Autonomous Scientific Discovery via Iterative Meta-Reflection

Bingchen Zhao[1] **Sara Beery**[2] **Oisin Mac Aodha**[1]
[1]University of Edinburgh [2]Massachusetts Institute of Technology

Abstract

Autonomous scientific discovery systems offer the potential to accelerate research by automating the process of hypothesis generation and validation. However, current systems operate within constrained search spaces or require predefined research questions, limiting their capacity for true open-ended inquiry. Furthermore, while they generate hypotheses iteratively, they largely lack the ability to explicitly synthesize their own accumulated findings to uncover complex, interconnected phenomena. We introduce DiscoPER, an autonomous large language model-powered framework that conducts open-ended research by dynamically generating and executing code to explore datasets without pre-specified research objectives. To ensure rigorous scientific validity, every proposed discovery must pass statistical testing. To overcome the limitations of isolated search, our framework introduces a second-order reasoning mechanism that periodically analyzes its own accumulated discoveries. By treating prior discoveries as empirical data, DiscoPER identifies structural patterns, confounds, and epistemic gaps, actively redirecting hypothesis exploration toward uncharted regions of the search space. The search space is further expanded by incorporating tool use, enabling the system to explore hypotheses beyond structured metadata by seamlessly processing and extracting useful information from multimodal sources like images. Evaluated on iNatDisco, a new multimodal ecological knowledge benchmark with pattern-level ground truth obtained from peer-reviewed literature, DiscoPER recovers 8 of 9 known patterns with a 72.7% hypothesis support rate, outperforming both classical causal discovery and LLM-guided baselines. Ablations show that DiscoPER scales with more data, and confirms the benefits of second-order "meta-reflection". Code and benchmarks are open-sourced to facilitate reproducible evaluation[1].

1 Introduction

Scientific discovery is a cumulative process. Researchers build on prior findings, notice gaps in what has been previously tested, and revise theories based on accrued evidence. Most existing LLM-based discovery systems (e.g., [36, 18, 15]) treat each hypothesis generation step independently, i.e., the system is provided with a dataset or research question, it generates a hypothesis, and the loop resets. There is no mechanism for the system to survey what it has already learned and to reason about what remains unknown. A human biologist, for instance, might notice that several pairwise correlations all involve the same variable and hypothesize a common cause, or review hundreds of field photographs to identify a seasonal migration pattern that no single observation reveals. This meta-level reflection across findings is central to science, yet is an aspect current discovery systems have largely ignored.

Moreover, existing systems are either restricted in what they can express or require external guidance to operate. Classical structure learning methods [39, 47, 8] search exhaustively over pairwise variable relationships but cannot express higher-order patterns such as interactions, mediation chains, or

[1]Code is available at https://github.com/DTennant/DiscoPER

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

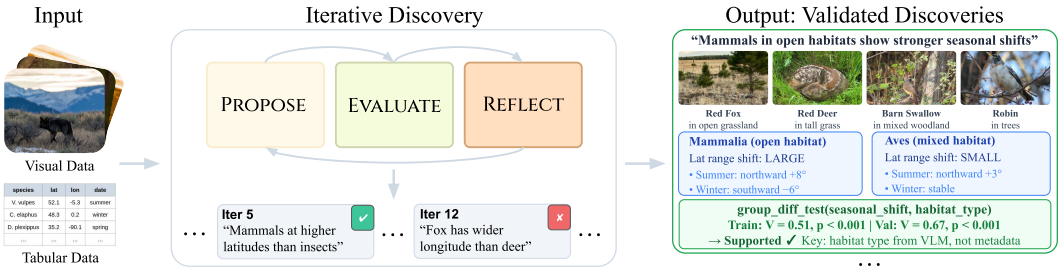


Figure 1: We introduce DiscoPER, an iterative approach for autonomous scientific discovery that takes multimodal data as input and generates a set of validated discoveries pertaining to the input data as output. At each iteration, the system proposes hypotheses, executes statistical tests on the underlying data, and accepts only discoveries that pass held-out validation. Periodically, the REFLECT module analyzes the accumulated accepted and rejected claims to identify gaps, confounds, and promising compound hypotheses, which then guide the next round of exploration.

confound structures. Full-pipeline AI scientist systems [30, 15, 14] can generate rich hypotheses from literature and domain knowledge, but require initial research objectives or seed hypotheses to guide their search. Code-based concurrent works [18, 36] write executable analyses, but their search is constrained by externally-supplied questions or task summaries. None of these systems perform fully open-ended discovery, i.e., (i) starting from raw unstructured data alone without a pre-specified question, (ii) expressing arbitrary testable hypotheses as executable code, and (iii) autonomously deciding what to investigate next based on what has already been discovered.

We present DiscoPER, an autonomous discovery system that satisfies all three requirements (see Fig. 1). At its core, DiscoPER contains LLM agents that receive data and propose hypotheses without being told what to look for a priori. Each hypothesis is realized as executable code that invokes statistical tools to ground every claim in a formal test on real data, rather than in LLM confidence estimates [25, 23]. Claims that pass statistical effect size and significance thresholds enter a persistent claim store. Periodically, a reflection step analyzes the current claims to identify under-explored variables and to detect recurring confounds across them. From this, it generates targeted guidance for the next model iteration, all without access to ground truth data.

In summary, we formalize open-ended discovery as a generalized PROPOSE–EVALUATE–REFLECT loop and show that existing systems are restricted instances of this framework, lacking either a flexible hypothesis space, unrestricted hypothesis generation, or reflective iterative search over the hypothesis space. We instantiate this framework as a multimodal, code-based system in which agents write and execute statistical analyses grounded in real data, validated on held-out splits. As existing benchmarks are often restricted by the research questions they ask, we construct a new benchmark from real scientific data where "ground-truth" patterns have been identified and the data itself allows for open-ended discovery. Our new multimodal benchmark dataset, called iNatDisco, is sourced from public citizen science data [1] that contains literature-grounded ecological claims. In our experiments, we demonstrate that DiscoPER obtains the best pattern rediscovery performance across multiple benchmarks, and that reflection outperforms compute-matched controls by enabling DiscoPER to systematically explore under-tested regions of the hypothesis space. Human evaluations further confirm that 91.7% of accepted claims are semantically valid and 85.0% of reflection outputs provide actionable scientific guidance, and we show that DiscoPER is able to rediscover over 60% of real-world literature-backed patterns across the iNatDisco benchmarks.

2 Related work

Causal discovery and structure learning. Our discovery framework bridges classical structure learning and recent LLM-augmented discovery. Traditional algorithms, including constraint-based [39], score-based [8], and continuous optimization methods [47], are fundamentally restricted to searching a predefined space that only contains the edges between variables in a dataset. More recently, LLM-based causal discovery methods [25, 23, 22] have emerged to improve search efficiency or inject domain knowledge. However, these approaches primarily replace a statistical testing oracle with LLM reasoning prompts or heuristic semantic judgments while remaining rigidly confined to the same edge space. Our approach fundamentally departs from these methods by using the LLM

strictly as a semantic generator and meta-reasoner to explore the vastly larger and more expressive space of executable hypotheses, while grounding actual verification in formal, data-driven, statistical testing to prevent hallucination.

Autonomous scientific discovery systems. Recent systems aspire to automate the full scientific process but differ fundamentally in how they source hypotheses and whether they perform data-driven validation. The AI Scientist [30] generates machine learning research ideas, writes code, runs experiments, and produces full papers, but operates on machine learning tasks rather than empirical data analysis and evaluates novelty via simulated review, rather than statistical validation. The AI Co-Scientist [15] uses multi-agent debate with tournament evolution to generate biomedical hypotheses that have achieved wet-lab validation, but its hypotheses originate from the literature and domain knowledge, rather than from data-driven statistical analysis. Robin [14] closes the loop with physical experiments, discovering a novel therapeutic, but is similarly guided by research objectives and literature context. SciAgents [13] traverses ontological knowledge graphs to uncover interdisciplinary connections in materials science, generating and critiquing hypotheses through multi-agent reasoning, but does not ground its claims in statistical tests on raw data. In laboratory settings, the Robot Scientist [26] demonstrated closed-loop hypothesis generation and physical experimentation in yeast genomics, and Coscientist [6] automates chemical synthesis planning and execution with LLM-driven tool use. However, both operate in domains with physical feedback loops rather than observational datasets. BioDiscoveryAgent [37] designs gene perturbation panels to achieve target cell phenotypes, but addresses a constrained optimization task rather than open-ended pattern discovery.

A parallel line of work focuses specifically on LLM-driven hypothesis generation and testing. [48] propose a framework for generating hypotheses with LLMs, while AutoDiscovery [2] drives open-ended discovery through Bayesian surprise, measuring belief shifts after evidence collection and uses Monte Carlo Tree Search (MCTS) to explore paths in a hypothesis tree. In contrast to tree-based path search guided by surprisal, DiscoPER maintains all accepted and rejected hypotheses in a shared claim store and uses a second-order REFLECT module to synthesize the full history, explicitly identifying coverage gaps, recurring confounds, and compound hypotheses across prior findings. [43] task LLMs with discovering transformation rules from input-output pairs via code execution, and [21] identify sub-hypotheses and design falsification experiments executed through code, though their approach remains largely linear and non-iterative. Concurrent to our work, ExperiGen [18] unifies hypothesis generation with experimental validation, incorporating statistical tests and a short-term memory module. While ExperiGen represents a step toward autonomous discovery, it requires descriptive seed hypotheses and task summaries to guide search, and its memory is conditioned implicitly on prior pairs rather than performing structured meta-analysis of what they collectively imply.

All of the above systems either require external research direction, lack a mechanism for structured reflection over accumulated findings, or both. Our approach differs in that it operates data-first, grounds every claim in held-out statistical validation, and uses reflective accumulation to reason about what the collection of findings implies for future exploration.

Benchmarks for scientific discovery. The shift from static LLM predictions to dynamic, code-driven agents has necessitated new environments and benchmarks for evaluating scientific discovery [31]. Recent datasets like BioDSA [44], HeurekaBench [36], BLADE [17], and DiscoveryBench [32] provide structured environments to evaluate agentic research capabilities. However, these benchmarks typically evaluate whether an agent can answer a predefined research question or navigate a highly constrained scenario. DiscoPER is designed for unconstrained, autonomous discovery, and as a result, standard task-oriented evaluation falls short. To address this, we introduce our own open-ended evaluation paradigm utilizing complex multimodal ecological data from the citizen science platform iNaturalist [1] to assess an agent's ability to autonomously formulate, test, and synthesize supported scientific claims from scratch.

3 Method

We posit that existing scientific discovery systems can be viewed as restricted instances of a single generalized framework (see Table 1 for a summary). In this section we introduce each component using our DiscoPER approach, illustrated in Fig. 2, as the primary example.

Setup. Let $\mathcal{X}$ be a dataset of N observations, optionally accompanied by additional data such as images. Let $\mathcal{P}$ denote any information beyond $\mathcal{X}$ available before the discovery loop begins: $\mathcal{P} = \emptyset$

Table 1: Existing autonomous scientific discovery systems can be viewed as instances of our generalized framework in Sec. 3. $\mathcal{H}$: hypothesis space ($\mathcal{H}_{\text{edge}}$ = pairwise variable edges, $\mathcal{H}_{\text{code}}^{\text{guided}}$ = executable code with external guidance, $\mathcal{H}_{\text{code}}^{\text{open}}$ = executable code without guidance). PROPOSE: how hypotheses are generated. EVALUATE: how hypotheses are tested. REFLECT: whether the system reasons over accumulated findings. $\mathcal{P}$: prior information required ($\emptyset$ = none, Partial = task description or seed hypotheses, Full = specific research questions or objectives).

Method	$\mathcal{H}$	PROPOSE	EVALUATE	REFLECT	$\mathcal{P}$
PC [39] / GES [8]	$\mathcal{H}_{\text{edge}}$	Enum. $(V_i, V_j, \mathbf{S})$	CI test; $p < \alpha$	$\mathcal{G} = \emptyset$	$\emptyset$
NOTEARS [47]	$\mathcal{H}_{\text{edge}}$	Grad. desc. on W	LS + DAG pen.	$\mathcal{G} = \emptyset$	$\emptyset$
GPT-4 BFS [23]	$\mathcal{H}_{\text{edge}}$	BFS; LLM judges	LLM confidence	$\mathcal{G} = \emptyset$	Full
ExperiGen [18]	$\mathcal{H}_{\text{code}}^{\text{guided}}$	LLM; refine $\leq T$	ReAct; Bonferroni	Short term memory	Partial
HeurekaBench [36]	$\mathcal{H}_{\text{code}}^{\text{guided}}$	$\emptyset$ (given)	Agent Eval	$\mathcal{G} = \emptyset$	Full
AI Scientist [30]	$\mathcal{H}_{\text{code}}^{\text{guided}}$	LLM ideation	ML metrics	$\mathcal{G} = \emptyset$	Full
DiscoPER (Ours)	$\mathcal{H}_{\text{code}}^{\text{open}}$	LLM($\mathcal{X}$, $\mathcal{C}_{t\text{-}1}$, $\mathcal{G}_{t\text{-}1}$)	Code	$\mathcal{G}_t = \text{REFLECT}(\mathcal{C}_t, \hat{\mathcal{C}}_t)$	$\emptyset$

(i.e., none), $\mathcal{P}$ = Partial (e.g., a dataset summary), or $\mathcal{P}$ = Full (e.g., specific questions to answer). A *hypothesis* h is a program that takes $\mathcal{X}$ as input and returns a judgment $b \in \{\text{supported}, \text{rejected}\}$ together with statistical evidence e. The *hypothesis space* $\mathcal{H}$ is the set of all such programs the system can express. A hypothesis that has been executed and accepted is called a *claim* / *discovery*. The *claim set* $\mathcal{C}_t \subseteq \mathcal{H}$ collects all claims accepted after t iterations, and we use $\hat{\mathcal{C}}_t$ to denote the set of rejected claims at t. The *claim set* on its own can accumulate discovered patterns, but it cannot conduct a meta-level analysis and use the accumulated knowledge to guide the next step of discovery. To address this, additional *guidance* is needed. The *guidance* $\mathcal{G}_t$ is a structured summary produced by analyzing $\mathcal{C}_t$ and $\hat{\mathcal{C}}_t$, operating at a higher level than individual claims. Specifically, it describes what has been explored, what is missing, and what to try next.

The discovery loop. Starting from $\mathcal{C}_0 = \emptyset$ and $\mathcal{G}_0 = \emptyset$, the system iterates:

$$h_t^{(1)}, \ldots, h_t^{(K)} \sim \text{PROPOSE}(\mathcal{X},\ \mathcal{C}_{t-1},\ \mathcal{G}_{t-1},\ \mathcal{P}) \tag{1}$$

$$\Delta\mathcal{C}_t, \Delta\hat{\mathcal{C}}_t = \text{EVALUATE}(\{h_t^{(k)}\}_{k=1}^{K},\ \mathcal{X},\ \mathcal{C}_{t-1},\ \hat{\mathcal{C}}_{t-1}) \tag{2}$$

$$\mathcal{C}_t = \mathcal{C}_{t-1} \cup \Delta\mathcal{C}_t,\ \hat{\mathcal{C}}_t = \hat{\mathcal{C}}_{t-1} \cup \Delta\hat{\mathcal{C}}_t \tag{3}$$

$$\mathcal{G}_t = \text{REFLECT}(\mathcal{C}_t, \hat{\mathcal{C}}_t) \tag{4}$$

At each iteration, PROPOSE generates K hypotheses conditioned on the current claims $\mathcal{C}_{t-1}$ and guidance $\mathcal{G}_{t-1}$. EVALUATE tests each hypothesis and adds those that pass to the claim set. REFLECT then analyzes the updated claims to produce updated guidance $\mathcal{G}_t$, which steers the next round of hypothesis proposal using PROPOSE. Systems without reflective reasoning set $\mathcal{G}_t = \emptyset$ for all t.

PROPOSE: multimodal hypothesis generation. An LLM agent receives a tabular summary of $\mathcal{X}$ (schema, column statistics, sample rows) and, when available, additional data such as sample images along with minimal domain framing (e.g., noting that images depict ecological scenes with vegetation, terrain, or animals; see Appendix C.5). Conditioned on this input, the current claim set $\mathcal{C}_{t-1}$ and the guidance $\mathcal{G}_{t-1}$, the agent outputs K structured hypotheses, each consisting of a natural-language statement, the variables involved, and accompanying Python code that implements a statistical test to verify each hypothesis. The guidance $\mathcal{G}_{t-1}$ steers DiscoPER toward under-explored regions of the hypothesis space without dictating specific hypotheses. By default, DiscoPER operates with $\mathcal{P} = \emptyset$, meaning no pre-specified research questions, target patterns, or seed hypotheses are provided, though the generator naturally draws on the base LLM's general pretraining knowledge when formulating candidate ideas.

EVALUATE: code execution and held-out validation. Each of the K hypothesis programs is executed on a training split of $\mathcal{X}$, invoking statistical tests such as correlation analysis, group comparison, predictive modeling, clustering with enrichment testing, or stratified subgroup re-analysis. The same code is then re-executed on a held-out validation split. A hypothesis is accepted into $\mathcal{C}_t$ only if the effect size exceeds a minimum threshold and the p-value falls below a significance level on both train and validation splits. The system is able to tune the code used for validating the hypothesis on a training set, but can only evaluate the code once on the validation set. This design is essential

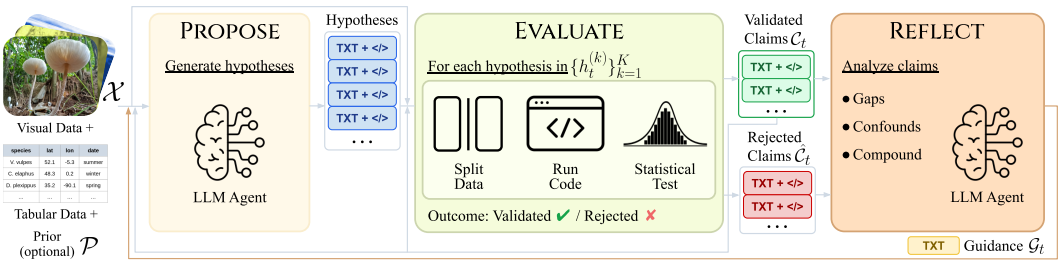


Figure 2: DiscoPER is an iterative scientific discovery system consisting of three core modules: PROPOSE generates hypotheses based on the data $\mathcal{X}$, and optional prior knowledge $\mathcal{P}$, and generates a set of candidate hypotheses $\{h_t^{(k)}\}_{k=1}^K$, where each is a natural language expression with accompanying code. EVALUATE generates code to test each hypothesis to either validate or reject them based on statistical evidence supported by the data. REFLECT analyzes the validated and rejected claims ($\mathcal{C}_t$ and $\hat{\mathcal{C}}_t$) to produce guidance $\mathcal{G}_t$ which steers the next round of hypothesis generation.

to avoid p-hacking where the system collect data or refines the analysis until nonsignificant results become significant [19]. As DiscoPER expresses hypotheses as executable Python programs over the statistical tool suite, its hypothesis space $\mathcal{H}_{\text{code}}^{\text{open}}$ encompasses multi-variable, conditional, stratified, and multimodal relationships rather than just fixed pairwise edges. This contrasts with classical causal discovery methods (e.g., [47, 8, 39]), which are restricted to $\mathcal{H}_{\text{edge}}$, i.e., programs that test a single directed edge $V_i \to V_j$ via a fixed conditional independence test, yielding $|\mathcal{H}_{\text{edge}}| = O(d^2)$. Code-based concurrent works such as ExperiGen [18] and HeurekaBench [36] also write programs, but they require an externally-supplied research question or target pattern ($\mathcal{P} \neq \emptyset$): HeurekaBench evaluates agents on pre-specified research questions paired with ground-truth answers and ExperiGen similarly operates on user-specified hypotheses. Without such external guidance these systems have no basis for selecting what to investigate as guidance is a structural requirement for these systems. DiscoPER can also accept optional guidance, but it is designed to operate with $\mathcal{P} = \emptyset$, freely exploring the full open space. We denote the constrained space $\mathcal{H}_{\text{code}}^{\text{guided}}$. These spaces form a hierarchy: $\mathcal{H}_{\text{edge}} \subset \mathcal{H}_{\text{code}}^{\text{guided}} \subseteq \mathcal{H}_{\text{code}}^{\text{open}}$.

REFLECT: meta-level guidance. Without explicit guidance, the hypothesis generator tends to revisit variants of previously explored hypotheses. Our REFLECT module addresses this by periodically analyzing the full claim set to redirect the search. After each round of evaluation, a separate LLM agent receives the accepted claims $\mathcal{C}_t$ and rejected claims $\hat{\mathcal{C}}_t$ and produces structured guidance $\mathcal{G}_t$. This guidance is a piece of text that prompts PROPOSE to explore different aspects of the dataset. In practice, the guidance takes diverse forms depending on the state of the claim store. It may identify *gaps*, where variables or relationships that are underexplored in $\mathcal{C}_t$. It may flag *confounds*, where a variable appears as a moderator across multiple accepted claims, suggesting that existing findings may be driven by an uncontrolled factor. Or it may propose *compound hypotheses*, where two or more accepted claims share overlapping variables in ways that imply a higher-order relationship that has not yet been tested. At the next iteration, PROPOSE conditions on $\mathcal{G}_t$, as is steered towards these opportunities without dictating specific hypotheses. REFLECT has no access to ground truth and reasons entirely from the system's own output.

This explicit reflection over accumulated evidence distinguishes DiscoPER from concurrent work: ExperiGen [18] conditions on prior hypotheses individually but never analyzes them collectively and HeurekaBench [36] has no mechanism to revise the search direction based on its findings. Our DiscoPER approach, displayed in Fig. 2, is the first to combine $\mathcal{H}_{\text{code}}^{\text{open}}$, $\mathcal{P} = \emptyset$, and use explicit reflection. This flexibility makes it well suited for open-ended discovery.

4 Experiments

Here we present quantitative and qualitative results comparing DiscoPER to alternative autonomous scientific discovery approaches. Results on additional datasets can be found in Appendix A.

4.1 Implementation details

Models. We evaluate DiscoPER with Claude Sonnet 4.5 as the default LLM, and additionally report results with Claude Opus 4.6, GPT 5.4, and DeepSeek V4 Pro. We run 100 iterations per

Table 2: Open-ended pattern rediscovery results on iNatDisco. *Recall (Exact)*: fraction of peer-reviewed ground-truth patterns rediscovered (with exact-only score-2 recall in parentheses). *Support rate*: fraction of all proposed hypotheses that pass held-out statistical validation, not applicable to causal methods (–) as they output a fixed graph rather than iteratively proposing hypotheses. Higher scores are better for both metrics.

Dataset	Method	Type	$\mathcal{H}$	Reflect	Prior Info	Recall (Exact)	Support Rate
iNatDisco-800	LLM+PC	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	0/9	-
	LLM+NOTEARS	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	0/9	-
	GPT-4 BFS	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	1/9	-
	Random Diversification	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✗	$\emptyset$	4/9	61.3%±4%
	HeurekaBench [36]	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✗	Partial	3/9	62.2%±6%
	HeurekaBench (Iterative)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✗	Partial	4/9	64.0%±5%
	ExperiGen [18]	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✓	Partial	3/9 (2/9)	56.6%±5%
	ExperiGen (Iterative)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✓	Partial	5/9	61.2%±4%
	DiscoPER w/o Reflect	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✗	$\emptyset$	7/9	70.0%±2%
	DiscoPER (Ours)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✓	$\emptyset$	**8/9 (7/9)**	**72.7%±3%**
iNatDisco-50K	LLM+PC	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	0/12	-
	LLM+NOTEARS	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	1/12	-
	GPT-4 BFS	Causal	$\mathcal{H}_{\text{edge}}$	✗	$\emptyset$	1/12	-
	Random Diversification	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✗	$\emptyset$	3/12	60.8%±4%
	HeurekaBench [36]	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✗	Partial	2/12	64.7%±4%
	HeurekaBench (Iterative)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✗	Partial	3/12	65.8%±4%
	ExperiGen [18]	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✓	Partial	3/12 (2/12)	67.8%±5%
	ExperiGen (Iterative)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{guided}}$	✓	Partial	4/12	68.9%±4%
	DiscoPER w/o Reflect	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✗	$\emptyset$	6/12	66.6%±3%
	DiscoPER (Ours)	Code / LLM	$\mathcal{H}_{\text{code}}^{\text{open}}$	✓	$\emptyset$	**8/12 (6/12)**	**74.2%±3%**

experiment with reflective accumulation every five iterations. At each iteration, DiscoPER outputs one hypothesis. A hypothesis is accepted only if the effect size $|\delta| \geq 0.2$ and $p \leq 0.05$ on both training and held-out validation splits, with an overfitting check requiring $|\delta_{\text{val}}| \geq 0.6 \cdot |\delta_{\text{train}}|$. We allow DiscoPER to call other LLMs or vision models in evaluating proposed hypotheses so that it can freely explore multimodal datasets. Specifically, when calling vision language models (VLMs) to identify visual attributes from images it uses the same base LLM for the vision model. Meaning that if the default LLM is Claude Sonnet 4.5, then it will use the same model for reading images. We run each experiment three times to calculate the mean and standard deviation. Please refer to Appendix C for additional implementation details.

Benchmark data. Existing evaluation protocols for scientific discovery are inadequate for our open-ended systems. Edge-recovery metrics (SHD, edge F1) cannot evaluate complex interactions such as mediation chains or confound structures. Question-answering benchmarks tell the agent what to look for, making them incompatible with $\mathcal{P} = \emptyset$. Therefore we construct two new multimodal benchmarks which target open-ended scientific discovery. The goal is to rediscover ecological patterns (i.e., relationships, trends, associations, etc.) using data sourced from research-grade observations from the citizen science platform iNaturalist [1]. Each observation on iNaturalist contains an image, latitude and longitude, positional accuracy, date of observation, and species name and taxonomic hierarchy. This makes it an ideal candidate for evaluating open-ended discovery from multimodal data. To curate ground-truth targets from the academic literature (e.g., "Monarch Butterfly observations show northward latitude shifts during spring-summer, consistent with their documented migration corridors" [7]), we started with 34 candidate patterns from 24 representative ecology papers, and excluded 6 where required variables were unavailable in iNaturalist, 5 where the finding could not be mapped to a clear target, and 6 where the pattern was not statistically present in our sampled data. This yielded 17 unique patterns across two datasets (with 3 shared between both, totaling 21 dataset–pattern entries): (i) *iNatDisco-800* which has 800 observations across eight species and nine ecological patterns obtained from peer-reviewed literature, and (ii) *iNatDisco-50K* which has 50,000 observations spanning 9,776 species with twelve ecological patterns. All 21 dataset–pattern entries remain statistically supported when verified with alternative statistical tests outside the system's default tool suite. Additional dataset construction details can be found in Appendix B.1.

4.2 Evaluation of open-ended discovery on iNatDisco

Here we demonstrate that DiscoPER, by combining code-driven hypothesis testing with meta-reflection, achieves higher performance on real world scientific pattern rediscovery. We evaluate using

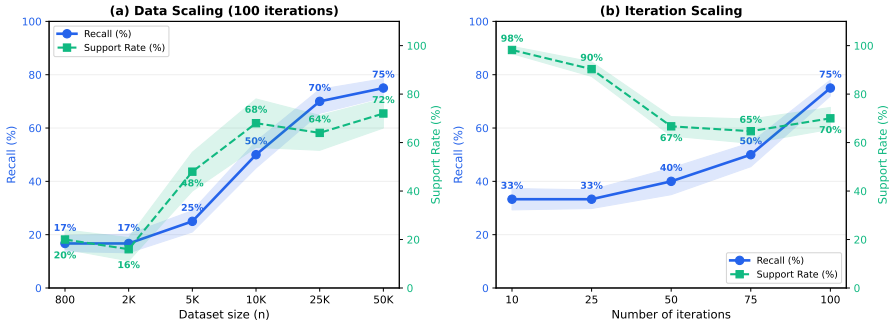


Figure 3: Scaling behavior on iNatDisco-50K. **(a)** Providing more data improves recall and yields more supported insights. **(b)** More model iterations increases recall but the support rate decreases as the model moves on from easy hypotheses and starts to propose more speculative ones.

two complementary metrics: (i) the *recall* of patterns from the peer-reviewed literature annotations, which measures validated rediscovery of known ecological relationships, and (ii) the *support rate* of the proposed hypotheses. The support rate, defined as the fraction of proposed hypotheses that pass held-out statistical validation, measures how reliably DiscoPER proposes hypotheses that are empirically present in the dataset via code. However, statistical significance alone does not measure scientific value and could be gamed by proposing trivial hypotheses that easily pass validation. Pairing the support rate with literature-pattern recall, exact-match scoring, untouched-test validation, and human evaluation therefore provides a more rigorous assessment.

The results are shown in Table 2. Classical causal discovery methods (LLM+PC, LLM+NOTEARS, GPT-4 BFS), even when equipped with an LLM to extract variables from raw data, recover at most 1/9 patterns on iNatDisco-800. Their edge-level hypothesis space cannot express the interaction effects, mediation chains, and multi-variable ecological patterns that dominate our ground truth. To test whether performance is explained simply by random coverage over the variable-and-test grid, we evaluate a *Random Diversification* baseline that randomly selects compatible variables, groups, temporal conditions, and test types from our tool suite under the same 100-hypothesis budget. It achieves 4/9 recall (61.3% support) on iNatDisco-800 and 3/12 recall (60.8% support) on iNatDisco-50K, showing that random grid coverage explains part of the performance but not the full gain. Guided LLM methods (HeurekaBench, ExperiGen) recover 3/9 patterns on iNatDisco-800 and 2–3/12 on iNatDisco-50K in their standard settings. When augmented with an iterative summary of prior results across all 100 iterations (*Iterative*), HeurekaBench improves to 4/9 and 3/12, and ExperiGen improves to 5/9 and 4/12, as their search remains constrained by the seeded task framing. DiscoPER discovers 8/9 patterns on iNatDisco-800 with a support rate of 72.7%, meaning nearly three-quarters of all hypotheses the system proposes are validated on held-out data. On iNatDisco-50K, which contains 12 patterns across 9,776 species, DiscoPER recovers 8/12 patterns with an even higher support rate of 74.2%, suggesting that larger datasets enable the system to ground its hypotheses more reliably even as the search space grows. Compared to the ablation without REFLECT, we observe consistent gains in both recall (7/9→8/9 on iNatDisco-800, 6/12→8/12 on iNatDisco-50K) and support rate, confirming that REFLECT broadens what the system investigates while improving proposal quality.

Exact-match recall, judge robustness, and untouched-test validation. Our LLM judge (Appendix C.1) scores matches as exact (score 2) or partial (score 1). Across the four main LLM discovery conditions, 29 matches receive score 2 and 11 receive score 1. For DiscoPER, the breakdown is 7 exact and 1 partial on iNatDisco-800, and 6 exact and 2 partial on iNatDisco-50K. Under strict exact-only recall, DiscoPER still achieves the best performance at 7/9 and 6/12 (compared to 2/9 and 2/12 for ExperiGen). Rescoring all outputs with GPT-5.4 as an independent judge yields 94.1% agreement with the Claude Sonnet 4.5 judge and preserves the same method ranking. Furthermore, to test whether adaptive search across 100 iterations overfits the validation split, we evaluated all final discoveries once on a strictly untouched test split that was never accessed by PROPOSE, EVALUATE, REFLECT, or any baseline during search. Of the 16 patterns recovered by DiscoPER, 15/16 remain supported under the original thresholds and 14/16 remain significant after Holm step-down correction across the final claims, yielding Holm-corrected untouched-test recall of 7/9 (iNatDisco-800) and 7/12 (iNatDisco-50K), compared to 6/9 and 5/12 without REFLECT.

In Fig. 3, we further investigate the scaling behavior of DiscoPER w.r.t dataset size and the number of iterations. We took iNatDisco-50K data and subsampled it to create data subsets of different scales

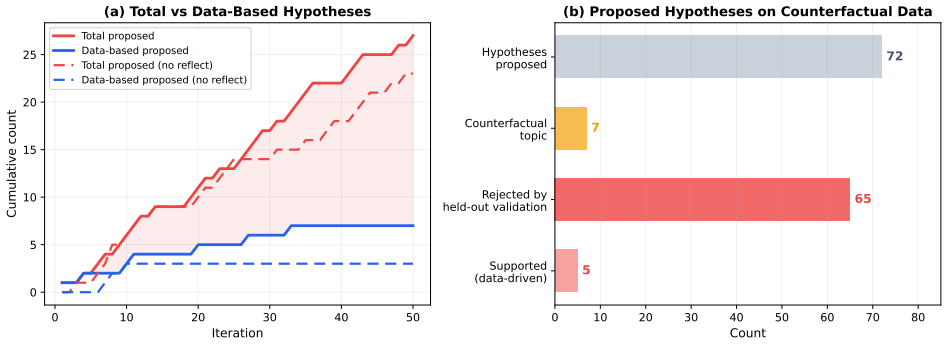


Figure 4: Experiments on our iNatDisco-800-CF counterfactual dataset. **(a)** Cumulative number of proposed hypotheses over 50 iterations, separated into all hypotheses and data-based hypotheses, with and without REFLECT. REFLECT increases the number of hypotheses grounded in the observed input data rather than only in the model's priors. **(b)** Distribution of proposed hypotheses. Most proposed hypotheses are rejected by held-out validation, the supported discoveries are data-driven and follow the counterfactual patterns present in the modified data.

to probe the scaling behavior of DiscoPER w.r.t dataset size. We observe that more data makes subtle cross-kingdom patterns more discoverable, and DiscoPER is able to recover more supported patterns from the data (Fig. 3 (a)). Additionally, we investigated the behavior when scaling the number of iterations, where each iteration proposes and evaluates one hypothesis and REFLECT runs every 5 iterations to analyze the claim store and steer subsequent proposals. Increasing the number of iterations reduces the support rate because the model proposes increasingly speculative hypotheses after early iterations have exhausted the easy ones (Fig. 3 (b)). In contrast, recall consistently increases as we increase the number of iterations, suggesting that the REFLECT-guided exploration continues to surface novel patterns even as per-hypothesis success rates decline.

4.3 Counterfactual evaluation

The ground truth patterns in iNatDisco are drawn from the peer-reviewed ecology literature, and thus we cannot guarantee they were absent from the base LLMs' training data. A system that simply recalls memorized facts, rather than testing hypotheses against data, would score well on this benchmark despite lacking genuine discovery capabilities. To address this, we constructed iNatDisco-800-CF, a counterfactual variant in which five well-known ecological relationships are deliberately reversed in the tabular data while the image set remains unchanged. Concretely, we modify observation timestamps, coordinates, and species-level metadata in the tabular data so that the original patterns no longer hold and the opposite relationships emerge instead. For example, we fix bird latitude to a constant value regardless of month (removing seasonal migration), resample 70% of fungal observations into spring months (reversing the autumn fruiting pattern), and narrow mammal geographic ranges while widening insect ranges (inverting the body-size and range relationship). As the images are not altered, the VLM still sees the same species in the same habitats, but the tabular data now contradicts real-world ecology. If a system relied only on LLM priors, it would report the real-world patterns (e.g., "fungi peak in autumn"). Instead, if it relies on data, it would report the counterfactual patterns actually present in the modified data. See Appendix B.2 for a detailed description of the specific data modifications we applied.

In Fig. 4 (a) we classify the proposed hypotheses into whether or not they are based on the observed data by examining if the PROPOSE process involves tool calling that examines the dataset itself. The gap between the data-based hypotheses and the total number of hypotheses can be roughly seen as the number of hypotheses that the PROPOSE module proposed mainly based on the LLM's internal knowledge. We observe that as the number of iterations increases, the LLM keeps proposing hypotheses that originate from its internal knowledge and a smaller fraction of the proposed hypotheses are based on the data. Additionally, the REFLECT module results in more data-driven hypotheses than without, demonstrating its efficacy at helping to learn from past hypotheses and their tested validity. As shown in Fig. 4 (b), the LLM proposes many prior-based hypotheses (e.g., "fungi peak in autumn"), but these are rejected by held-out validation because the data shows the opposite, whereas 7 proposals directly test the modified relationships and all 5 counterfactual patterns are validated.

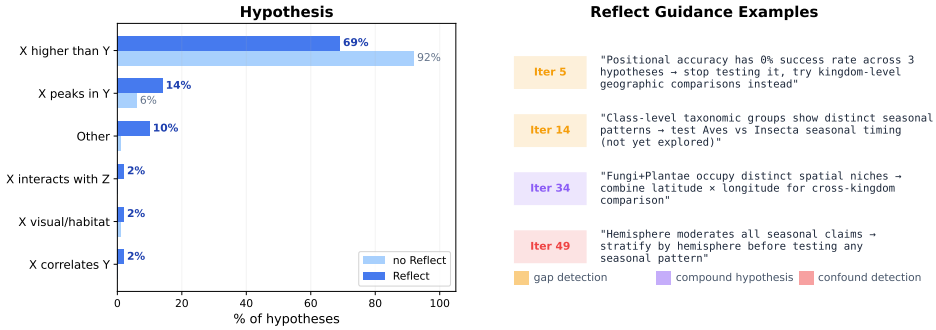


Figure 5: Ablation of reflection. **Left:** distribution of generated hypothesis with and without REFLECT. Without reflection, hypotheses are dominated by simple pairwise comparisons, while REFLECT produces a broader set of seasonal, interaction, visual, and correlation-based hypotheses. **Right:** examples of guidance produced by REFLECT, including gap detection, compound hypothesis generation, and confound detection. These guidance messages redirect later proposal steps toward under-explored variables and higher-order relationships.

Therefore the final supported discoveries are grounded in empirical data rather than memorized LLM priors, even when the data is counterfactual w.r.t. the LLM's internal knowledge.

4.4 Ablations

We analyze the internal behavior of DiscoPER to explore how our REFLECT module steers discovery and how vision contributes to hypothesis generation. We also evaluate compute-matched controls for REFLECT, report human expert verification of claims and reflections, and perform ablations on the backbone LLM and optional user-provided context (Appendix A).

Reflection guidance. Fig. 5 (right) illustrates concrete examples of REFLECT guidance produced during a 100-iteration run on iNatDisco-50K. The guidance falls into three categories, each addressing a distinct failure mode of unguided search: *Gap detection* identifies variables or relationships with zero coverage in the claim store. At iteration 5, REFLECT observes that positional accuracy has a 0% support rate across attempted hypotheses and recommends abandoning it in favor of kingdom-level geographic comparisons, redirecting search away from an unproductive direction. *Compound hypothesis generation* combines insights from multiple claims to propose higher-order hypotheses. At iteration 34, REFLECT observes that Fungi and Plantae have been tested separately against latitude and longitude, and proposes testing their joint latitude$\times$longitude spatial niche separation. *Confound detection* flags variables that moderate multiple existing claims. At iteration 49, REFLECT notices that hemisphere appears as a moderator in all seasonal claims and recommends stratifying by hemisphere before testing any seasonal pattern, preventing it from reporting confounded associations.

Without REFLECT, the distribution of hypotheses DiscoPER generates is heavily skewed toward simple pairwise comparisons (Fig. 5 (left)): 92% of base hypotheses follow the "X is higher/greater than Y" form, with no interaction or correlation hypotheses generated at all. REFLECT substantially diversifies the hypothesis space and reduces the dominance of direct pairwise comparison hypotheses from 92% to 69%, shifting DiscoPER toward more structured relational hypotheses. These include interaction hypotheses ($0\% \rightarrow 2\%$), correlation tests ($0\% \rightarrow 2\%$), and more specific seasonal or visual patterns, such as "X peaks in Y" ($6\% \rightarrow 14\%$). These shifts are modest in magnitude but important in effect, e.g., the compound hypothesis at iteration 34 ("test Fungi$\times$Plantae spatial niche separation") is one of only two interaction-type hypotheses in the entire run, yet it produces a statistically supported discovery ($p < 0.001$) that the baseline never formulates.

Vision capabilities. DiscoPER is designed to support discovery of patterns in scientific databases beyond text or tabular metadata. By allowing our method to query a vision language model (VLM) using code, we access visual processing capabilities that enable DiscoPER to find patterns that require understanding information uniquely stored in the *images* in our database. Fig. 6 illustrates two example results. In Fig. 6 (left), the VLM describes mammal observations as occurring in expansive terrain such as scrubland and mountains, while plant observations are described as "sessile and rooted in place". These visual cues motivate DiscoPER to formulate a species range-size hypothesis comparing mammals and plants. The resulting statistical test uses longitude information and finds that mammals have wider longitudinal ranges than plants in the data. While the longitudinal spread itself can be computed from metadata alone, the visual descriptions provide ecological context for why

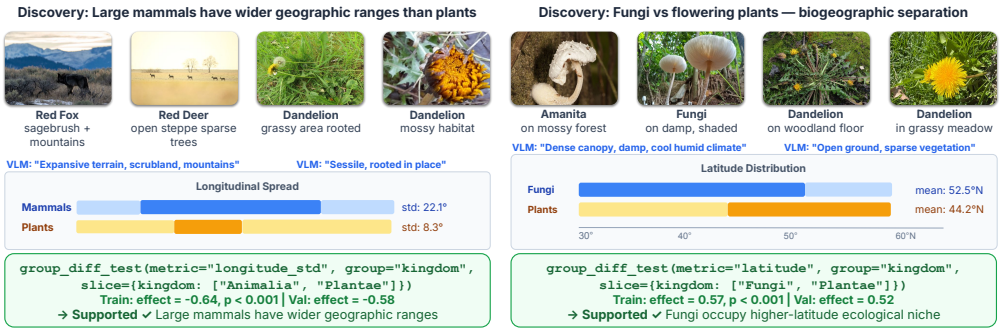


Figure 6: Examples of vision-grounded discoveries produced by DiscoPER. DiscoPER can use visual evidence extracted from images to formulate and validate hypotheses that are not directly available from tabular metadata. **Left:** VLM-derived descriptions of open terrain for mammals versus sessile, rooted growth for plants motivate a range-size hypothesis, which is then statistically validated on longitudinal metadata showing that mammals occupy broader longitudinal ranges than plants. **Right:** visual cues about canopy density and ground cover lead to a supported biogeographic separation between fungi and flowering plants, with fungi concentrated at higher latitudes. In both cases, candidate insights are converted into executable statistical tests and accepted only after validation on held-out data.

this comparison is meaningful, i.e., the system connects the geographic pattern to image-level cues about mobility, habitat openness, and rootedness. Additionally in Fig. 6 (right), the VLM observes fungi on "mossy forest floors with dense canopy" versus dandelions in "open ground with sparse vegetation," motivating DiscoPER to formulate a biogeographic niche hypothesis comparing fungi and flowering plants. The resulting statistical test uses latitude metadata and finds that fungi occupy a higher-latitude niche than flowering plants. Neither habitat type nor vegetation density appears in the metadata as both are extracted entirely from the images.

Limitations. DiscoPER does not control the input data generation process and thus cannot make discoveries if they are not supported by the data available. For example, there are known spatial, temporal, and observer biases in citizen science data which mean that statistical patterns in the raw observations may not always reflect underlying ecological phenomena [24]. Furthermore, passing statistical significance and effect-size thresholds confirms that a pattern is empirically supported in the data, but does not by itself establish causal validity or scientific novelty. As a result, human expert verification of any proposed discoveries is essential.

Our iNatDisco datasets evaluate open-ended inquiry through the rediscovery of claims validated in the academic literature rather than de novo scientific breakthroughs. Moreover, the set of annotated claims is not complete, i.e., there will be other valid patterns that a model could propose that are not annotated in the data, and our automated translation from published papers to testable benchmark claims when constructing the evaluation dataset may be imperfect. Importantly, the claims output by DiscoPER are selected based on passing held-out and untouched-test statistical significance tests, which ensures that they are grounded in empirical evidence in the data. In Appendix A we also report results on other datasets where the complete set of valid claims is known.

5 Conclusion

Autonomous scientific discovery requires more than proposing plausible hypotheses; it requires open-ended exploration, empirical validation, and the ability to reason over accumulated evidence. We introduced DiscoPER, a code-driven discovery framework that formulates testable open-ended hypotheses from multimodal data, validates them on held-out splits, and periodically reflects over its own claims to identify gaps, confounds, and compounds derived insights. Across ecological and causal discovery benchmarks, this combination improves both the recall and the diversity of supported discoveries over systems constrained by fixed edge spaces, predefined questions, or short-term memory. More broadly, our results suggest that progress in AI-assisted science depends on building agents that do not only generate isolated insights, but can also organize and reason over an evolving body of evidence. DiscoPER is still limited by the scope and bias of the data it observes, and its discoveries require human scrutiny. Nevertheless, reflective, evidence-grounded agents offer a promising path toward scientific tools that help researchers surface overlooked patterns and formulate new testable questions.

Acknowledgements. This work was in part supported by a Royal Society Research Grant, a Schmidt Sciences AI2050 Early Career Fellowship, an NSF CAREER Grant (Award No. 2441060), and the NSF and NSERC AI Biodiversity Change Global Center (NSF Award No. 2330423 and NSERC Award No. 585136).

References

[1] iNaturalist. `https://www.inaturalist.org`. Accessed on 2026-05-05.

[2] Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, et al. Autodiscovery: Open-ended scientific discovery via bayesian surprise. In *NeurIPS*, 2025.

[3] Roberto Ambrosini, Diego Rubolini, Anders Pape Møller, Luciano Bani, Jacquie Clark, Zsolt Karcza, Didier Vangeluwe, Chris du Feu, Fernando Spina, and Nicola Saino. Climate change and the long-term northward shift in the african wintering range of the barn swallow hirundo rustica. *Climate Research*, 2011.

[4] Jonathan J. Bennie, James P. Duffy, Richard Inger, and Kevin J. Gaston. Biogeography of time partitioning in mammals. *PNAS*, 2014.

[5] Lynne Boddy, Ulf Büntgen, Simon Egli, Alan C Gange, Einar Heegaard, Paul M Kirk, Aqilah Mohammad, and Håvard Kauserud. Climate variation effects on fungal fruiting. *Fungal Ecology*, 2014.

[6] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 2023.

[7] Lincoln P Brower. Monarch butterfly orientation: missing pieces of a magnificent puzzle. *Journal of Experimental Biology*, 1996.

[8] David Maxwell Chickering. Optimal structure identification with greedy search. *JMLR*, 2002.

[9] Frank-M. Chmielewski and Thomas Rötzer. Response of tree phenology to climate change across europe. *Agricultural and Forest Meteorology*, 2001.

[10] Natalie Cooper and Andy Purvis. Body size evolution in mammals: complexity in tempo and mode. *The American Naturalist*, 2010.

[11] Kevin J. Gaston. *The Structure and Dynamics of Geographic Ranges*. Oxford University Press, 2003.

[12] Valerius Geist. *Deer of the World: Their Evolution, Behaviour, and Ecology*. Stackpole Books, 1998.

[13] Alireza Ghafarollahi and Markus J Buehler. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *Advanced Materials*, 2025.

[14] Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv:2505.13400*, 2025.

[15] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, et al. Towards an AI co-scientist. *arXiv:2502.18864*, 2025.

[16] Elmer Gray, Eugene M. McGehee, and Don F. Carlisle. Seasonal variation in flowering of common dandelion. *Weed Science*, 1973.

[17] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. In *EMNLP (Findings)*, 2024.

[18] Jishu Sen Gupta, Harini SI, Somesh Kumar Singh, Syed Mohamad Tawseeq, Yaman Kumar Singla, David Doermann, Rajiv Ratn Shah, and Balaji Krishnamurthy. Accelerating social science research via agentic hypothesization and experimentation. *arXiv:2602.07983*, 2026.

[19] Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS Biology*, 2015.

[20] Helmut Hillebrand. On the generality of the latitudinal diversity gradient. *The American Naturalist*, 2004.

[21] Kexin Huang, Ying Jin, Ryan Li, Michael Y Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *ICML*, 2025.

[22] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *ICLR*, 2024.

[23] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv:2402.01207*, 2024.

[24] Alison Johnston, Eleni Matechou, and Emily B Dennis. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 2023.

[25] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *TMLR*, 2024.

[26] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 2004.

[27] Christian Körner. The use of 'altitude' in ecological research. *Trends in Ecology and Evolution*, 2007.

[28] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1988.

[29] Mark V. Lomolino, Brett R. Riddle, Robert J. Whittaker, and James H. Brown. *Biogeography*. Sinauer Associates, 4th edition, 2010.

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

[31] Erpai Luo, Jinmeng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Lei Wei, and Xuegong Zhang. Benchmarking ai scientists in omics data-driven biological research. *arXiv:2505.08341*, 2025.

[32] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *ICLR*, 2025.

[33] Annette Menzel, Tim H. Sparks, Nicole Estrella, Elisabeth Koch, Anto Aasa, Rein Ahas, Kerstin Alm-Kübler, Peter Bissolli, Ol'ga Braslavska, Agrita Briede, Frank M. Chmielewski, Zalika Crepinsek, Yannick Curnel, Aslog Dahl, Claudio Defila, Alison Donnelly, Yolanda Filella, Katarzyna Jatczak, Finn Mage, Antonio Mestre, Oyvind Nordli, Josep Penuelas, Pentti Pirinen, Viera Remisova, Helfried Scheifinger, Martin Striz, Andreja Susnik, Arnold J. H. van Vliet, Frans-Emil Wielgolaski, Susanne Zach, and Ana Zust. European phenological response to climate change matches the warming pattern. *Global Change Biology*, 2006.

[34] Montague H. C. Neate-Clegg and Morgan W. Tingley. Adult male birds advance spring migratory phenology faster than females and juveniles across north america. *Global Change Biology*, 2023.

[35] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *NeurIPS*, 2020.

[36] Siba Smarak Panigrahi, Jovana Videnović, and Maria Brbić. Heurekabench: A benchmarking framework for ai co-scientist. In *ICLR*, 2026.

[37] Yusuf Roohani et al. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *ICLR*, 2025.

[38] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 2005.

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

[40] Leho Tedersoo, Mohammad Bahram, Sergei Põlme, et al. Global diversity and geography of soil fungi. *Science*, 2014.

[41] N. Vanderhoff, P. Pyle, M. A. Patten, R. Sallabanks, and F. C. James. American robin (Turdus migratorius), version 1.0. In *Birds of the World*. Cornell Lab of Ornithology, 2020.

[42] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. In *NeurIPS - Datasets and Benchmarks*, 2024.

[43] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *ICLR*, 2024.

[44] Zifeng Wang, Benjamin Danek, and Jimeng Sun. Biodsa-1k: Benchmarking data science agents for biomedical research. *arXiv:2505.16100*, 2025.

[45] Kentwood D. Wells. *The Ecology and Behavior of Amphibians*. University of Chicago Press, 2007.

[46] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *ICML*, 2019.

[47] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *NeurIPS*, 2018.

[48] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Workshop on NLP for Science (NLP4Science)*, 2024.

Appendix

A Additional results

A.1 Additional ablations

Compute-matched controls for REFLECT. Because REFLECT introduces periodic LLM calls and summarizes prior outcomes, we evaluate three compute-matched controls on iNatDisco-800 and iNatDisco-50K using the same 100 evaluated hypotheses, number of model calls, context length, and output-token budget: (i) *Random guidance*, which injects random exploration prompts that may be unrelated to the claim history (achieving 6/9 recall and 68.1% support on iNatDisco-800, and 5/12 recall and 65.2% support on iNatDisco-50K); (ii) *Memory-only summary*, which summarizes accepted and rejected hypotheses without identifying gaps, confounds, or compound directions (6/9 recall and 69.0% support on iNatDisco-800, and 6/12 recall and 67.4% support on iNatDisco-50K); and (iii) *Additional-proposal control*, which replaces each REFLECT call with an extra hypothesis-generation call (7/9 recall and 70.4% support on iNatDisco-800, and 6/12 recall and 68.1% support on iNatDisco-50K). While extra calls, context, and raw history access each provide some benefit over no reflection, none matches REFLECT (8/9 and 72.7%, and 8/12 and 74.2%), showing that the remaining improvement stems from explicitly diagnosing coverage gaps, repeated failures, confounds, and compound links across prior findings.

Human verification of claims and reflections. To evaluate the scientific validity of DiscoPER's discoveries and reflections beyond automated metrics, we conducted blinded human evaluations with two annotators across three settings: (i) *Partial-match verification:* Two annotators blinded to the generating method and judge score reviewed all 11 score-1 (partial match) claims across the main LLM methods and confirmed 10/11 as valid partial matches (90.9% inter-annotator agreement), complementing our 95.0% human agreement on score-2 exact matches. (ii) *Accepted-claim validity:* Across a sample of 60 accepted hypotheses, human annotators judged 91.7% to be semantically valid and 86.7% to be scientifically interpretable, with 90.0% agreement. (iii) REFLECT *guidance quality:* Across 40 REFLECT outputs sampled across both iNatDisco benchmarks and different stages of search (shown with the prior accepted/rejected claim history but blinded to downstream hypotheses and outcomes), annotators judged 92.5% of reflections to be faithful to the earlier results and 85.0% to be useful and actionable for guiding the next proposal (87.5% agreement), with failure cases mainly consisting of overly general suggestions or unverified confound claims.

Base LLM comparison. Table A1 (left) shows that DiscoPER is compatible with different backbone LLMs, but the choice of model substantially impacts discovery performance. Claude Sonnet 4.5 achieves the highest recall, recovering 8/9 patterns with a 72.7% support rate. Claude Opus 4.6 obtains a slightly higher support rate (76.5%) but recovers only 4/9 patterns, suggesting a more conservative search behavior that validates a larger fraction of its proposed hypotheses, but explores fewer of the benchmark patterns. GPT-5.4 and DeepSeek V4 Pro (without vision) recover 3/9 and 2/9 patterns, respectively. The nonzero recall of the text-only DeepSeek variant indicates that the tabular analysis pipeline alone contains substantial signal, while the stronger vision-enabled models benefit from the additional image-derived evidence.

Controllability. Table A1 (right) shows that DiscoPER can be steered by providing optional user context without modifying the underlying discovery loop. In the prior-knowledge setting, the user provides only a small set of natural-language facts that should be treated as already known, such as "fungi peak in autumn" and "monarchs migrate," without revealing the benchmark targets or expected discoveries. In the guided setting, the user provides a single broad research interest, such as "investigate observer bias in GPS accuracy," rather than a concrete hypothesis or test to execute. Providing prior knowledge related to already-known facts increases topic adherence from 43% to 54% while preserving the same recall (7/9), indicating that the agent can incorporate external context without losing coverage of the benchmark patterns. Providing a specific research focus produces a stronger steering effect, increasing topic adherence to 68%, but reduces recall from 7/9 to 6/9 and lowers the support rate. This reflects the expected trade-off of controllable open-ended search, i.e., user context can redirect the system toward a desired region of the hypothesis space, but stronger steering narrows exploration and may reduce the number of broadly supported discoveries.

Compute-matched controls for REFLECT. Table A2 compares DiscoPER against three compute-matched controls that use the exact same 100 evaluated hypotheses, number of LLM calls, context length, and output-token budget: (i) *Random guidance*, which injects random exploration prompts unrelated to the claim history; (ii) *Memory-only summary*, which summarizes prior accepted and rejected hypotheses without identifying coverage gaps, confounds, or compound directions; and (iii) *Additional-proposal control*, which replaces each REFLECT call with an extra hypothesis-generation call. Across both iNatDisco-800 and iNatDisco-50K, DiscoPER with structured REFLECT outperforms all three controls on both recall and support rate.

Table A1: Model and controllability ablations on iNatDisco-800. **Left:** Recall across four LLMs. † denotes no vision used (i.e., text only). **Right:** Impact of providing prior knowledge or a research focus to the system. Topic adherence measures the fraction of hypotheses related to the user-specified interest.

Model	Supp. Rate	Recall
Sonnet 4.5	72.7%	8/9
Opus 4.6	76.5%	4/9
DeepSeek V4 Pro†	65.2%	2/9
GPT-5.4	70.1%	3/9

	Default	Prior Know.	Guided
User provides	*nothing*	*known facts*	*focus area*
Supp. Rate	72.7%	45.5%	36.7%
Recall	8/9	7/9	6/9
Topic adherence	N/A	54%	68%

Table A2: Compute-matched controls for REFLECT on iNatDisco-800 and iNatDisco-50K. All methods use the same 100-hypothesis evaluation budget, number of LLM calls, context length, and token budget.

	iNatDisco-800		iNatDisco-50K	
Method	**Recall**	**Supp. Rate**	**Recall**	**Supp. Rate**
DiscoPER w/o Reflect	7/9	70.0%	6/12	66.6%
Random guidance (compute-matched)	6/9	68.1%	5/12	65.2%
Memory-only summary (compute-matched)	6/9	69.0%	6/12	67.4%
Additional-proposal (compute-matched)	7/9	70.4%	6/12	68.1%
DiscoPER (Ours)	**8/9**	**72.7%**	**8/12**	**74.2%**

A.2 Classical causal discovery benchmarks

Here we evaluate DiscoPER on two standard causal discovery benchmarks with validated causal DAGs: SACHS [38] (11 proteins, 17 edges) and ASIA [28] (8 variables, 8 edges). For each benchmark, we compare against classical structure learning methods (PC, GES, NOTEARS, DAG-GNN, GOLEM) on edge recovery, and report our method with and without REFLECT. Table A3 reports edge-recovery F1 against the published ground truth DAGs. DiscoPER achieves the highest F1 on SACHS (0.83) and is competitive on ASIA (0.86), outperforming all classical structure learning methods by a substantial margin. On SACHS, the gap is particularly large: classical methods plateau at 0.33–0.48 F1, constrained by their edge-level hypothesis space, while DiscoPER can express and test richer relational patterns that map onto the same edges. Removing REFLECT consistently degrades performance (0.83→0.72 on SACHS, 0.86→0.80 on ASIA), confirming that reflective accumulation helps even on classical benchmarks by identifying under-tested variable pairs and reducing redundant proposals.

A.3 Synthetic visual benchmark

The iNatDisCo benchmarks evaluate discovery on real-world data, but they cannot isolate the contribution of vision from metadata as every visually-grounded hypothesis could potentially be proposed by an LLM that has memorized ecological knowledge. To disentangle these factors, we construct a synthetic benchmark where (i) the visual features are novel (colored shapes on backgrounds, not real species), ensuring the LLM has no prior knowledge about the patterns, and (ii) some ground truth patterns are only discoverable by analyzing the images (e.g., "red shapes are more common in autumn"), providing a controlled test of visual feature extraction.

We construct a synthetic benchmark to test DiscoPER's ability to extract visual features from images for hypothesis testing. The dataset contains 5,000 programmatically generated images of colored

Table A3: Edge recovery F1 on classical causal discovery benchmarks. The results for classical methods are obtained from the respective publications. Higher scores are better.

Method	SACHS	ASIA
DiscoPER	0.83±0.03	0.86±0.05
DiscoPER (no Reflect)	0.72±0.04	0.80±0.03
PC [39]	0.48	0.77
GES [8]	0.42	0.73
NOTEARS [47]	0.36	0.68
DAG-GNN [46]	0.33	0.65
GOLEM [35]	0.38	0.69
GPT-4 BFS [23]	0.74	0.93

shapes on backgrounds. Fig. A1 illustrates some representative examples. Each image has visual variables (i.e., color, shape, size, texture, background, count) and tabular metadata (i.e., category, region, season, temperature, elevation). Eight ground truth patterns span metadata-only, vision-only, and cross-modal relationships are constructed (see Table A4).

Results for DiscoPER on the synthetic dataset are presented in Table A5. At 100 iterations, the system recovers 3 of 8 ground truth patterns with a support rate of 54.2%. Both metadata-only patterns (P1, P2) are reliably discovered, confirming that the code-driven pipeline handles standard tabular relationships well. One vision-only pattern is recovered, where the agent identifies the color-season association (P3) by leveraging the VLM during hypothesis generation. However, the remaining vision-only and cross-modal patterns are not recovered: the agent often proposes correct visual hypotheses (e.g., "large shapes appear more often on grass backgrounds") but the current vision tool pipeline lacks the statistical power to validate them on held-out data, as individual image classification followed by a chi-squared test introduces substantial noise. This result highlights that the bottleneck for multimodal discovery is not hypothesis generation but statistical validation of visual features, motivating future work on richer vision-to-tabular feature extraction.

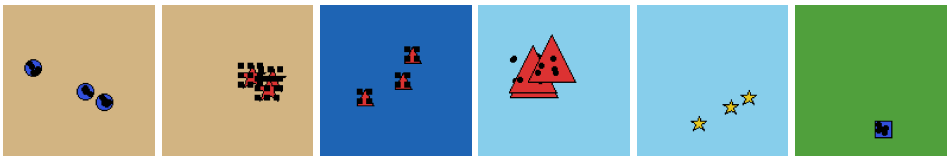


Figure A1: Sample images from the synthetic visual benchmark. Shapes vary in color, size, and texture; backgrounds vary between sky, grass, water, and sand.

Table A4: Ground truth patterns in the synthetic visual benchmark.

#	Pattern	Type	Description
P1	Temperature ∼ elevation	Metadata	$r = -0.7$
P2	Category A warmer	Metadata	Higher mean temperature
P3	Red in autumn	Vision	67% of autumn images are red (vs 25%)
P4	Size ∼ background	Vision	Large on grass (61%), small on water
P5	Shape-texture exclusion	Vision	Circles/squares never striped
P6	Blue ∼ elevation	Cross-modal	Blue at elevation 2402 vs 1484
P7	Region ∼ size	Cross-modal	North: 72% large; South: 18%
P8	Spotted + cat. C → cold	Cross-modal	Anomalously low temperature

B Additional dataset details

B.1 iNatDisco benchmark construction

We construct two ecological discovery benchmarks from iNaturalist research-grade observations, using the same set of images as the INQUIRE dataset [42]: (i) iNatDisco-800, containing 800 observations across 8 species with 9 ground-truth patterns, and (ii) iNatDisco-50K, containing 50,000

Table A5: Synthetic visual benchmark results.

# Iterations	Method	Type of Pattern	Supp. Rate	Recall
-	HeurekaBench	All	34.5%	0/8
-	ExperiGen	All	41.1%	1/8
50 iter	Ours	All	56.4%	2/8
100 iter	Ours	All	54.2%	3/8
100 iter	Ours	Metadata	–	2/8
100 iter	Ours	Vision	–	1/8
100 iter	Ours	Cross-Modal	–	0/8

observations across 9,776 species with 12 ground-truth patterns. A candidate pattern is a relationship reported in a peer-reviewed ecology paper that can be tested using fields available in iNaturalist, such as species, observation date, geographic coordinates, or images. Each candidate was converted into a structured benchmark target specifying the variables involved, expected direction, and any temporal or taxonomic conditions. We began with 34 candidate patterns drawn from 24 representative ecology papers. We excluded 6 candidates because the required variables were unavailable in iNaturalist, 5 because the paper's finding could not be mapped to a clear testable target, and 6 because the pattern was not statistically present in our sampled data ($p < 0.05$, effect size ≥ 0.2). This filtering left 17 unique patterns across the two datasets (Table A6): 9 in iNatDisco-800 and 12 in iNatDisco-50K, with 3 patterns shared across both datasets (rows 1, 6, and 9) and row 10 grouping both the general latitudinal diversity gradient and tropical insect concentration from [20], totaling 21 dataset–pattern entries.

Importantly, benchmark construction is strictly separated from system evaluation: DiscoPER receives neither the target patterns nor the code, statistical tests, or results used during the recoverability check, and all LLM baselines share the exact same tool suite and validation protocol. As an additional sensitivity check, we repeated benchmark validation using alternative statistical tests that were not used in the original construction process; all 21/21 dataset–pattern entries remain statistically supported, and the method ranking in Table 2 is unchanged. Note that while some published findings focus on specific taxa or fine-grained conditions, citizen-science observations in iNaturalist do not always provide that level of specificity, so we standardize broader taxonomic or seasonal slices where needed to form testable targets. Examples of hypotheses proposed by DiscoPER that were rejected by held-out statistical validation on iNatDisco are shown in Fig. A2.

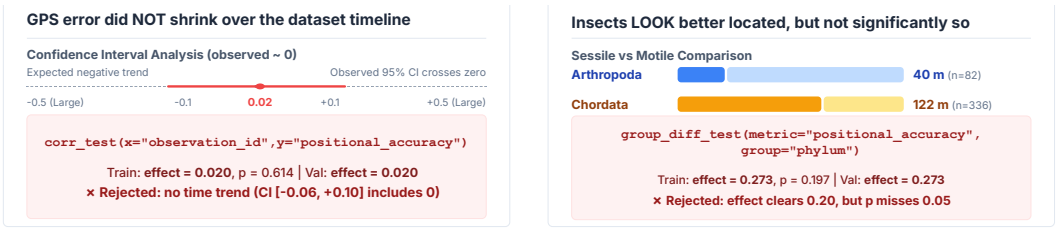


Figure A2: Example rejected claims on iNatDisco.

B.2 iNatDisco-800-CF: counterfactual dataset construction

Here we describe the iNatDisco-800-CF dataset that is used in Sec. 4.3 in the main paper. We construct iNatDisco-800-CF by modifying iNatDisco-800 to reverse five well-established ecological relationships (Table A7). For each relationship, we alter the underlying data so that the real-world pattern is no longer statistically present and the reversed pattern holds instead.

Each counterfactual was verified to be statistically present in the modified dataset. For example, after modification CF1, the Spearman correlation between bird latitude and month is $r = 0.02$, $p = 0.66$ (no significant seasonal shift). After CF2, 77% of fungal observations fall in spring months. After CF4, mammal latitude standard deviation is 2.5° versus 14.8° for insects.

Table A6: Ground-truth ecological patterns used in iNatDisco-800 and iNatDisco-50K. Three patterns are present in both datasets.

#	Dataset	Pattern	Ecological Basis	Reference
1	800, 50K	Fungi autumn fruiting	*A. muscaria* / fungal observations concentrated Sep–Nov	"Climate variation effects on fungal fruiting" (Boddy et al., 2014) [5]
2	800	Monarch migration	*D. plexippus* latitude increases Mar–Aug	"Monarch butterfly orientation: missing pieces of a magnificent puzzle" (Brower, 1996) [7]
3	800	Barn Swallow migration	*H. rustica* shifts northward in spring	"Climate change and the long-term northward shift in the African wintering range of the barn swallow" (Ambrosini et al., 2011) [3]
4	800	Red Fox wide range	*V. vulpes* spans 25–70°N	"Body size evolution in mammals: complexity in tempo and mode" (Cooper and Purvis, 2010) [10]
5	800	Robin partial migration	*T. migratorius* northward shift spring–summer	"American Robin (*Turdus migratorius*), version 1.0" (Vanderhoff et al., 2020) [41]
6	800, 50K	Plant seasonality / phenological offset	Plantae concentrated in spring–summer; plants peak Apr–Jun vs animals May–Aug	"European phenological response to climate change matches the warming pattern" (Menzel et al., 2006) [33]
7	800	Dandelion early flowering	*T. officinale* peaks Mar–May	"Seasonal variation in flowering of common dandelion" (Gray et al., 1973) [16]
8	800	Red Deer northern habitat	*C. elaphus* concentrated 45–60°N	"Deer of the World: Their Evolution, Behaviour and Ecology" (Geist, 1998) [12]
9	800, 50K	Hemisphere season inversion	Seasonal patterns invert between NH and SH	"Response of tree phenology to climate change across Europe" (Chmielewski & Rötzer, 2001) [9]
10	50K	Latitudinal diversity gradient / insect tropical concentration	Species richness increases toward the equator; insects more concentrated at $<30°$ latitude	"On the generality of the latitudinal diversity gradient" (Hillebrand, 2004) [20]
11	50K	Bird latitudinal migration	Birds shift northward spring–summer	"Adult male birds advance spring migratory phenology faster than females and juveniles across North America" (Neate-Clegg et al., 2023) [34]
12	50K	Amphibian spring emergence	Amphibians peak sharply Mar–May	"The Ecology and Behavior of Amphibians" (Wells, 2007) [45]
13	50K	Lepidoptera wide latitude	Butterflies span wider range than other insects	"The Structure and Dynamics of Geographic Ranges" (Gaston, 2003) [11]
14	50K	Fungi temperate concentration	Fungi concentrated 40–60°N	"Global diversity and geography of soil fungi" (Tedersoo et al., 2014) [40]
15	50K	Mammal temporal uniformity	Mammals more uniform across months than birds	"Biogeography of time partitioning in mammals" (Bennie et al., 2014) [4]
16	50K	Continental endemism	Certain families show continental endemism	"Biogeography," 4th ed. (Lomolino et al., 2010) [29]
17	50K	Elevation-latitude proxy	Alpine plants at higher latitudes in mid-lat bands	"The use of 'altitude' in ecological research" (Körner, 2007) [27]

The counterfactual evaluation tests a critical property of code-driven discovery: because every claim must pass held-out statistical validation, the system cannot "hallucinate" a pattern that is not in the data, even when the LLM's prior knowledge strongly suggests it should be there. As reported in Fig. 4, across 72 hypotheses proposed on iNatDisco-800-CF, the LLM proposes many prior-based hypotheses (e.g., "fungi peak in autumn"), but 65 proposals are rejected by held-out validation because the data shows the opposite, while 7 proposals directly test the modified relationships and 5 counterfactual patterns are validated.

C Additional implementation details

C.1 Evaluation protocol: LLM-as-a-judge for recall

Evaluating open-ended discovery requires matching free-form natural-language claims against ground truth patterns that may be worded differently, use different species exemplars, or operate at a different level of generality. String matching and keyword overlap fail in this setting, e.g., a system claim "Migratory insects shift northward in spring" should match the ground truth "Monarch butterfly latitude increases March–August" even though they share few similar words. We therefore use an LLM-as-a-judge to perform semantic matching.

Input filtering. Only *supported* claims are submitted to the judge. Rejected and inconclusive claims are excluded. This ensures that the recall metric measures not just whether the system *proposed* a pattern, but whether it proposed a pattern *and* produced code that validates it on held-out data.

Batched scoring. We present the judge with the full list of ground truth patterns and one supported claim at a time. For each claim, the judge outputs:

- `best_pattern_id`: the ground truth pattern ID that best matches the claim, or `none` if no match exists.
- `score`: an integer on a 0–2 scale:
  - 2 (exact match): the claim captures the same underlying ecological mechanism as the ground truth, even if worded differently or using different species exemplars (e.g., "Migratory insects shift northward in spring" matches "Monarch butterfly latitude increases March–August").
  - 1 (partial match): the claim is related but captures only part of the relationship, is too vague, or describes a consequence rather than the core pattern.
  - 0 (miss): the claim describes a completely different phenomenon.
- `reasoning`: a one-sentence explanation of the scoring decision.

The judge is instructed to focus on *semantic equivalence of the underlying ecological mechanism*, not surface-level wording. A generalized version of a ground truth pattern (e.g., using a higher taxonomic level or a different exemplar species) receives score 2 if the core mechanism is identical. Each claim may match at most one ground truth pattern. The judge is also instructed not to match claims to prior-knowledge or obvious-tier patterns, ensuring that recall is computed only over novel patterns.

Aggregation. For each ground truth pattern, we record the highest score received from any supported claim. A pattern is considered "discovered" if its best score is ≥ 1 (partial or exact match), and we also report exact-only recall (requiring score $= 2$) in Table 2. Across the four main LLM discovery conditions on iNatDisco, 29 matches receive score 2 (exact) and 11 receive score 1 (partial). For DiscoPER, the breakdown is 7 exact and 1 partial on iNatDisco-800 (7/9 exact-only recall), and 6 exact and 2 partial on iNatDisco-50K (6/12 exact-only recall), compared to 2/9 and 2/12 exact-only recall for ExperiGen. We report recall as fractions (e.g., 8/9) rather than percentages to make the denominator explicit.

Table A7: Counterfactual inversions in iNatDisco-800-CF. Each row shows the real-world relationship (which the LLM is expected to know), the counterfactual we inject into the data, and the modification applied.

#	Real World	Counterfactual	Data Modification
CF1	Birds migrate north in spring	Birds show no seasonal latitude shift	Set bird latitude to species mean $\pm\, \mathcal{N}(0, 2)$ regardless of month
CF2	Fungi fruit in autumn (Sep–Nov)	Fungi peak in spring (Mar–May)	Resample 70% of fungal observations to months 3–5
CF3	Species richness increases toward equator	Richness decreases toward equator	Remove 60% of observations at $\lvert\text{lat}\rvert < 30°$
CF4	Large mammals have wide ranges	Mammals narrow, insects wide	Set mammal lat $\sim \mathcal{N}(\mu, 1.5)$; insect lat $\sim \mathcal{N}(\mu, 15)$
CF5	Plants peak in spring/summer	Plants peak in winter (Nov–Feb)	Resample 65% of plant observations to months 11, 12, 1, 2

Judge model and cross-judge robustness. We use Claude Sonnet 4.5 as the primary judge model via structured JSON output (function calling). The same judge model is used across all experiments and baselines to ensure comparability. To verify that results are not sensitive to judge selection, we rescored all outputs using GPT-5.4 as an independent judge model and found 94.1% agreement with the Claude Sonnet 4.5 judge, with the exact same ranking across all methods.

Human verification of judge scores, accepted claims, and REFLECT guidance. We conducted three blinded human verification studies with two annotators to validate both the evaluation protocol and the scientific quality of DiscoPER's intermediate and final outputs:

- *Exact and partial match verification:* Manually inspecting all score-2 matches across iNatDisco-800 and iNatDisco-50K showed 95.0% human agreement with the judge. In addition, two annotators blinded to the generating method and original judge score reviewed all 11 score-1 (partial match) claims across the main LLM methods and confirmed 10/11 as valid partial matches, with 90.9% inter-annotator agreement.
- *Accepted-claim validity:* Across a sample of 60 accepted hypotheses generated by DiscoPER, two blinded annotators evaluated whether each claim was semantically valid (aligned with the executed statistical test and result) and scientifically interpretable. Annotators judged 91.7% of accepted claims to be semantically valid and 86.7% to be scientifically interpretable, with 90.0% inter-annotator agreement.
- REFLECT *guidance evaluation:* We sampled 40 REFLECT outputs across both iNatDisco benchmarks and different stages of the search. Two annotators were shown the accepted and rejected claims available when each reflection was produced, together with the resulting guidance, but were blinded to later hypotheses and whether the guidance led to a supported discovery. Annotators assessed whether each reflection was *faithful* (basing its statements about coverage gaps, repeated failures, or possible confounds on the supplied claim history) and *useful* (suggesting a specific, scientifically reasonable analysis testable with the available data and tools, rather than giving a generic recommendation or repeating an earlier hypothesis). Annotators judged 92.5% of reflections to be faithful to prior results and 85.0% to be useful for guiding the next proposal, with 87.5% agreement.

C.2 LLM-based baseline construction

We implement two LLM baselines as ablations of our framework by varying the prior information $\mathcal{P}$ provided to the system (see Table 1 in the main paper). A central claim of our work is that open-ended discovery (i.e., $\mathcal{P} = \emptyset$) combined with reflective accumulation yields better results than concurrent guided approaches. To test this rigorously, we implement ExperiGen-like and HeurekaBench-like baselines *within our own framework*, ensuring that the only differences are (i) the prior information $\mathcal{P}$ provided to the hypothesis generator and (ii) whether REFLECT is enabled. All other components (e.g., the backbone LLM (Claude Sonnet 4.5)), the statistical tool suite (Appendix C.4), the held-out validation protocol (effect size ≥ 0.2, $p \leq 0.05$ on both train and validation splits), the experiment planner, and the insight evaluator are identical across all conditions. This controlled design isolates the effect of guidance and reflection from confounds due to model choice, tool availability, or evaluation criteria.

HeurekaBench-like ($\mathcal{P}$ = Full). HeurekaBench [36] evaluates agents on pre-specified research questions with known ground-truth answers. We simulate this by replacing the open-ended hypothesis generator prompt with nine specific research questions drawn from our iNatDisco-800 ground truth. The agent is instructed to pick one unanswered question per iteration and design a rigorous statistical test for it. REFLECT is disabled, as HeurekaBench has no accumulation mechanism. The full system prompt is:

HeurekaBench-like Baseline Prompt

```
You are analyzing a biodiversity dataset.
TASK DESCRIPTION: This dataset contains research-grade wildlife observations
across species with GPS coordinates, dates, and taxonomic classification.
The goal is to find statistically significant relationships between species
traits, geographic distribution, and temporal patterns.
DATASET SUMMARY:
- 800 observations across 8 species
- Variables:  latitude, longitude, month, season, kingdom, class_name,
species_name, positional_accuracy
- Species include:  Vulpes vulpes, Cervus elaphus (mammals), Hirundo rustica,
Turdus migratorius (birds), Danaus plexippus (insect), Amanita muscaria
(fungi), Helianthus annuus, Taraxacum officinale (plants)
- Geographic range:  global, with concentration in Northern Hemisphere
Generate hypotheses about ecological patterns in this data.  Test each with
appropriate statistical methods.
```

This setup provides the agent with *maximum* guidance: it knows exactly what to look for. Despite this advantage, the agent recovers only 3/9 patterns in the standard setting (see Table 2), because (i) many questions are phrased at the species level while the ground-truth patterns involve cross-taxon interactions, and (ii) the fixed question list cannot adapt to emerging evidence.

ExperiGen-like ($\mathcal{P}$ = Partial). ExperiGen [18] provides agents with a task description, dataset summary, and seed hypotheses to accelerate discovery. We simulate this by replacing the hypothesis generator prompt with a structured task description that includes the dataset name and size (800 observations, eight species across three kingdoms), all variable names and types, the species list with taxonomic classification, the geographic scope, and a general instruction to "find statistically significant relationships between species traits, geographic distribution, and temporal patterns." REFLECT is disabled, as ExperiGen uses implicit conditioning on prior hypothesis-evidence pairs rather than explicit meta-analysis. The full system prompt is:

ExperiGen-like Baseline Prompt

```
You are analyzing a biodiversity dataset.
TASK DESCRIPTION: This dataset contains research-grade wildlife observations
across species with GPS coordinates, dates, and taxonomic classification.
The goal is to find statistically significant relationships between species
traits, geographic distribution, and temporal patterns.
DATASET SUMMARY:
- 800 observations across 8 species
- Variables:  latitude, longitude, month, season, kingdom, class_name,
species_name, positional_accuracy
- Species include:  Vulpes vulpes, Cervus elaphus (mammals), Hirundo rustica,
Turdus migratorius (birds), Danaus plexippus (insect), Amanita muscaria
(fungi), Helianthus annuus, Taraxacum officinale (plants)
- Geographic range:  global, with concentration in Northern Hemisphere
Generate hypotheses about ecological patterns in this data.  Test each with
appropriate statistical methods.
```

This setup provides *partial* guidance: the agent knows the variable space and the general direction of inquiry, but must decide which specific hypotheses to pursue. It recovers 3/9 patterns on iNatDisco-800 (see Table 2), similar to the HeurekaBench-like baseline, because the generic task framing tends to produce simple pairwise comparisons rather than the compound cross-taxon patterns that dominate the ground truth.

DiscoPER without reflection ($\mathcal{P} = \emptyset$, no Reflect). This baseline uses the same open-ended setting as our full system ($\mathcal{P} = \emptyset$) but with $\mathcal{G}_t = \emptyset$ for all t, i.e., the hypothesis generator never receives guidance from the REFLECT step. This isolates the contribution of reflective accumulation while keeping all other components identical, including the open-ended system prompt that prioritizes cross-kingdom and interaction hypotheses.

Iterative LLM baselines and random diversification. To test whether simple access to prior history allows guided baselines to match DiscoPER, we also evaluate iterative variants (*HeurekaBench (Iterative)* and *ExperiGen (Iterative)* in Table 2) that receive a summary of previously accepted and rejected hypotheses across all 100 iterations. Additionally, to separate structured reflection from exhaustive coverage over the variable-and-test space, we implement a *Random Diversification* baseline that randomly samples compatible variables, taxonomic groups, temporal slices, and statistical test primitives from Table A8 without replacement under the same 100-hypothesis budget and held-out acceptance criteria.

Controlling for fairness. We emphasize three design choices that ensure a fair comparison: *(i)* All baselines use the same backbone LLM, so differences in recall are not attributable to model capability. *(ii)* All baselines use the same held-out validation protocol, so differences in support rate reflect genuine differences in hypothesis quality, not evaluation stringency. *(iii)* The HeurekaBench-like baseline is given research questions that directly correspond to ground-truth patterns, an *advantage* over our open-ended system, which must discover these patterns from scratch. That our system achieves higher recall despite this disadvantage strengthens the case for open-ended exploration with reflective accumulation.

C.3 Additional details

Classical causal discovery baselines. For the SACHS and ASIA benchmarks, we compare against classical structure learning algorithms: PC [39], GES [8], NOTEARS [47], DAG-GNN [46], and GOLEM [35]. For PC and GES, we used the `causal-learn` Python package with default hyperparameters. For NOTEARS, DAG-GNN, and GOLEM, we report published results from their respective papers on the same SACHS and ASIA datasets.

Each classical method outputs a set of directed edges (e.g., PKA $\rightarrow$ Raf). To evaluate these against our pattern-level ground truth, we convert each edge to a natural-language claim of the form "`[Variable A]` has a direct causal effect on `[Variable B]`." These converted claims are then scored by the same LLM judge used for all other methods, ensuring a fair comparison on identical ground truth.

Variable extraction for causal baselines. On SACHS, the 11 protein variables (Raf, Mek, Erk, Akt, PKA, PKC, P38, JNK, Plcg, PIP2, PIP3) are used directly as nodes in the causal graph. On ASIA, the 8 binary variables (asia, tub, smoke, lung, bronc, either, xray, dysp) are used directly. No feature engineering or variable selection is applied, the methods receive the raw variables as defined in the original benchmark publications.

C.4 Statistical tools

The EVALUATE step executes hypothesis code using a fixed set of seven statistical primitives. The hypothesis generator selects which tool to invoke and specifies its parameters; no manual intervention is required. Table A8 describes each tool.

All tools return a standardized result dictionary containing `effect_size`, `p_value`, and `status`. A hypothesis is accepted into the claim store only if the effect size exceeds 0.2 and $p \leq 0.05$ on both the training and held-out validation splits.

C.5 Prompts

Here we provide the full prompts used in each component of DiscoPER. The same prompts are used for iNatDisco datasets as well as the causal discovery and synthetic datasets.

C.5.1 Hypothesis generation (PROPOSE)

The hypothesis generator receives the dataset summary, optional prior knowledge, recent discovery results, and a list of hypotheses to avoid. In our default open-ended setting ($\mathcal{P} = \emptyset$), the prior knowledge, recent discovery results, and hypotheses to avoid are initialized to empty at the start, though the prompt provides minimal domain framing (e.g., identifying the setting as ecological data analysis and, when images are present, noting general scene elements such as vegetation, water, animals, terrain, and sky). On subsequent iterations, the prompt includes the claim store and guidance from REFLECT.

Table A8: Statistical tools available to DiscoPER. The first five operate on tabular data and the last two use a vision language model to extract visual features from images before applying statistical tests.

Tool	Description
`corr_test`	Tests the correlation between two continuous variables using Spearman (rank-based, non-parametric) or Pearson (linear) correlation. Returns the correlation coefficient as effect size and the associated p-value.
`group_diff_test`	Tests whether a continuous metric differs between exactly two groups. Computes Cliff's delta (non-parametric) or Cohen's d (parametric) as effect size, with significance assessed via permutation test ($n = 1{,}000$ shuffles), Mann–Whitney U, or Student's t-test.
`predictive_test`	Trains a random forest classifier to predict a target variable from a set of features, evaluated via 5-fold cross-validation. Returns AUC as effect size, testing whether the features carry predictive signal about the target.
`cluster_and_enrich`	Applies k-means clustering to the feature space, then tests whether cluster membership is associated with a categorical variable using a chi-squared test. Returns Cramér's V as effect size.
`stratified_retest`	Re-runs a primary test (e.g., `corr_test`) separately within each stratum of a stratifying variable. Reports per-stratum results and a stability score measuring how consistently the effect holds across strata. Used to test for confounds and interaction effects.
`visual_attribute_test`	Sends individual images to a VLM, which classifies a specified visual attribute (e.g., habitat type: forest / grassland / urban). The resulting labels are tested against a metadata grouping variable using a chi-squared test with Cramér's V as effect size.
`visual_group_comparison`	Sends images from two groups to a VLM simultaneously and asks it to identify systematic visual differences. Returns a structured list of differences with per-difference confidence scores; the mean confidence serves as the effect size.

Hypothesis Generation Prompt

You are a scientific hypothesis generator for ecological data analysis.
Your task is to generate a NOVEL testable hypothesis about patterns in the provided dataset.
Dataset Summary: {dataset_summary}
Prior Knowledge (DO NOT re-test): {prior_knowledge}
Recent Discovery Results: {recent_insights}
Hypotheses to AVOID: {hypotheses_to_avoid}
Instructions:

1. Propose ONE specific, testable hypothesis that is DIFFERENT from previous attempts
2. If previous hypotheses about a topic were rejected, try a completely different angle
3. Focus on unexplored variable combinations
4. Look for interaction effects, threshold effects, or conditional relationships

Output a JSON object with: statement, scope, variables, expected_direction, risk_flags.

When images are available, the prompt is augmented with:

Visual Hypothesis Augmentation

{n_images} sample images from the dataset are attached. These are the ACTUAL ecology scene images.
Look at them to understand:

- What visual elements vary across scenes (vegetation, water, animals, terrain, sky)
- How those visual elements might relate to the numeric features
- Patterns you can SEE that the numeric features might not fully capture

Your hypothesis should be informed by what you SEE in the images, not just the numbers.

C.5.2 Experiment planning (EVALUATE)

The experiment planner receives the hypothesis and selects a statistical tool with appropriate parameters.

Experiment Planning Prompt

You are an ecological statistician planning experiments.
CRITICAL RULES FOR CODE-CLAIM ALIGNMENT:

1. The grouping variable MUST match what the hypothesis compares (species → species_name, class → class_name, kingdom → kingdom)
2. The metric MUST measure what the claim describes ("seasonal shift" = difference in latitude between seasons, not raw latitude)
3. The data slice MUST include exactly the populations the claim describes
4. If the hypothesis is about a visual property, use visual_attribute_test

Available tools: corr_test, group_diff_test, visual_attribute_test, visual_group_comparison, predictive_test, stratified_retest.
Output a JSON object with: method, feature_spec, dataset_slice_spec.

C.5.3 Reflective accumulation (REFLECT)

The REFLECT agent receives all accumulated claims and produces structured guidance.

Meta-Insight Discovery Prompt

You are a meta-science analyst. Your task is to discover META-INSIGHTS.
A Meta-Insight is NOT another finding about raw data. It is an OBSERVATION ABOUT THE PATTERN OF INSIGHTS that provides ACTIONABLE GUIDANCE for future hypothesis generation.
Types of meta-insights:

- **Confound Pattern:** A variable appears across multiple claims, suggesting it should be controlled
- **Variable Cluster:** Variables that always co-occur in supported claims, suggesting interaction effects
- **Gap:** Variables or relationships with zero coverage in the claim store
- **Success Pattern:** Certain hypothesis types have higher support rates than others
- **Contradiction:** Two claims report opposite effects, requiring stratified analysis
- **Interaction Hint:** Overlapping variables across claims imply a compound hypothesis

Dataset Summary: {dataset_summary}
Accumulated Base Insights: {meta_analysis}
Generate {n_meta_insights} meta-insights, each with: observation, actionable_recommendation, affected_variables, meta_type, priority, source_insight_ids.

The meta-insights are then used to generate guided hypotheses:

Guided Hypothesis Prompt

You are a hypothesis generator. Generate NEW TESTABLE HYPOTHESES that are GUIDED BY META-INSIGHTS.
Meta-insights tell you: what confounds to control for, what variable interactions to explore, what gaps to fill, what contradictions to resolve.
Meta-Insights: {meta_insights}
Dataset Summary: {dataset_summary}
Generate {n_hypotheses} hypotheses that directly address the meta-insights. Each hypothesis should cite which meta-insight it addresses.

Table A9: Compute budget for all experiments.

Resource	Amount
Total completed runs	∼86
Total API calls	∼8,250
Estimated input tokens	∼33M
Estimated output tokens	∼8M
Estimated API cost (Sonnet pricing)	∼$220
Wall-clock time (all experiments)	∼72 hours

D Compute resources

LLM API calls take the majority of the computing resources of this work. Each iteration of DiscoPER involves approximately three LLM API calls: one for PROPOSE (hypothesis generation), one for experiment planning, and one for REFLECT (every K iterations). A 50-iteration run completes in 30–60 minutes depending on model latency and a 100-iteration runs take 1–2 hours. Table A9 summarizes the total compute budget for all experiments reported in the paper. The multi-model experiments (Section 4.4) used Claude Sonnet 4.5, Claude Opus 4.6, GPT-5.4, and DeepSeek V4 Pro via their respective APIs. Classical causal discovery baselines (PC, GES) ran locally in under one minute each using the `causal-learn` package.

E Broader impact

DiscoPER could accelerate discovery on large observational datasets (e.g., citizen science platforms) where human analysis capacity is limited, complementing domain expertise. The held-out validation and counterfactual evaluation provide safeguards against hallucinated findings. However, system's outputs should be treated as candidates for expert review, not established facts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the contribution of our work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitation of our work in the experiments section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our work is not a theoretical work.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide our code as supplementary material and will open source them.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will open source our code and the generated data.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all details in the appendix, as well as in our accompany codes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars on our results on scaling the proposed system. The variance of the results seems to be small.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information for compute resources in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and confirm our work respects the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide discussions on the broader impacts in the appendix

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work proposes a framework for automating scientific discovery in real world datasets. No pretrained models or harmful data are generated.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the used assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We propose new datasets that is iNatDisco-800 and iNatDisco-50K, they are documented in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects are involved in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: DiscoPER uses LLMs as a central part for hypotheses generation, but the core method development doesn't involve any usage of LLMs. LLMs have been used in the implementation process in developing DiscoPER, one human author have verified all codes generated by LLMs in method development. LLMs have also been used in drafting this paper, human authors have reviewed and rephrased all LLMs generated texts.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.