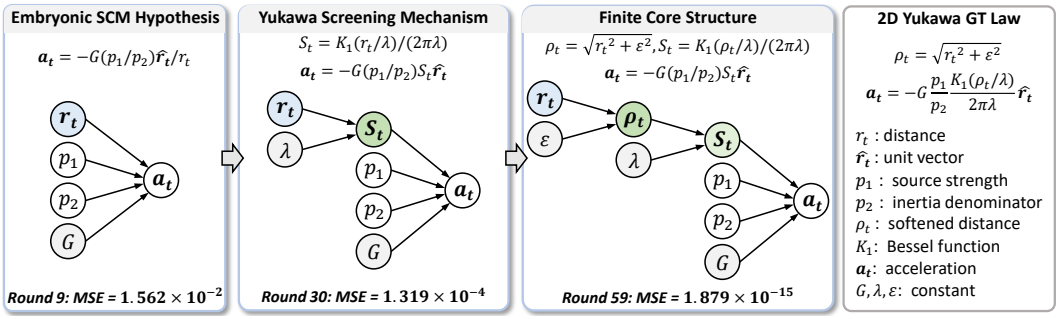

Figure 4: **SCM hypothesis evolution on the Yukawa world in DiscoverPhysics** (Wiemann et al., 2026). EvoSCM progressively recovers the 2D Yukawa ground-truth law through physically meaningful intermediate states, rather than merely fitting parameters along a smooth refinement path. (1) *Round 9*: the agent starts from an embryonic Coulomb-like hypothesis ($\boldsymbol{a}_t \propto 1/r_t$); (2) *Round 30*: a revision introduces a screening factor $S_t = K_1(r_t/\lambda)/(2\pi\lambda)$, coinciding with the Yukawa screening mechanism (Yukawa, 1935), resolving intermediate-range overprediction; (3) *Round 59*: a further revision introduces a softened distance $\rho_t = \sqrt{r_t^2 + \varepsilon^2}$, coinciding with the finite core model (Jastrow, 1951), resolving residual short-range discrepancies.

Table 5: **Ablation study on EvoSCM.** We evaluate SCM representation, discriminative action, and structured revision.

Method	⟨Explanation⟩ ↑	norm ⟨MSE⟩ ↓
Full EvoSCM	**75.09%**	**2.77e-4**
w/o SCM (Free-Form Textual Hypotheses)	48.18%	7.82e-2
w/o Discriminative Action (Random Action)	47.45%	1.05e-1
w/o Structured Revision (Free-Form SCM Rewriting)	70.73%	5.45e-4

Table 6: **Ablation study on the SCM hypothesis population size** (K).

Population Size	⟨Explanation⟩ ↑	norm ⟨MSE⟩ ↓
$K = 1$	63.27%	6.78e-3
$K = 4$	74.86%	2.59e-4
$K = 6$	**75.09%**	**2.77e-4**
$K = 8$	75.01%	2.71e-4

screening alone cannot explain. This observation suggests that the evidence-driven causal model evolution does not merely fit parameters along a smooth loss landscape but discovers structurally distinct intermediate hypotheses, each capturing a qualitatively new aspect of the underlying law.

Ablation Study on EvoSCM. Table 5 evaluates the contribution of each core component. Replacing SCMs with free-form textual hypotheses causes the largest drop, highlighting the value of an explicit causal representation. Removing discriminative action selection also substantially degrades performance, while free-form SCM rewriting yields a smaller but consistent decline, supporting targeted hypothesis revision. Table 6 studies population size: performance improves from $K = 1$ to $K = 4$ and remains stable for larger K, suggesting that maintaining several competing hypotheses is beneficial while EvoSCM is not sensitive to larger populations.

5 Conclusion

We present EvoSCM, a framework that represents scientific beliefs as explicit and persistent structural causal models, enabling scientific agents to revise not only how they reason, but also what they believe about the world. EvoSCM maintains a population of competing SCM hypotheses and evolves them through a closed loop of abduction, intervention, induction, and deduction, supporting a structured, falsifiable, and cumulatively revisable process of scientific discovery. Across physics, chemistry, and biology, EvoSCM consistently improves scientific discovery over baseline agents and existing evolution methods, yielding more accurate explanations and predictions while using experimental budgets more effectively. The evolved SCMs transfer across base architectures, suggesting that the resulting causal models encode reusable scientific knowledge beyond the reasoning process of any particular model. Together, these results support explicit causal model evolution as a promising direction for scientific agents that refine their understanding through experimentation.

Limitation and Future Direction. Our evaluation focuses on simulated scientific environments. Extending EvoSCM to real-world settings, where observations are noisy, interventions are costly, and ground truth is unavailable, remains an important next step.

AI USE STATEMENT

Large language models are an integral component of the proposed EvoSCM framework and its experimental evaluation. EvoSCM uses LLMs as the reasoning backbone of scientific agents, performing abductive inference, causal experiment design, inductive rule extraction, structural revision, and deductive validation, as described in Sections 3.2 and 3.3. We evaluate GPT-5.4, GPT-5.5, GPT-5.6-Luna, GPT-5.6-Sol, and Qwen3.6-35B-A3B as LLM backbones, executing the full scientific discovery loop on three benchmarks (DiscoverPhysics, LLEMA, and ActiveSciBench-GRN), as described in Section 4. All LLM-generated scientific outputs, including causal graphs, structural equations, predictions, experimental decisions, and revision decisions, are quantitatively evaluated against the corresponding ground-truth benchmarks. These uses of LLMs are intrinsic to the EvoSCM system studied and evaluated in this work.

Separately, generative AI tools were used to assist with polishing portions of the manuscript for clarity and presentation. Outside the execution and evaluation of the proposed EvoSCM framework, generative AI tools were not used to determine the core methodological contributions or to generate or alter the reported experimental results. All AI-assisted text was reviewed, revised where appropriate, and verified by the authors for accuracy, correctness, and originality. The authors take full responsibility for the final content of this work, including all text, claims, methods, experimental results, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

EvoSCM is a general framework for scientific discovery through causal model evolution and does not involve human subjects, personal data, or sensitive information. Our experiments are conducted in computational scientific-discovery benchmark environments (DiscoverPhysics, LLEMA, and ActiveSciBench-GRN) and do not involve the collection or processing of personally identifiable information or data requiring human-subject consent.

EvoSCM is intended to assist scientific reasoning by representing hypotheses as explicit, inspectable structural causal models, not to replace human scientific judgment. The resulting SCMs, predictions, and revision traces can be inspected and evaluated, facilitating verification by domain experts before any downstream use. While the benchmark evaluations in this work introduce no direct human-subject or privacy risks, automated scientific discovery systems may produce incorrect, incomplete, or misleading hypotheses. Conclusions generated by such systems should therefore not be treated as independently validated scientific findings without appropriate empirical verification and human oversight, particularly in safety-critical or real-world scientific applications.

REPRODUCIBILITY STATEMENT

We provide an anonymized implementation of EvoSCM as part of the supplementary material. The supplementary code includes the EvoSCM framework implementation and the configurations used in our experiments. We intend to publicly release the source code upon publication.

The EvoSCM framework, including its SCM representation, experiment design, inference, and hypothesis revision procedures, is described in Sections 3.1, 3.2, 3.3, and 3.4. Experimental settings and evaluation protocols are provided in Section 4.1 and the Appendix, including the LLM backbones, population sizes, experimental budgets, random seeds, and other implementation details required to reproduce the experiments. We use the official evaluation protocols of DiscoverPhysics (Wiemann et al., 2026), LLEMA (Abhyankar et al., 2026), and ActiveSciBench-GRN (Kabra et al., 2026); benchmark-specific configurations are documented in Section 4.1, the Appendix, and the supplementary code.

REFERENCES

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan Reddy. Llema: Evolutionary search with llms for multi-objective materials discovery. In *International Conference on Learning Representations*, pp. 139415–139442, 2026.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. 2023. URL https://arxiv.org/abs/2311.07361.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6709–6738, 2025.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

George EP Box and WILLIAM J Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. 2023. URL https://arxiv.org/abs/2308.10848.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. 2025. URL https://arxiv.org/abs/2508.07407.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 13481–13544, 2024.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Qihan Ren, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence. *Transactions on Machine Learning Research*, 2026.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations*, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. 2025. URL https://arxiv.org/abs/2502.09858.

Robert Jastrow. On the nucleon-nucleon interaction. *Physical Review*, 81(2):165, 1951.

Sanchit Kabra, Nikhil Abhyankar, Saaketh Desai, Prasad Iyer, and Chandan K Reddy. Llm-autoscilab: Closed-loop scientific discovery via active experimentation with llms. 2026. URL https://arxiv.org/abs/2605.24043.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. 2024. URL https://arxiv.org/abs/2408.06292.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 2023.

Michael Nguyen, Quoc Nguyen, and Paul Vuong. Recursive self-evolving agents via held-out selection. 2026. URL https://arxiv.org/abs/2606.28374.

OpenAI. Introducing GPT-5.4, March 2026a. URL https://openai.com/index/introducing-gpt-5-4/.

OpenAI. Introducing GPT-5.5, April 2026b. URL https://openai.com/index/introducing-gpt-5-5/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026c. URL https://openai.com/index/gpt-5-6/.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Karl Popper. *The logic of scientific discovery*. routledge, 2005.

Yingming Pu, Tao Lin, and Hongyu Chen. Principle-evolvable scientific discovery via uncertainty minimization. 2026. URL https://arxiv.org/abs/2602.06448.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. 2026. URL https://arxiv.org/abs/2503.24047.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Martiño Ríos-García, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Ai scientists produce results without reasoning scientifically. 2026. URL https://arxiv.org/abs/2604.18805.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 2023.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable ai scientists: A hypothesis evolution protocol for llm agents. 2026. URL https://arxiv.org/abs/2607.09195.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. A survey on self-evolution of large language models. 2024. URL https://arxiv.org/abs/2404.14387.

Stefan Wahl, Raphaela Schenk, Ali Farnoud, Jakob H. Macke, and Daniel Gedon. A probabilistic framework for llm-based model discovery. 2026. URL https://arxiv.org/abs/2602.18266.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. 2023. URL https://arxiv.org/abs/2305.16291.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of computer science*, 18(6):186345, 2024a.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 2024b.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, Shuangjia Zheng, Ning Ding, Wanli Ouyang, Nanqing Dong, Yu Cheng, Siqi Sun, Lei Bai, and Bowen Zhou. From ai for science to agentic science: A survey on autonomous scientific discovery. 2025. URL https://arxiv.org/abs/2508.14111.

Matt L. Wiemann, Lindsay M. Smith, Peter Melchior, Siddharth Mishra-Sharma, Andrew Gordon Wilson, Pavel Izmailov, and Carolina Cuesta-Lázaro. Discoverphysics: Benchmarking llms for out-of-the-box scientific thinking. 2026. URL https://arxiv.org/abs/2605.26087.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. 2023. URL https://arxiv.org/abs/2308.08155.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. Skillopt: Executive strategy for self-evolving agent skills. 2026. URL https://arxiv.org/abs/2605.23904.

Hideki Yukawa. On the interaction of elementary particles. i. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 17:48–57, 1935.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

APPENDIX

TABLE OF CONTENTS

A Additional Results **15**

A.1 Additional Results on DiscoverPhysics 15

A.2 Additional Results on LLEMA 15

A.3 Additional Results on ActiveSciBench-GRN 15

A.4 LLM Token Usage 16

B Additional Analysis **17**

B.1 Additional Ablation Study 17

B.2 Additional Case Studies 18

C Algorithmic Details **20**

C.1 SCM Representation 20

C.2 Prompt Design 21

C.3 Pseudocode 22

D Benchmark Details **23**

D.1 DiscoverPhysics 23

D.2 LLEMA 24

D.3 ActiveSciBench-GRN 25

E Experiment Details **26**

E.1 DiscoverPhysics 26

E.2 LLEMA 28

E.3 ActiveSciBench-GRN 30

F Yukawa World Details **32**

G Additional Related Work **33**

G.1 Self-Evolving LLM Agents 33

G.2 Scientific Agents 34

A Additional Results

A.1 Additional Results on DiscoverPhysics

Table 7 reports additional DiscoverPhysics (Wiemann et al., 2026) results using GPT-5.6-Sol (OpenAI, 2026c) as the base model. EvoSCM outperforms the baseline across all metrics, improving explanation score from 63.64% to 79.13%, reducing prediction error by over five times, and requiring roughly one-third fewer experimental episodes (3,543 vs. 5,296). EvoSCM nevertheless produces consistent gains on top of this stronger base, confirming that its improvements are not limited to weaker models and that explicit causal model evolution provides complementary benefits beyond what a more capable LLM backbone alone can achieve.

Table 7: **Physical law discovery on DiscoverPhysics** (Wiemann et al., 2026). EvoSCM consistently outperforms all baselines in explanation score, prediction error, success rate, and experimental efficiency on GPT-5.6-Sol (OpenAI, 2026c).

Method	⟨Explanation⟩ ↑	norm ⟨MSE⟩ ↓	pass@1 ↑	pass@2 ↑	pass@3 ↑	pass@4 ↑	pass@5 ↑	Episodes ↓
Baseline	63.64%	7.79e-4	34.76%	49.37%	57.51%	61.68%	63.64%	5296
EvoSCM (Ours)	**79.13%**	**9.96e-5**	**43.82%**	**61.39%**	**68.30%**	**70.96%**	**72.73%**	**3543**

A.2 Additional Results on LLEMA

Table 8 reports additional LLEMA (Abhyankar et al., 2026) results using GPT-5.6-Luna (OpenAI, 2026c) as the base model, comparing EvoSCM against the LLEMA baseline agent. EvoSCM achieves a higher hit rate and stability on all 14 material design tasks, consistent with the Qwen3.6-35B-A3B results reported in the main paper (Table 2). The stability advantage is again particularly notable: the baseline achieves zero stability on 10 of 14 tasks, while EvoSCM achieves nonzero stability on every task, with the largest margins on SAW/BAW Acoustic Substrates (19.17% vs. 0.00%) and Solid-State Electrolytes (29.17% vs. 1.67%). These results confirm that EvoSCM's improvements generalize across LLM backbones and are not specific to a particular base architecture.

Table 8: **Chemistry & Materials discovery on LLEMA** (Abhyankar et al., 2026). EvoSCM consistently outperforms the LLEMA Baseline in hit rate (H.R., %) and stability (Stab., %) based on GPT-5.6-Luna (OpenAI, 2026c).

Method	Wide-Bandgap Semicond.		SAW/BAW Acoustic Substrates		High-k Dielectrics		Solid-State Electrolytes		Piezo Energy Harvesters		Transparent Conductors		Insulating Dielectrics	
	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑
Baseline	10.00	3.33	29.17	0.00	0.00	0.00	6.67	1.67	54.17	0.00	1.67	1.67	0.83	0.00
EvoSCM (Ours)	**16.67**	**10.00**	**64.17**	**19.17**	**8.33**	**6.67**	**31.67**	**29.17**	**79.17**	**5.00**	**7.50**	**7.50**	**6.67**	**4.17**
Method	**Photovoltaics Absorbers**		**Hard Coating Materials**		**Hard, Stiff Ceramics**		**Aerospace Materials**		**Acousto-optic Hybrids**		**Low Density Structures**		**Perovskite Oxides**	
	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑	H.R ↑	Stab. ↑
Baseline	0.00	0.00	0.83	0.00	25.83	0.00	47.50	0.00	8.33	0.83	0.83	0.83	1.67	0.00
EvoSCM (Ours)	**8.33**	**7.50**	**8.33**	**6.67**	**58.33**	**9.17**	**62.50**	**5.83**	**17.50**	**6.67**	**5.00**	**5.00**	**7.50**	**4.17**

A.3 Additional Results on ActiveSciBench-GRN

Tables 9 and 10 report per-regime results on ActiveSciBench-GRN (Kabra et al., 2026) across three predefined kinetic regimes (Easy, Medium, Hard) on Qwen3.6-35B-A3B (Qwen Team, 2026) and GPT-5.6-Luna (OpenAI, 2026c), respectively. The regimes reflect progressively sharper and more nonlinear parameter settings: Easy tasks exhibit quasi-linear responses where small perturbations reveal the graph clearly, Medium tasks introduce saturation effects that obscure weak edges, and Hard tasks feature bistability and switching behavior where the graph is identifiable only through carefully designed multi-node perturbations (Kabra et al., 2026). EvoSCM substantially outperforms all baselines on edge F1 and exact graph accuracy across every regime and both backbones, with particularly large margins in exact graph accuracy, the most demanding metric requiring every

edge to be recovered correctly. On Qwen3.6, EvoSCM achieves perfect sign accuracy on Medium and Hard regimes and near-perfect on Easy, while on GPT-5.6-Luna it achieves competitive sign accuracy across all regimes, with ReasoningBank matching or slightly exceeding on two regimes but trailing substantially on edge F1 and exact graph accuracy. EvoSCM's performance remains stable across regimes: the gap between Easy and Hard is minimal on both backbones, whereas several baselines (particularly ModelSMC and PiEvo) degrade on harder regimes. This robustness suggests that EvoSCM's causal experiment design is particularly effective in the Hard regime, where the correct graph structure can only be identified through targeted multi-node interventions rather than simple single-variable perturbations, exactly the setting where population-level disagreement guides the agent toward the most revealing experiments.

Table 9: **Biological network inference across three predefined kinetic regimes on ActiveSciBench-GRN** (Kabra et al., 2026). EvoSCM consistently outperforms baselines in edge F1 (%), exact graph accuracy (%) and sign accuracy (%) on Qwen3.6-35B-A3B (Qwen Team, 2026).

Method	Easy			Medium			Hard		
	Edge F1 ↑	Exact Graph ↑	Sign ↑	Edge F1 ↑	Exact Graph ↑	Sign ↑	Edge F1 ↑	Exact Graph ↑	Sign ↑
Baseline	81.59	33.33	95.04	82.24	33.33	95.93	81.29	17.78	95.37
GEPA	83.63	31.11	96.67	85.00	44.44	99.44	84.20	33.33	98.15
ReasoningBank	82.40	33.33	96.04	85.59	44.44	97.70	84.30	28.89	97.41
ModelSMC	77.25	33.33	95.85	77.87	28.89	97.33	74.60	24.44	97.04
PiEvo	77.73	26.67	99.00	77.10	24.44	97.70	78.52	26.67	96.04
EvoSCM (Ours)	**96.19**	**68.89**	**99.56**	**96.43**	**71.11**	**100.00**	**95.74**	**71.11**	**100.00**

Table 10: **Biological network inference across three predefined kinetic regimes on ActiveSciBench-GRN** (Kabra et al., 2026). EvoSCM substantially improves edge F1 (%) and exact graph accuracy (%), achieving competitive sign accuracy (%) on GPT-5.6-Luna (OpenAI, 2026c).

Method	Easy			Medium			Hard		
	Edge F1 ↑	Exact Graph ↑	Sign ↑	Edge F1 ↑	Exact Graph ↑	Sign ↑	Edge F1 ↑	Exact Graph ↑	Sign ↑
Baseline	80.02	13.33	97.52	81.28	24.44	99.00	83.17	17.78	98.44
GEPA	85.62	20.00	97.56	85.17	26.67	98.89	85.65	28.89	98.44
ReasoningBank	85.04	28.89	97.89	86.95	37.78	**100.00**	86.13	35.56	**99.56**
ModelSMC	82.75	17.78	98.26	85.09	28.89	97.11	79.93	17.78	96.52
PiEvo	78.28	12.22	99.00	76.71	18.89	96.78	77.14	21.11	97.59
EvoSCM (Ours)	**96.81**	**80.00**	**99.96**	**97.70**	**80.00**	99.96	**96.31**	**71.11**	**99.56**

A.4 LLM Token Usage

Figure 5 reports the total LLM token consumption for each method across the three benchmarks. Despite achieving consistently superior discovery performance, EvoSCM maintains moderate to low token usage relative to other evolution methods. On LLEMA, EvoSCM consumes the fewest tokens among all evolution methods, using only marginally more than the unevolved baseline. On ActiveSciBench-GRN, EvoSCM uses fewer tokens than even the baseline agent, likely because its more causal experiment design and targeted revisions lead to faster convergence, reducing the total number of reasoning steps required. On DiscoverPhysics, EvoSCM uses more tokens than the baseline and PiEvo but substantially fewer than GEPA, ReasoningBank, and ModelSMC. This efficiency can be attributed to the structured SCM representation: causal graphs, structural equations, and parameters encode scientific knowledge compactly, whereas free-form textual hypotheses, reasoning memories, and verbose model descriptions require substantially more tokens to represent and manipulate. Overall, these results indicate that EvoSCM's performance gains are not achieved by simply consuming more computational resources but rather by using tokens more effectively through structured causal representations and targeted revisions.

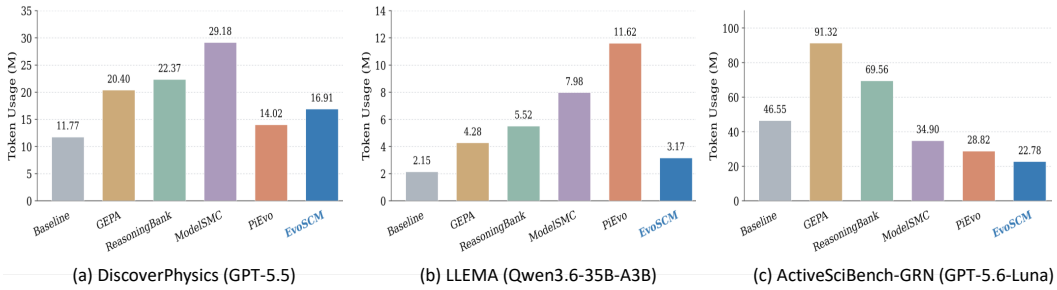

Figure 5: **LLM token usage (millions) across three benchmarks.** EvoSCM achieves the strongest scientific discovery performance (Tables 1, 2, and 3) while maintaining competitive token usage compared to other evolution methods across all three domains.

B Additional Analysis

B.1 Additional Ablation Study

Table 11 extends the main ablation study (Table 5) with three additional ablations that complete the coverage of EvoSCM's discovery loop stages.

w/o Induction (No Correction Rules). Raw prediction-observation discrepancies are passed directly to the Revision step without being distilled into correction rules. The LLM must simultaneously identify the failure pattern and determine the appropriate structural edit, rather than receiving a distilled directive that separates diagnosis from repair. Performance degrades moderately, confirming that the inductive abstraction step produces more targeted revisions by decoupling *what went wrong* from *how to fix it*.

w/o Pre-committed Prediction (Only Observations). The agent observes experimental outcomes before formulating its explanation, rather than committing to hypothesis-specific predictions in advance. Without pre-commitment, the agent can rationalize any observation post-hoc, weakening the discrepancy signal that drives revision. This ablation produces a notable decline, indicating that pre-commitment is important not because it improves the individual predictions, but because it ensures that each experiment generates an unambiguous, falsifiable test of the current hypotheses.

w/o Abduction (No Latent State Inference). Predictions are derived directly from the structural equations without conditioning on accumulated evidence to infer latent variable values. The core discovery loop remains intact, but predictions become less calibrated because each hypothesis is not grounded in what previous experiments revealed about the latent state. This ablation produces the mildest degradation among the three, suggesting that abduction primarily refines prediction quality rather than driving structural discovery.

Table 11: **Ablation study on EvoSCM.** We evaluate the effects of SCM representation, discriminative action, structured revision, deductive validation, induction, pre-committed prediction, and abductive inference.

Method	⟨**Explanation**⟩ ↑	**norm** ⟨**MSE**⟩ ↓
Full EvoSCM	**75.09%**	**2.77e-4**
w/o SCM (Free-Form Textual Hypotheses)	48.18%	7.82e-2
w/o Discriminative Action (Random Action)	47.45%	1.05e-1
w/o Structured Revision (Free-Form SCM Rewriting)	70.73%	5.45e-4
w/o Deduction (No Evidence Validation or Consistency Checking)	67.09%	8.83e-4
w/o Induction (No Correction Rules)	66.37%	9.94e-4
w/o Pre-committed Prediction (Only Observations)	63.47%	4.79e-3
w/o Abduction (No Latent State Inference)	69.58%	5.74e-4

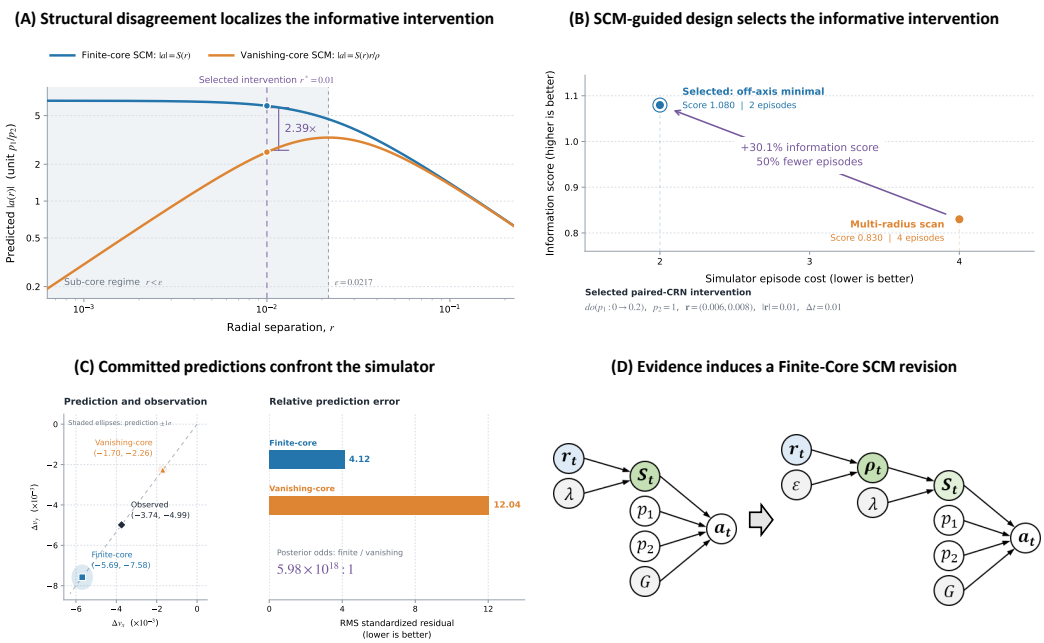

Figure 6: **Detailed case study: finite core recovery on the Yukawa world in Discover-Physics** (Wiemann et al., 2026). (A) Surviving hypotheses diverge at short range, localizing a discriminative intervention at $r^* = 0.01$. (B) EvoSCM selects the higher-information, lower-cost experiment. (C) Pre-committed predictions favor the finite-core hypothesis. (D) A localized revision inserts $r_t \to \rho_t \to S_t$ while preserving the rest of the causal structure.

B.2 Additional Case Studies

Detailed case study on the Yukawa world. Figure 6 zooms into a single revision event to show how the four stages of EvoSCM's loop jointly recover the finite core from Figure 4. (A) After identifying the screened Yukawa mechanism, the surviving hypotheses agree in the far field but diverge at short range: predicted accelerations differ by $2.39\times$ at $r^* = 0.01$, making this the maximally discriminative intervention point. (B) EvoSCM selects a paired-CRN off-axis intervention over a multi-radius scan for its higher information score and lower cost (1.080 vs. 0.830; 2 vs. 4 episodes). (C) The finite-core hypothesis produces a substantially smaller residual than the vanishing-core alternative (4.12σ vs. 12.04σ), enabling decisive model selection. (D) The resulting revision inserts a softened distance $\rho_t = \sqrt{r_t^2 + \varepsilon^2}$, replacing the direct edge $r_t \to S_t$ with $r_t \to \rho_t \to S_t$ while preserving the learned screening kernel and coupling structure.

Case Study on LLEMA. Figure 7 illustrates the evolution of EvoSCM's task-specific epistemic model for solid-state electrolyte discovery on LLEMA (Abhyankar et al., 2026). At Stage A, after 16 oracle-evaluated observations, EvoSCM forms its first actionable hypothesis: higher halogen content and coordination adaptability are associated with lower energy above the convex hull, while higher mean electronegativity favors formation energy and lattice anisotropy introduces an energetic penalty. Four additional experiments, including informative chemical counterexamples, lead to Stage B, where unsupported relations are revised and composition-related mechanisms are promoted: metallic fraction becomes positively associated with hull energy, while electronegativity contrast complements mean electronegativity in explaining formation energy. After another four boundary-refinement experiments, Stage C integrates elemental diversity, coordination variation, halogen content, electronegativity statistics, and lattice anisotropy into a multi-factor stability hypothesis. The model continues to abstain from positing band-gap relations because their estimated reliability remains below the acceptance threshold, demonstrating calibrated epistemic restraint rather than forcing a complete explanation. These structures should be interpreted as local,

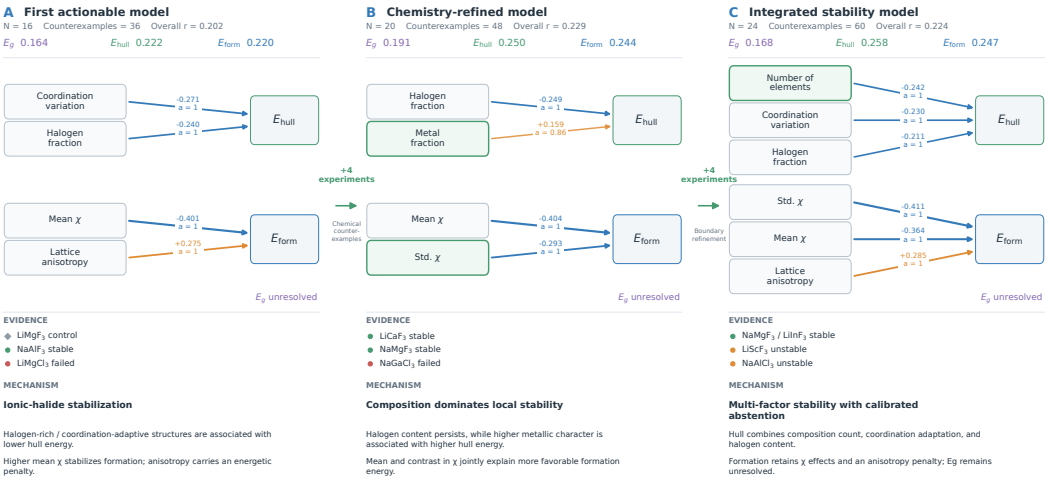

Figure 7: **SCM hypothesis evolution for solid-state electrolyte discovery on LLEMA** (Abhyankar et al., 2026). EvoSCM progressively refines its causal model of material stability across three stages of active experimentation. Edge color denotes the sign of the standardized coefficient, edge width its magnitude, opacity indicates sign agreement across the population of $K = 7$ candidate SCMs, and green-outlined variables denote newly promoted mechanisms.

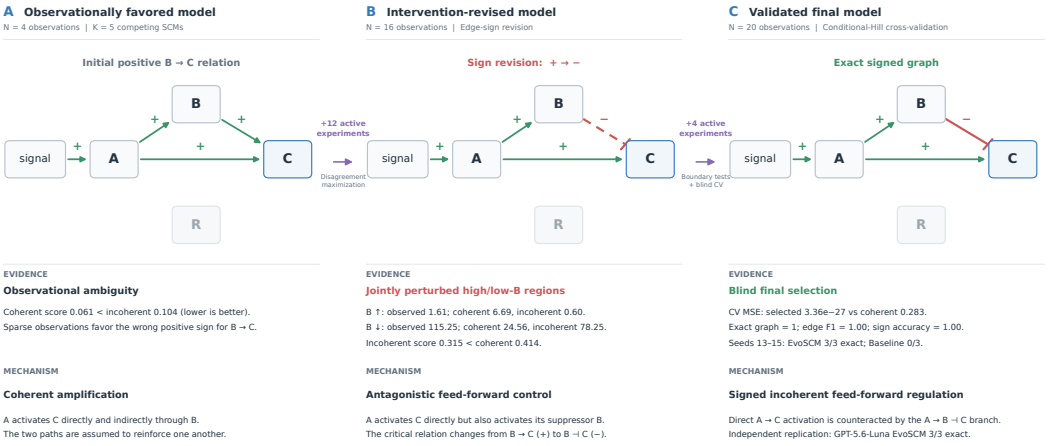

Figure 8: **SCM hypothesis evolution for gene regulatory network inference on ActiveSciBench-GRN** (Kabra et al., 2026). EvoSCM corrects an initial coherent feed-forward hypothesis to the ground-truth incoherent feed-forward motif through targeted soft interventions and sign revision.

intervention-guiding epistemic relations learned within the frozen ALIGNN evaluation environment, not as recovery of universal causal laws for solid-state electrolytes.

Case Study on ActiveSciBench-GRN. Figure 8 illustrates how EvoSCM revises a mechanistic hypothesis on ActiveSciBench-GRN (Kabra et al., 2026). With only four initial observations, the coherent feed-forward model is preferred over the incoherent alternative (fit score 0.061 vs. 0.104, lower is better), leading to the incorrect hypothesis that B activates C. EvoSCM then selects continuous soft interventions in regions where the candidate SCMs disagree most strongly, jointly varying signal and perturbation multipliers rather than changing a single variable in isolation. The resulting observations increasingly contradict the coherent explanation: for example, in two selected high- and low-B regions, the observed reporter values are 1.61 and 115.25, compared with coherent-model predictions of 6.69 and 24.56 and incoherent-model predictions of 0.60

and 78.25, respectively. After 16 observations, the working hypothesis revises the critical relation from $B \to C\,(+)$ to $B \dashv C\,(-)$, yielding an antagonistic feed-forward mechanism in which A activates C directly while also activating its suppressor B. Following four additional boundary-testing experiments, the conditional-Hill cross-validation procedure selects the complete signed graph $\{\text{signal} \to A,\ A \to B,\ A \to C,\ B \dashv C\}$ with a MSE of 3.36×10^{-27}, compared with 0.283 for the coherent alternative. Because coherent and incoherent motifs both belong to the public candidate grammar, this case demonstrates active discrimination, sign revision, and validation among competing biological mechanisms rather than unconstrained de novo discovery.

C Algorithmic Details

C.1 SCM Representation

EvoSCM represents each SCM hypothesis $\mathcal{H}_k^t = (\mathbf{V}, \mathbf{U}_k^t, G_k^t, \mathbf{F}_k^t, \mathbf{\Theta}_k^t)$ as a structured JSON object that the LLM can read, reason over, and edit. The representation encodes three core components of the SCM tuple alongside operational metadata.

Causal Graph. The graph G_k^t is stored as an edge list of directed causal relationships. Each variable is annotated with its role (`state`, `parameter`) and observability, enabling the Revision step to identify which variables can be added or removed:

```
"variables": [
{"name": "r_t", "role": "state", "observed": true},
{"name": "C", "role": "parameter", "observed": false}
],
"edges": [
{"source": "r_t", "target": "a_t"},
{"source": "p1", "target": "a_t"}
]
```

Structural Equations. Each mechanism $f_i \in \mathbf{F}_k^t$ is stored as a symbolic expression string together with its parent variables:

```
"mechanisms": [{
"target": "a_t",
"parents": ["r_t", "p1", "p2"],
"equation": "a_t = -C*(p1/p2)*r_vec/
(r**2 + eps**2)"
}]
```

Storing the symbolic expression alongside its declared parents allows the LLM to reason about functional dependencies and supports direct numerical evaluation during the Prediction step.

Parameters. Parameters $\mathbf{\Theta}_k^t$ are stored with point estimates and optional bounds:

```
"parameters": {
"C": {"estimate": 0.1585,
"lower": 0.155, "upper": 0.163}
}
```

Separating parameters from functional forms enables the Update Parameter operator to adjust quantitative values without modifying the structural equations.

Operational Metadata. Beyond the formal SCM tuple, each hypothesis carries metadata that supports the discovery loop: a natural-language description summarizing the causal explanation, a list of assumptions (e.g., `"isotropic"`, `"static"`), falsification conditions that would trigger revision, the hypothesis status (`active` or `pruned`), and a revision counter tracking how many times the hypothesis has been edited. This metadata is not part of the formal SCM but provides context that helps the LLM generate more targeted revisions and more interpretable explanations. A complete example of the JSON representation is provided in Listing 1.

Listing 1: Example SCM hypothesis representation (Yukawa environment).

```
{
  "id": "H_screened_K1",
  "name": "Screened 2D Helmholtz field",
  "description": "Static isotropic screened field
                  with signed p1/p2 coupling.",
  "variables": [
    {"name": "pos2_t", "role": "state",
     "observed": true},
    {"name": "p1", "role": "source_control",
     "observed": true},
    {"name": "p2", "role": "response_inertia",
     "observed": true},
    {"name": "lambda", "role": "parameter",
     "observed": false}
  ],
  "edges": [
    {"source": "pos2_t",
     "target": "field_gradient_t"},
    {"source": "lambda",
     "target": "field_gradient_t"},
    {"source": "p1", "target": "a_t"},
    {"source": "p2", "target": "a_t"},
    {"source": "field_gradient_t",
     "target": "a_t"},
    {"source": "a_t",
     "target": "velocity2_next"}
  ],
  "mechanisms": [{
    "target": "a_t",
    "parents": ["pos2_t", "p1", "p2",
                "G", "lambda"],
    "equation": "a_t = -G*(p1/p2)*K1(r/lambda)/(2*pi*lambda)*pos2_t/r"
  }],
  "parameters": {
    "G": {"estimate": 0.9992,
          "lower": 0.05, "upper": 5.0},
    "lambda": {"estimate": 1.9999,
               "lower": 0.2, "upper": 10.0},
    "epsilon": {"estimate": 0.035,
                "lower": 1e-6, "upper": 0.2}
  },
  "assumptions": [
    "static",
    "isotropic",
    "velocity independent",
    "visible signs of p1 and p2 are causal"
  ],
  "falsifiers": [
    "far-range responses follow a constant
     power law rather than exponential
     K1 suppression"
  ],
  "status": "active",
  "revision": 6
}
```

Population Management. The population $\mathcal{P}_t = \{\mathcal{H}_1^t, \ldots, \mathcal{H}_K^t\}$ is maintained as an indexed collection of JSON objects. At each round, the serialized population and the accumulated evidence $\mathcal{D}_t$ are provided to the LLM so that inter-hypothesis comparisons (e.g., prediction divergence for casual experiment design) can be performed within the same call. In practice, the population sizes used in our experiments ($K \leq 8$) remain well within the context limits of all evaluated LLM backbones.

C.2 Prompt Design

Each stage of the EvoSCM discovery loop is implemented as a structured LLM call. All calls share a common preamble containing the serialized hypothesis population $\mathcal{P}_t$, the accumulated evidence $\mathcal{D}_t$, and a description of the observable variables $\mathbf{V}$. Below we summarize the stage-specific instructions; full prompt templates are included in the released code.

Abduction. The LLM receives each hypothesis $\mathcal{H}_k^t$ with the accumulated evidence and is instructed to infer latent variable values consistent with the observed data:

> *Given the following SCM hypothesis and experimental evidence, infer the values of the latent exogenous variables that best explain the observed outcomes. For each latent variable, provide a point estimate and a brief justification based on the structural equations and evidence.*

Action. The LLM receives the full population $\mathcal{P}_t$ and is instructed to select the intervention that maximally discriminates between competing hypotheses:

> *You have K competing SCM hypotheses. Design an intervention $\mathrm{do}(\mathbf{X} = \mathbf{x})$ that maximizes the disagreement among their predicted outcomes. Identify the variable(s) to intervene on and the value(s) to set, prioritizing interventions where different hypotheses make the most divergent predictions.*

Prediction. For each hypothesis $\mathcal{H}_k^t$, the LLM derives the predicted outcome using that hypothesis's structural equations and abduced latent states. The prompt enforces pre-commitment:

> *Using the structural equations and parameters of this hypothesis, and the latent variable values inferred during abduction, compute the predicted outcome of the intervention $\mathrm{do}(\mathbf{X} = \mathbf{x})$. Commit to a specific numerical prediction before observing the experimental result.*

Induction. The LLM receives each hypothesis's committed prediction alongside the observed outcome and is instructed to identify systematic patterns of failure:

> *Compare the predicted outcome with the experimental observation. Identify any systematic discrepancy and distill it into a concise correction rule describing the empirical pattern that the current hypothesis fails to capture (e.g., "the relationship between X and Y follows an inverse-square law rather than a linear law").*

Revision. The LLM receives the correction rules and is instructed to apply targeted structural edits. The prompt explicitly enumerates the available operators:

> *Based on the correction rule, revise the SCM hypothesis using one or more of the following operators: (1) Add/Remove Edge: insert or remove a directed edge; (2) Add/Remove Latent: introduce or remove a latent variable; (3) Update Mechanism: change the functional form of a structural equation; (4) Update Parameter: adjust the numerical value of an existing parameter. Apply the minimal set of operators sufficient to accommodate the correction rule. Prefer shallow revisions (parameter and mechanism updates) when possible, resorting to structural edits only when the evidence warrants them.*

Deduction. The LLM receives each revised hypothesis and performs two validation checks:

> *Validate the revised hypothesis: (1) Evidence Validation: derive predictions for all past interventions in the accumulated evidence and verify that the revised hypothesis accounts for previous observations without introducing new contradictions. (2) Consistency Check: verify that the causal graph is a valid DAG, all structural equations are well-defined given the revised graph, and the posited mechanisms are mutually compatible. Report whether the hypothesis passes both checks.*

C.3 PSEUDOCODE

Algorithm 1 presents the complete EvoSCM discovery loop. Each iteration comprises two phases: Causal Experiment Design (Section 3.2) selects and executes a discriminative intervention, and SCM Hypothesis Revision (Section 3.3) transforms the population through inductive correction, structured revision, and deductive validation. After T rounds, the evolved population $\mathcal{P}_T$ is returned for downstream inference (Section 3.4).

Algorithm 1 EvoSCM Discovery Loop

Require: Environment $\mathcal{E}$, budget T, population size K, LLM backbone $\mathcal{L}$
Ensure: Evolved population $\mathcal{P}_T$

Initialize $\mathcal{P}_0 = \{\mathcal{H}_1^0, \ldots, \mathcal{H}_K^0\}$
$\mathcal{D}_0 \leftarrow \emptyset$
for $t = 0$ **to** $T - 1$ **do**
  % Casual Experiment Design (§3.2)
  for $k = 1$ **to** $|\mathcal{P}_t|$ **do**
    $\hat{\mathbf{U}}_k^t \leftarrow \textsc{Abduction}(\mathcal{H}_k^t, \mathcal{D}_t, \mathcal{L})$
  end for
  $\mathbf{x}_{t+1}^* \leftarrow \textsc{Action}(\mathcal{P}_t, \mathcal{D}_t, \mathcal{L})$
  for $k = 1$ **to** $|\mathcal{P}_t|$ **do**
    $\mathbf{y}_k^{\text{pre}} \leftarrow \textsc{Prediction}(\mathcal{H}_k^t, \hat{\mathbf{U}}_k^t, \mathbf{x}_{t+1}^*, \mathcal{L})$
  end for
  $\mathbf{y}_{t+1}^{\text{obs}} \sim P^*(\mathbf{Y} \mid \text{do}(\mathbf{X}{=}\mathbf{x}_{t+1}^*))$
  $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\mathbf{x}_{t+1}^*, \mathbf{y}_{t+1}^{\text{obs}})\}$
  % SCM Hypothesis Revision (§3.3)
  $\mathcal{P}_{t+1} \leftarrow \emptyset$
  for $k = 1$ **to** $|\mathcal{P}_t|$ **do**
    $\mathcal{R}_k \leftarrow \textsc{Induction}(\mathcal{H}_k^t, \mathbf{y}_k^{\text{pre}}, \mathbf{y}_{t+1}^{\text{obs}}, \mathcal{D}_{t+1}, \mathcal{L})$
    $\tilde{\mathcal{H}}_k^{t+1} \leftarrow \textsc{Revision}(\mathcal{H}_k^t, \mathcal{R}_k, \mathcal{L})$
    $v_k \leftarrow \textsc{Deduction}(\tilde{\mathcal{H}}_k^{t+1}, \mathcal{D}_{t+1}, \mathcal{L})$
    if $v_k = \texttt{pass}$ **then**
      $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_{t+1} \cup \{\tilde{\mathcal{H}}_k^{t+1}\}$
    end if
  end for
end for
return $\mathcal{P}_T$

D Benchmark Details

D.1 DiscoverPhysics

Task definition and composition. DiscoverPhysics (Wiemann et al., 2026) is an interactive benchmark for discovering hidden laws of motion in simulated worlds whose physics deliberately departs from familiar physical assumptions. The complete benchmark comprises 22 curated worlds, including 11 publicly released worlds and 11 private worlds. Table 12 summarizes the public worlds, which cover unusual force laws, hidden sources and particle types, preferred spatial directions, and time-dependent interactions.

Experimental interface. A discovery session begins with initial trajectory observations and an experimental protocol, without revealing the underlying force law or hidden particle types. Each experiment specifies controllable particles through their initial positions, initial velocities, generalized charges, and requested measurement times. The simulator evolves these particles together with any existing particles and returns the requested positions and velocities. Observations are generated on demand rather than retrieved from a fixed dataset. The interface also supports fitting the free parameters of a proposed law against previously collected trajectories. Observation noise is configurable, whereas the held-out trajectories used for prediction evaluation are noise-free.

Evaluation. The final submission consists of a natural-language explanation and an executable Python implementation of the inferred dynamics. Predictive accuracy is evaluated on held-out particle trajectories. For a world w, let $\mathcal{I}_w$ denote the evaluated particle–time pairs, and let $\mathbf{r}_i(t)$ and $\widehat{\mathbf{r}}_i(t)$ denote the reference and predicted positions, respectively. The normalized trajectory error is

$$\text{nMSE}_w = \frac{1}{V_w |\mathcal{I}_w|} \sum_{(i,t)\in\mathcal{I}_w} \|\widehat{\mathbf{r}}_i(t) - \mathbf{r}_i(t)\|_2^2 , \tag{1}$$

Table 12: **Public worlds on DiscoverPhysics**. The table summarizes the principal discovery targets of the 11 publicly released worlds, rather than the full 22-world benchmark.

World	Principal discovery target
`gravity`	Two-dimensional attraction with force magnitude proportional to $1/r$.
`yukawa`	A screened interaction with exponential suppression at large distances.
`fractional`	An anomalous power-law interaction governed by a fractional operator.
`circle`	A fractional interaction observed through a multi-particle ring configuration.
`three_species`	Hidden particle classes with distinct attractive and repulsive couplings.
`dark_matter`	Additional forces generated by unobserved sources.
`ether`	A central interaction combined with a preferred-direction body force.
`hubble`	A central interaction combined with an outward, position-dependent contribution.
`oscillator`	A time-dependent coupling that periodically switches between attraction and repulsion.
`extra_dimensions`	A crossover from $1/r$ force scaling at large distances to $1/r^2$ scaling at short distances.
`coulomb_easy`	An attractive inverse-square central interaction.

where V_w is the benchmark's reference trajectory variance for that world. Normalization accounts for differences in the natural motion scales of different worlds, and trajectory errors are aggregated geometrically across worlds.

Explanation quality is evaluated separately by an LLM judge using an expert-written, world-specific rubric, with scores reported on a normalized scale. Discovery success requires both sufficient predictive accuracy and sufficient explanation quality. The $\mathrm{pass@}k$ metric measures the expected fraction of worlds with at least one successful discovery among k independent attempts.

D.2 LLEMA

Task definition and composition. LLEMA (Abhyankar et al., 2026) introduces a suite of 14 application-driven materials-discovery tasks, publicly released as LLEMABench. Here, LLEMA refers to this benchmark suite, rather than the evolutionary method introduced in the same work. Each task specifies an application goal together with numerical property constraints and, where applicable, compositional requirements. The objective is to identify candidate materials satisfying these requirements, rather than to recover a unique reference equation or causal graph. The released task specifications are summarized in Table 13.

Candidate representation and property evaluation. Candidates are represented by chemical compositions and crystallographic information files (CIFs), which describe lattice geometry, symmetry, and atomic sites. The evaluation pipeline combines available Materials Project data with pretrained property predictors, including CGCNN and ALIGNN, to obtain task-relevant material properties. These include electronic, mechanical, dielectric, and thermodynamic quantities, with the required property set determined by the task. Candidate validity is established by checking the resulting properties and composition against the task specification.

The benchmark therefore evaluates computationally screened materials. Its property estimates and stability assessments should not be interpreted as direct experimental measurements or confirmation of successful synthesis.

Evaluation. Hit rate measures the percentage of generated candidates satisfying all task-specific constraints. The stability metric measures the percentage that are both task-valid and classified as thermodynamically stable. For M generated candidates, let $v_i \in \{0, 1\}$ indicate whether candidate i satisfies the task constraints, and let $s_i \in \{0, 1\}$ indicate whether it passes the stability screen. The two metrics, expressed as percentages, are

$$\text{H.R.} = \frac{100}{M} \sum_{i=1}^{M} v_i, \qquad \text{Stab.} = \frac{100}{M} \sum_{i=1}^{M} v_i s_i. \tag{2}$$

Table 13: **Materials-discovery tasks on LLEMA.** Core property targets follow the publicly released task descriptions. The table summarizes task-level targets; thermodynamic stability is additionally assessed by the stability evaluator.

Task	Core property targets
Hard, stiff ceramics	$100 \le K \le 300$; $60 \le G \le 200$.
Hard coating materials	$K \ge 200$; $E_f \le -1$; $E_g \ge 3$.
Structural materials for aerospace	$\rho \le 5$; $K \ge 100$; $G \ge 40$.
High-k dielectrics	$10 \le \kappa \le 90$; $2.5 \le E_g \le 6.5$.
SAW/BAW acoustic substrates	$25 \le G \le 150$; $3.7 \le \kappa \le 95$.
Solid-state electrolytes†	$E_f \le -1$; $E_g \ge 2$.
Stable wide-bandgap semiconductors	$E_g \ge 2.5$; $E_f \le -1$.
Photovoltaic absorbers‡	$0.7 \le E_g \le 2$; $E_f \le 0$.
Piezo energy harvesters	$d \ge 8$; $10 \le \kappa \le 8000$.
Acousto-optic hybrids	$3 \le d \le 9$; $8 \le \kappa \le 85$.
Electrically insulating dielectrics	$E_g \ge 2.5$; $\kappa \ge 8$.
Transparent conductors	$E_g > 3$; $50 \le \sigma \le 5000$.
Low-density structural materials	$\rho \le 3.5$; $65 \le G \le 195$.
Toxic-free perovskite oxides*	$E_g \ge 2$; $90 \le K \le 135$.

K and G: bulk and shear moduli in GPa; E_g: band gap in eV; E_f: formation energy in eV/atom; ρ: density in $\mathrm{g\,cm^{-3}}$; κ: dimensionless dielectric constant; d: piezoelectric proxy in $\mathrm{pC\,N^{-1}}$; σ: electrical conductivity in $\mathrm{S\,cm^{-1}}$. For the two piezoelectric tasks, d and κ denote the corresponding proxies in the released specification. †Requires at least one of Li, Na, K, Mg, Ca, or Al. ‡Restricted to earth-abundant elements satisfying the benchmark's non-toxicity requirement. *Excludes Pb, Cd, Hg, Tl, Be, As, Sb, Se, U, and Th, and favors stable ABO_3 oxides.

Thus, the stability metric uses all generated candidates as its denominator, rather than only the valid subset. The original benchmark reports an energy-above-hull threshold of 0.1 eV/atom for thermodynamic stability screening. It additionally uses Pareto-front analysis to assess trade-offs among competing material properties.

D.3 ACTIVESCIBENCH-GRN

Task definition and composition. ActiveSciBench-GRN, introduced with LLM-AutoSciLab (Kabra et al., 2026), evaluates the recovery of hidden gene-regulatory networks through active perturbation experiments. The reference suite comprises five regulatory motif families, three benchmark versions, and three difficulty levels, yielding $5 \times 3 \times 3 = 45$ task configurations per random seed. Table 14 summarizes the motif families. The discovery target is a signed directed graph specifying which regulatory interactions exist, their directions, and whether they are activating or repressing. The inclusion of feedback motifs means that the target graphs are not restricted to directed acyclic graphs.

Experimental interface. An experiment sets an upstream signal and multiplicative perturbations of nodes A, B, C, and R. Perturbation factors below or above one represent knockdown or overexpression, respectively; one denotes the reference setting. The public simulator returns a reporter response and a marker panel for these nodes. Responses are generated by motif-specific nonlinear regulatory mechanisms rather than retrieved from a fixed observational dataset.

Difficulty levels. Difficulty is controlled through nonlinear response regimes. Easy instances exhibit approximately linear responses; medium instances introduce saturation that can obscure weak regulatory effects; and hard instances emphasize stronger nonlinearities and, for the relevant motifs, feedback-dependent switching behavior. The governing graph and dynamical parameters are withheld from the learner.

Evaluation. Let $\mathbf{A}^{\star}, \widehat{\mathbf{A}} \in \{-1, 0, +1\}^{d \times d}$ denote the reference and predicted signed adjacency matrices, where $+1$, -1, and 0 represent activation, repression, and absence of an edge, respec-

Table 14: **Regulatory motif families in ActiveSciBench-GRN.** Each family has three benchmark versions and three difficulty levels.

Motif family	Regulatory structure
Activation chain	A sequential activation cascade from an upstream signal to downstream regulators.
Coherent feedforward loop	Direct and indirect activating paths with consistent regulatory effects.
Incoherent feedforward loop	Competing paths combining activation and repression.
Negative-feedback circuit	Downstream activation induces a repressive branch that limits the response.
Toggle-switch circuit	Mutual repression supporting switching and state selection.

tively. Define the corresponding signed edge sets $\mathcal{E}^\star$ and $\widehat{\mathcal{E}}$, whose elements are $(\text{source}, \text{target}, \text{sign})$ triples. The public evaluator computes

$$P = \frac{|\widehat{\mathcal{E}} \cap \mathcal{E}^\star|}{|\widehat{\mathcal{E}}|}, \qquad R = \frac{|\widehat{\mathcal{E}} \cap \mathcal{E}^\star|}{|\mathcal{E}^\star|}, \qquad F_1 = \frac{2PR}{P+R}, \tag{3}$$

with undefined ratios assigned zero. Thus, an edge is counted as correct only when its direction and sign both match.

Sign accuracy measures sign agreement over directed edges present in both graphs and is zero when this overlap is empty. Exact graph accuracy requires equality of the complete signed adjacency matrices. The evaluator additionally provides motif accuracy, measuring whether the predicted motif-family label is correct.

E Experiment Details

E.1 DiscoverPhysics

Mapping SCM variables to physical quantities. The SCM maintained by EvoSCM is an *epistemic* model of the environment, not a copy of the simulator's ground-truth causal program. For each hypothesis $\mathcal{H}_k^t = (\mathbf{V}, \mathbf{U}_k^t, G_k^t, \mathbf{F}_k^t, \mathbf{\Theta}_k^t)$, the variable set is partitioned as

$$\mathbf{V}_k^t = \mathbf{V}_{\text{obs}} \cup \mathbf{V}_{\text{act}} \cup \mathbf{V}_{\text{lat},k}^t \cup \mathbf{V}_{\text{der},k}^t.$$

In two-particle environments, $\mathbf{V}_{\text{obs}}$ contains particle positions $\mathbf{x}_t$, velocities $\mathbf{v}_t$, and the observed trajectory; $\mathbf{V}_{\text{act}}$ contains the public control scalars p_1 and p_2, the initial position and velocity of the mobile particle, and, when available, the absolute start time. The hypothesis may introduce derived quantities $\mathbf{V}_{\text{der},k}^t$ such as the separation r_t, radial direction $\hat{\mathbf{r}}_t$, field gradient $\nabla\phi_t$, and acceleration $\mathbf{a}_t$. The semantic roles of p_1 and p_2 are not hard-coded: competing hypotheses may interpret them as source strength, response inertia, charge, or nuisance controls, and these interpretations are resolved through interventions.

The latent variables $\mathbf{V}_{\text{lat},k}^t$ represent unobserved physical causes proposed by the LLM, such as hidden source locations, particle species, screening mechanisms, or time-dependent couplings. Hypothesis-specific parameters $\mathbf{\Theta}_k^t$ include quantities such as a coupling constant G, radial exponent q, screening length λ, and numerical regularization ϵ. These variables and parameters are inferred from observations; they are not populated from the benchmark's hidden ground truth.

For population environments, the observed variables instead contain positions and velocities of visible particles and probes. Action variables include probe positions, velocities, and masses, depending on the environment topology. Hidden dark sources and latent species assignments may appear in the SCM but are not directly manipulable by the agent.

Implementation of interventions. An intervention $\text{do}(\mathbf{X} = \mathbf{x})$ is implemented by assigning one or more fields of the public experiment schema before the trajectory is generated. EvoSCM cannot intervene on hidden physical laws, hidden simulator parameters, or unexposed latent state. Table 15 summarizes the available action spaces by environment topology.

Table 15: Public intervention variables in DiscoverPhysics. Measurement schedules determine when the trajectory is observed and are not treated as physical causes.

Topology	Manipulable inputs
Two-particle	p_1, p_2, initial position $\mathbf{x}_0$ and velocity $\mathbf{v}_0$ of the mobile particle, measurement schedule.
Time-modulated	Two-particle inputs plus the absolute `start_time`.
Circle	Ring radius, initial tangential velocity, measurement schedule.
Dark matter / three species	Initial positions and velocities of neutral probes, measurement schedule. Hidden sources and species assignments cannot be set.
Ether / Hubble	Probe positions, velocities, masses, measurement schedule.

EvoSCM supports both ordinary and paired-counterfactual experiments. For an ordinary intervention, the selected input is passed directly to the simulator. For a paired counterfactual, the agent specifies a factual input $\mathbf{x}$ and a set of sparse assignments $\{\delta_j\}$. The runtime evaluates

$$\mathbf{y}^{(0)} = \mathrm{Sim}(\mathbf{x}), \qquad \mathbf{y}^{(j)} = \mathrm{Sim}\big(\mathrm{do}_{\delta_j}(\mathbf{x})\big),$$

and returns both branch outputs and paired differences $\mathbf{y}^{(j)} - \mathbf{y}^{(0)}$. All branches share the same hidden environment and a restored random-number state, so paired differences use common random numbers and isolate the declared intervention's effect. The factual rollout and each counterfactual branch consume one simulator episode.

At each round, the LLM proposes two to four candidate experiment designs together with pre-committed predictions from the competing hypotheses. EvoSCM executes only the feasible design with the largest posterior-weighted predictive disagreement, ensuring that interventions are selected using predictions made before observing the corresponding outcome.

Population size, rounds, and backbones. The maximum population size is $K_{\max} = 6$. At the first round, the LLM initializes between three and six structurally distinct hypotheses with uniform evidence weights, irrespective of any confidence values reported by the LLM, to avoid favoring a familiar textbook law before any experiment has been performed.

Each session runs for at most $T_{\max} = 16$ LLM interaction rounds and may terminate earlier when the agent submits its final executable law. Since one counterfactual design can contain multiple simulator branches, the number of LLM rounds and simulator episodes are recorded separately. In paired comparisons, EvoSCM uses the same world, noise seed, LLM backbone, explanation judge, and maximum round count as the matched baseline session. Its simulator-episode budget is additionally capped by the number of episodes consumed by that baseline session.

We evaluate with three LLM backbones: GPT-5.4 (OpenAI, 2026a), GPT-5.5 (OpenAI, 2026b), and GPT-5.6-Sol (OpenAI, 2026c). The maximum output length is 8,192 tokens per call. Evaluation results and reference explanations are not exposed to the discovery agent.

Prompt construction. The system prompt is assembled from three components: (i) the benchmark-wide interactive discovery template, (ii) a topology-specific instruction file, and (iii) the EvoSCM protocol. The topology-specific instructions are inserted into the master template, after which the EvoSCM protocol is appended. At each round, the runtime supplies a compact JSON state containing the current hypothesis population, posterior weights, recent prediction errors, the latest abduction, induction and deduction records, completed intervention coverage, and remaining budget. The LLM returns structured blocks: `<scm_update>` for abduction and hypothesis revision, `<design_experiments>` for candidate interventions and pre-committed predictions, and `<final_law>`, `<explanation>`, `<scm_final>` at termination. Complete prompt templates and schemas are included in the released code.

Physics-specific adaptations. The EvoSCM loop is unchanged across environments, but the following adaptations tailor evidence collection to physical-law discovery.

Causal-invariance audits. For two-particle environments, the runtime tracks a set of systematic audits rather than trusting the most familiar initial law. These include a null-source intervention,

Table 16: Mapping from LLEMA tasks to target property variables.

Task	Outcomes in $\mathbf{Y}_\tau$
Hard, Stiff Ceramics	Bulk modulus; shear modulus
Hard Coating Materials	Bulk modulus; formation energy; band gap
Aerospace Materials	Density; bulk modulus; shear modulus
High-k Dielectrics	Dielectric constant; band gap
SAW/BAW Acoustic Substrates	Shear modulus; dielectric constant
Solid-State Electrolytes	Formation energy; band gap
Wide-Bandgap Semiconductors	Band gap; formation energy
Photovoltaic Absorbers	Band gap; formation energy
Piezo Energy Harvesters	Piezoelectric coeff.; dielectric response
Acousto-optic Hybrids	Piezoelectric coeff.; dielectric response
Insulating Dielectrics	Band gap; dielectric constant
Transparent Conductors	Band gap; electrical conductivity
Low-Density Structures	Density; shear modulus
Perovskite Oxides	Band gap; bulk modulus

controlled magnitude changes in p_2, sign reversals of p_1 and p_2, a multi-scale radius scan (at least four noise-cancelled responses spanning at least an eight-fold radius range), and, when absolute time is exposed, a phase sweep over `start_time`. The scale audit allows the agent to distinguish a constant power law from Yukawa screening or a short-distance dimensional crossover.

Paired counterfactuals for noise cancellation. Paired counterfactuals cancel common observation noise and estimate local causal responses. The runtime derives empirical quantities (source and response-control exponents, local radial slopes, sign relations, absolute-time responses) exclusively from experiments observed by the agent. These quantities are returned as evidence summaries; no hidden world label or ground-truth equation is used.

Topology-specific action schemas. For population environments, EvoSCM reasons about latent source populations through controlled probe placements, velocities, and masses rather than intervening on hidden particles. For the circle environment, experiments vary collective ring-level initial conditions instead of individual two-particle controls.

Parameter calibration. Continuous parameters can be calibrated using the benchmark's trajectory-level MSE-fitting tool on trajectories collected during discovery. The final output is an executable Python law, and the runtime checks its implied direction, control scaling, radial dependence, and time dependence against observation-derived causal evidence before accepting the submission.

E.2 LLEMA

Mapping SCM variables to material properties. For each task τ, EvoSCM maintains an epistemic SCM over

$$\mathbf{V}_\tau = \mathbf{X}_{\mathrm{mat}} \cup \mathbf{Y}_\tau,$$

where $\mathbf{X}_{\mathrm{mat}}$ contains composition and structure descriptors and $\mathbf{Y}_\tau$ contains the target properties required by the task. The causal direction is fixed by domain knowledge:

$$X_j \longrightarrow Y_{\tau,q}, \qquad X_j \in \mathbf{X}_{\mathrm{mat}}, \quad Y_{\tau,q} \in \mathbf{Y}_\tau.$$

EvoSCM therefore does not search over arbitrary causal DAGs in this domain. Instead, it evolves the active descriptor set, edge signs and magnitudes, uncertainty estimates, and the relative support for competing local mechanisms.

The descriptor variables include structural quantities (volume per atom, density, lattice anisotropy, minimum pair distance, nearest-neighbor statistics), compositional quantities (composition entropy, elemental fractions, electronegativity range), and atomic-level statistics (mean and standard deviation of atomic number, electronegativity, mass, and radius). All descriptors are deterministically computed from the generated crystal using PYMATGEN. For every task, energy above hull (E_{hull}) is included as an additional stability outcome even when it is not a primary objective. Table 16 shows the task-to-property mapping.

SCM hypothesis population. We maintain $K = 7$ competing SCM hypotheses (not to be confused with the number of material proposals). Each hypothesis is a bootstrap sparse ridge mechanism fitted independently for each property outcome. Features are standardized and screened per property, with a ridge coefficient of 2.0. Bootstrap resampling produces seven competing signed mechanisms; their median coefficient, sign agreement, and coefficient variance define the ensemble-level edge estimate and uncertainty.

The property-wise mechanisms are complemented by three additional components: (i) a calibrated LLM chemistry prior, (ii) non-parametric local mechanisms for the binary hit and stability outcomes, and (iii) a surrogate-support counterexample frontier. The numerical SCM starts with a reliability prior of 0.15, while the LLM chemistry prior is initialized at 0.25. Reliability is updated using pre-committed prediction error and parent-child delta-sign accuracy. Outcome-specific edges are withheld from the planner when their calibrated reliability falls below 0.20, allowing the system to retain a partially specified SCM rather than forcing unsupported edges.

From material design to SCM interventions. EvoSCM does not directly apply interventions such as $\mathrm{do}(\texttt{halogen_fraction} = x)$, because modifying a material generally changes several descriptors simultaneously. Instead, the action variable is a chemically explicit transformation

$$a_t : m_{\mathrm{parent}} \longrightarrow m_{\mathrm{child}},$$

where each proposed action records its type, changed factor, source and target values, and parent formula, making the intervention auditable. Supported transformations include:

- elemental substitutions (same-group, oxidation-state-preserving, or chemically similar);
- stoichiometric and redox-aware composition changes;
- crystal-prototype-preserving or prototype-changing interventions;
- canonical polymorph realization at fixed composition;
- isotropic lattice strain, polar displacement, and coordination changes at fixed composition;
- stability-boundary, Pareto-frontier, and controlled-disagreement experiments.

A deterministic candidate compiler converts each intervention intention into an evaluable crystal. The compiler can normalize the formula, ground a composition in a common prototype, generate bounded structural counterfactuals, and repair geometry errors. It never queries the property oracle. Each compiled candidate retains its source and all repair actions, so compiler transformations remain distinct from oracle-derived evidence.

Population size, rounds, and budget. We use $T = 5$ adaptive rounds. A shared initialization requests eight candidates from the LLM and evaluates four diverse candidates. In each subsequent round, the LLM proposes $K_{\mathrm{proposal}} = 8$ materials, the compiler constructs at most $K_{\mathrm{compiled}} = 40$ pre-oracle candidates, and the policy evaluates $B = 4$ candidates. Each task-seed session therefore uses $B_0 + TB = 4 + 5 \times 4 = 24$ oracle episodes and five adaptive LLM calls. Compiler-generated candidates that are not selected do not consume oracle budget.

The active memory retains at most six success and six failure exemplars in the prompt, while the fitted SCM uses all valid observations. Five deterministic island labels provide provenance and diversity tracking. The formula exclusion memory contains at most 96 previously tested formulas.

LLM Backbones. The main comparison uses Qwen3.6-35B-A3B (Qwen Team, 2026) with reasoning mode disabled, a maximum output length of 12,288 tokens, and strict JSON schema enforcement. The transfer experiment uses GPT-5.6-Luna (OpenAI, 2026c) with the same configuration: reasoning mode disabled, the same token limit, and the same budget.

Property oracle interface. After structural validation, each candidate is serialized as a CIF and evaluated on the properties required by the current task plus E_{hull}. All methods use a single frozen ALIGNN oracle version. Table 17 lists the property models.

A candidate is a *hit* if it is structurally valid and satisfies every task-specific constraint. It is *stable* if $E_{\mathrm{hull}} < 0.1\ \mathrm{eV/atom}$, and a *stable hit* if both conditions hold. For Transparent Conductors and

Table 17: ALIGNN property models used for oracle evaluation.

Property	**Model identifier**
Band gap	`jv_optb88vdw_bandgap_alignn`
Formation energy	`jv_formation_energy_peratom_alignn`
Energy above hull	`jv_ehull_alignn`
Bulk modulus	`jv_bulk_modulus_kv_alignn`
Shear modulus	`jv_shear_modulus_gv_alignn`
Dielectric response	`jv_epsx_alignn`
Piezoelectric (max d_{ij})	`jv_dfpt_piezo_max_dij_alignn`
Piezoelectric (dielectric)	`jv_dfpt_piezo_max_dielectric_alignn`
Density	Deterministic (from crystal geometry)
Conductivity	Derived: $\sigma = \mathrm{PF}/S^2$ (S/cm)

Low-Density Structures, stability is included in the hit definition per the task specification. H.R. is the fraction of evaluated candidates that are hits; Stab. is the fraction that are stable hits.

Oracle results include the property vector, failed constraints, and cache provenance. Cache keys depend on the CIF bytes, property model, and oracle version. Invalid structures receive no oracle call, and every prediction is stored before its outcome is observed.

Domain-specific adaptations. The EvoSCM loop is shared across all 14 tasks, but its generic intervention language is mapped to materials-specific actions through five frozen adapters.

1. Prototype grounding. Recognized compositions (AB, AB_2, ABX_3, A_3BX) are grounded in common crystal prototypes (rocksalt, zincblende, wurtzite, rutile, fluorite, perovskite, etc.). Piezoelectric ABX_3 hypotheses use a bounded polar tetragonal realization rather than centrosymmetric cubic.

2. Stability-aware portfolio allocation. The four evaluated slots per round are allocated to a chemistry-anchor candidate, a joint-feasibility hypothesis, a counterexample-repair or structural-counterfactual experiment, and a stability/Pareto experiment.

3. Fixed-composition structural counterfactuals. Stable hits and near misses can undergo bounded isotropic strain, prototype realization, coordination changes, or polar displacement while keeping composition fixed.

4. Property-family adapters. Band-gap/dielectric tasks use a bounded strain ladder from a stable near miss; acoustic tasks compare canonical realizations at fixed composition; piezoelectric tasks prioritize polar-displacement interventions; density/modulus tasks include explicit Pareto and objective-boundary probes.

5. Calibration controls. A frozen library of textbook structures (e.g., $LiMgF_3$/$LiCaF_3$ for electrolyte calibration, ZnO/AlN/$LiNbO_3$ for piezoelectric calibration, MgO/CaF_2 for dielectric calibration) supplies positive, negative, and boundary controls. These consume one oracle slot, are recorded as calibration interventions, and are not presented as novel discoveries.

All adapters are deterministic, frozen in the released code, and operate before oracle evaluation. They may use public chemistry, composition, geometry, task constraints, and previous observations, but never an unobserved oracle outcome.

E.3 ACTIVESCIBENCH-GRN

Mapping SCM variables to genes. EvoSCM represents each epistemic hypothesis as a steady-state SCM over the variable set

$$\mathbf{V} = \{S, X_A, X_B, X_C, X_R\},$$

where S is an externally controlled signaling input and X_A, X_B, X_C, X_R are the steady-state expression levels of four abstract genes. Gene C is the reporter gene: the scalar reporter intensity returned by the simulator is a scaled readout of X_C. In addition to this reporter, the agent observes the complete marker panel (X_A, X_B, X_C, X_R) after each experiment.

A candidate SCM contains a signed directed graph $G_k = (\mathbf{V}, E_k, \sigma_k)$, where $\sigma_{ij} = +1$ denotes activation and $\sigma_{ij} = -1$ denotes repression. Local mechanisms are represented using signed Hill-type functions. For each endogenous gene,

$$X_j = m_j \left[b_j + a_j \prod_{i \in \mathrm{Pa}_{G_k}(j)} H_{\sigma_{ij}}(X_i;\, K_{ij},\, n_{ij}) \right], \qquad j \in \{A, B, C, R\},$$

where H_{+1} and H_{-1} denote activating and repressing Hill functions, respectively. Feedback and toggle motifs are interpreted as equilibrium solutions of potentially cyclic regulatory graphs rather than DAGs. The fitted parameters are epistemic approximations used for prediction and model discrimination; they are not assumed to reproduce the simulator's hidden kinetic parameters.

Mapping interventions to gene perturbations. An experimental action is defined as

$$\mathbf{u} = (s,\, m_A,\, m_B,\, m_C,\, m_R),$$

with a continuous signal level $s \in [0.05, 10]$ and continuous perturbation multipliers $m_j \in [0.1, 4]$. A multiplier $m_j = 1$ denotes the unperturbed condition, $m_j < 1$ represents knockdown, and $m_j > 1$ represents overexpression. These are *soft interventions*: the agent applies $\mathrm{do}(m_j = c)$, which scales the production of gene j, rather than directly fixing its expression through $\mathrm{do}(X_j = c)$.

At each round, the intervention selector combines LLM-proposed experiments with 512 log-uniformly sampled points from the legal intervention space. It selects four experiments that maximize predictive disagreement among the current SCM hypotheses, with a diversity term that discourages redundant interventions. Each hypothesis's reporter predictions are committed before the selected experiments are submitted to the simulator.

Population size, rounds, and backbones. Each session begins with $N_0 = 4$ shared log-uniform initial experiments. EvoSCM then runs $T = 4$ active-discovery rounds, each proposing $K = 5$ structurally diverse candidate SCMs and up to five candidate interventions. After numerical fitting and graph deduplication, the best five distinct candidates form the working population. Four new experiments are executed per round, giving a total budget of $N_{\mathrm{exp}} = N_0 + T \times 4 = 20$. One session therefore makes four LLM calls and twenty simulator calls.

We evaluate two backbone configurations. The local-model experiments use Qwen3.6-35B-A3B (Qwen Team, 2026) with reasoning mode disabled, temperature 0.7, top-$p = 0.8$, top-$k = 20$, and a maximum output length of 4,096 tokens. The API-model experiments use GPT-5.6-Luna (OpenAI, 2026c) with a maximum output length of 4,096 tokens. Both configurations share the same K, T, intervention bounds, initial-design size, and simulator budget.

The fixed working population size $K = 5$ is distinct from the candidate archive maintained by the final selector. The archive retains every valid unique graph produced during the four rounds and therefore has a data-dependent size.

Relation between the epistemic SCM and the target network. The target regulatory network G^* and each epistemic candidate G_k share the same labeled node set and signed-edge representation. A candidate edge $A \xrightarrow{+} B$ directly expresses the hypothesis that A activates B, while $R \xrightarrow{-} C$ hypothesizes repression of the reporter gene. The predicted graph can therefore be compared directly against the hidden signed adjacency matrix of G^*. The agent receives a public universal hypothesis grammar describing all five possible motif families, together with the intervention interface and graph constraints. It is not given the current session's motif identity, ground-truth edges, kinetic parameters, evaluator scores, or task identifier. Candidate graphs may contain at most six signed edges, cannot contain self-loops, cannot use S as a destination, and must include a directed path from S to C. The same grammar is supplied to EvoSCM and the baseline. This experiment therefore evaluates active discrimination, revision, and selection within a public, biologically motivated hypothesis class; it is not unconstrained de novo topology discovery.

GRN-specific final selection. After the fourth round, EvoSCM retains all valid unique graph candidates encountered during the session. For each archived graph, it fits node-conditional signed Hill mechanisms using only the observed interventions and marker values. Candidates are ranked

by five-fold grouped cross-validation using the mean held-out squared log error over $\{A, B, C, R\}$. Ties are resolved first by preferring the graph with fewer edges, then by a deterministic canonical graph ordering. This selector uses neither the target graph nor evaluator metrics, and makes no additional LLM calls or simulator experiments. The selected graph is frozen before the evaluator is invoked and is used unchanged for all graph-recovery metrics. Exact Graph Accuracy requires equality of the complete signed adjacency matrix; Edge F1 is computed over signed edge triples; and Sign Accuracy measures sign agreement on edges present in both the prediction and target.

F YUKAWA WORLD DETAILS

This section relates the ground-truth law used by the Yukawa world to the standard continuum Yukawa theory and clarifies the three stages of the SCM evolution shown in Figure 4. The apparent difference from the commonly quoted Yukawa potential, $\exp(-r/\lambda)/r$, is caused by dimensionality: that expression is the Green function in three spatial dimensions, whereas the DiscoverPhysics Yukawa world is two-dimensional.

Continuum field equation. Let $\mathbf{x}_1$ be the fixed source position, let $\mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1$, $r = \|\mathbf{r}\|$, and $\widehat{\mathbf{r}} = \mathbf{r}/r$. With $m = \lambda^{-1}$, the static screened field is governed by the modified Helmholtz equation

$$(\nabla^2 - m^2)\phi(\mathbf{x}) = G p_1 \, \delta^{(d)}(\mathbf{x} - \mathbf{x}_1), \tag{4}$$

where G is the field coupling, p_1 is the source strength, and d is the number of spatial dimensions. Fourier transformation gives

$$\widetilde{\phi}(\mathbf{k}) = -\frac{G p_1}{\mathbf{k}^2 + m^2}. \tag{5}$$

Consequently, the d-dimensional Green function is

$$\mathcal{G}_d(r) \equiv \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{\mathbf{k}^2 + m^2} \tag{6}$$

$$= \frac{1}{(2\pi)^{d/2}} \left(\frac{m}{r}\right)^{d/2-1} K_{d/2-1}(mr), \tag{7}$$

where K_ν denotes the modified Bessel function of the second kind. The continuum potential is $\phi_d(r) = -G p_1 \mathcal{G}_d(r)$.

For $d = 3$, using $K_{1/2}(z) = \sqrt{\pi/(2z)} e^{-z}$ recovers the familiar expression

$$\phi_3(r) = -\frac{G p_1}{4\pi} \frac{e^{-r/\lambda}}{r}. \tag{8}$$

Thus, the usual exponentially screened inverse-distance potential is the three-dimensional member of the same family, rather than the appropriate ground truth for the present two-dimensional world.

Two-dimensional Yukawa potential and force. For $d = 2$, Equation 7 instead gives

$$\mathcal{G}_2(r) = \frac{1}{2\pi} K_0(mr), \qquad \phi_2(r) = -\frac{G p_1}{2\pi} K_0(r/\lambda). \tag{9}$$

The mobile particle has response/inertia parameter p_2 and obeys

$$\dot{\mathbf{x}}_2 = \mathbf{v}_2, \qquad \dot{\mathbf{v}}_2 = \mathbf{a}(\mathbf{r}) = -\frac{1}{p_2} \nabla \phi_2(r). \tag{10}$$

Using $\mathrm{d}K_0(z)/\mathrm{d}z = -K_1(z)$ yields

$$\boxed{\mathbf{a}(\mathbf{r}) = -G \frac{p_1}{p_2} \frac{K_1(r/\lambda)}{2\pi\lambda} \widehat{\mathbf{r}}}. \tag{11}$$

Equation 11 is therefore not an unrelated discrete empirical law: it is the radial force obtained directly by taking the gradient of the continuum two-dimensional Yukawa potential.

Operational ground truth in the simulator. The N-body backend regularizes close encounters with the softened radius

$$\rho = \sqrt{r^2 + \epsilon^2}. \tag{12}$$

It evaluates the radial kernel at ρ while retaining the physical radial direction $\widehat{\mathbf{r}}$. The operational law that generates the observed trajectories is therefore

$$\boxed{\mathbf{a}_{\text{sim}}(\mathbf{r}) = -G\frac{p_1}{p_2}\frac{K_1(\rho/\lambda)}{2\pi\lambda}\widehat{\mathbf{r}}, \qquad \rho = \sqrt{r^2+\epsilon^2}}. \tag{13}$$

The benchmark parameters are $G = 1$, $\lambda = 2$, and $\epsilon = 0.05$. The replacement $r \mapsto \rho$ is a simulator-level numerical regularization, not an additional prediction of the unregularized continuum Helmholtz equation. In particular, Equation 13 approaches the continuum law in Equation 11 when $r \gg \epsilon$, while its magnitude remains finite when $r \lesssim \epsilon$. The direction floor $\delta = 10^{-6}$ used in the learned executable law is likewise a numerical safeguard of the agent implementation and is not a physical ground-truth parameter.

The simulator advances the continuous state $(\mathbf{x}_t, \mathbf{v}_t)$ using a fourth-order Yoshida symplectic integrator with step size $\Delta t = 0.005$. Hence, the time-unrolled arrows $\mathbf{a}_t \to \mathbf{v}_{t+1} \to \mathbf{r}_{t+1}$ in Figure 4 represent the numerical state-transition layer, whereas Equation 13 is the scientific mechanism recovered by the epistemic SCM.

Interpretation of the SCM evolution. The three stages in Figure 4 follow directly from distinct asymptotic regimes of the same operational law.

First, for $\epsilon \ll r \ll \lambda$,

$$K_1(z) = \frac{1}{z} + \mathcal{O}(z \log z), \tag{14}$$

and therefore

$$\mathbf{a}_{\text{sim}}(\mathbf{r}) \simeq -G\frac{p_1}{p_2}\frac{1}{2\pi r}\widehat{\mathbf{r}}. \tag{15}$$

The initial unscreened inverse-radius SCM is thus a valid local approximation to the two-dimensional Yukawa force, rather than an arbitrary incorrect prior.

Second, for $r \gg \lambda$,

$$K_1(r/\lambda) \simeq \sqrt{\frac{\pi\lambda}{2r}}e^{-r/\lambda}, \tag{16}$$

so a sufficiently broad radius sweep reveals the exponentially suppressed tail. This evidence motivates the SCM revision that introduces the screening length λ and the K_1 mechanism.

Finally, for $r \lesssim \epsilon$, $\rho \simeq \epsilon$, and the force magnitude ceases to follow the divergent $1/r$ near-field approximation. Targeted sub-core interventions therefore identify the additional causal path

$$(r, \epsilon) \longrightarrow \rho \longrightarrow K_1(\rho/\lambda) \longrightarrow \mathbf{a}, \tag{17}$$

which explains the finite-core revision in the final SCM. This last step is a discovery of the simulator's operational regularization, whereas the preceding K_0/K_1 structure is the recovery of the underlying continuum two-dimensional Yukawa mechanism. Accordingly, the final model in Figure 4 is ground-truth-equivalent for the evaluated N-body environment while retaining a clear correspondence to the continuum Yukawa field theory.

G Additional Related Work

G.1 Self-Evolving LLM Agents

Self-evolving LLM agents can improve from experience without updating their parameters (Wang et al., 2024a; Tao et al., 2024; Fang et al., 2025), spanning approaches such as prompt optimization, memory distillation, skill acquisition, workflow search, and multi-agent collaboration.

Prompt and Strategy Optimization. OPRO (Yang et al., 2024) frames prompt optimization as a black-box optimization problem, using LLMs to iteratively generate and refine prompts based on task performance. PromptBreeder (Fernando et al., 2024) applies an evolutionary algorithm to co-evolve task prompts and mutation operators, demonstrating that self-referential improvement can outperform hand-crafted prompt engineering. EvoPrompt (Guo et al., 2024) similarly evolves discrete prompts using genetic algorithms and differential evolution, connecting prompt optimization to classical evolutionary computation. These methods optimize the instructions given to the LLM but do not modify the agent's model of the external world, which is the focus of EvoSCM.

Reflective and Experiential Learning. Reflexion (Shinn et al., 2023) introduces verbal reinforcement learning, where agents generate natural-language reflections on failures and store them in an episodic memory buffer to avoid repeating mistakes. ExpeL (Zhao et al., 2024) extracts transferable experiential insights from successful and failed trajectories, building a growing library of lessons that generalize across tasks. Self-Refine (Madaan et al., 2023) demonstrates that LLMs can iteratively critique and improve their own outputs through multi-round feedback, without any external training signal. These systems accumulate experiential knowledge that improves decision-making, but the accumulated knowledge takes the form of verbal reflections, heuristic rules, or trajectory summaries rather than executable scientific models with falsifiable predictions.

Multi-Agent and Workflow Evolution. MetaGPT (Hong et al., 2024) organizes multiple LLM agents into structured workflows with role-based specialization, demonstrating that procedural organization improves complex task performance. AutoGen (Wu et al., 2023) provides a general-purpose framework for multi-agent conversations with customizable interaction patterns. AgentVerse (Chen et al., 2023) explores how groups of agents can collaboratively evolve their problem-solving strategies through dynamic team composition. These approaches evolve the agent's organizational structure and communication protocols, a form of procedural self-evolution that is complementary to EvoSCM's epistemic self-evolution.

G.2 Scientific Agents

A separate body of work applies LLM agents to scientific research (AI4Science & Quantum, 2023; Wei et al., 2025; Ren et al., 2026), spanning hypothesis generation, experiment design, result interpretation, and domain-specific applications across physics, chemistry, and biology.

Laboratory and Experimental Agents. ChemCrow (M. Bran et al., 2024) augments LLMs with chemistry-specific tools for synthesis planning, safety assessment, and property prediction, demonstrating end-to-end chemical research assistance. SciAgents (Ghafarollahi & Buehler, 2025) employs multi-agent collaboration with graph-based reasoning for materials science discovery. Agent Laboratory (Schmidgall et al., 2025) automates the full research workflow from literature review through experimentation to report writing, using LLM-driven planning to coordinate each stage. These systems demonstrate that LLM agents can interact with real or simulated scientific environments, but they typically do not maintain an explicit, revisable model of the underlying mechanisms that can be tested against experimental outcomes.

Hypothesis Generation and Evolution. SciMON (Wang et al., 2024b) generates novel scientific ideas by mining literature for inspiration and iteratively refining hypotheses against existing work. ResearchAgent (Baek et al., 2025) generates research ideas by building entity-relation graphs from academic papers and using them to identify knowledge gaps. FunSearch (Romera-Paredes et al., 2024) evolves programs through LLM-guided search to discover novel mathematical and scientific solutions, demonstrating that iterative LLM-based refinement can yield genuine discoveries. While these systems contribute to automated hypothesis generation, they generally represent hypotheses in natural language rather than as executable causal models, limiting their ability to derive interventional predictions or trace prediction failures to specific structural commitments.