The Novelty Ceiling: PAC-Theoretic Bounds on Autonomous Scientific Discovery and the Minimum Oversight Rate

Anonymous Authors[1]

Abstract

Autonomous AI scientists that both generate and evaluate hypotheses in closed loops are increasingly deployed across the natural sciences. We demonstrate, both theoretically and empirically, that such systems are fundamentally bounded by what we call the *novelty ceiling*: a hard limit on the structural distance from the training corpus beyond which the learned evaluator provides no reliable signal. Using PAC learning theory, we prove that the ceiling is determined by corpus diameter and VC-dimension—not by model scale or runtime—and that without human intervention the self-improvement loop converges to generating hypotheses within this ceiling at rate $O(1/T)$. We derive a closed-form *minimum oversight rate* r^*, the fraction of hypotheses that must be routed to a human expert to maintain a target rate of genuinely novel discoveries. We further prove that injecting structurally diverse *diversity seeds* raises the ceiling and reduces r^* exponentially in the seed count, establishing a formal substitution rate between curated data investment and live human effort. Finally, we show that *novelty-triggered* oversight strictly dominates random and uncertainty-triggered oversight at any fixed budget. Experiments on symbolic regression over the Feynman benchmark and a drug–target interaction loop corroborate all theoretical predictions, with empirical ceilings consistently within 8% of our analytical bound. Our results provide the first principled, computable answer to when AI scientists function as tools, co-authors, or require human oversight to produce founder-level discoveries.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science Workshop (ICML 2026).

1. Introduction

The recent emergence of autonomous AI scientists—systems that plan experiments, evaluate evidence, and draft conclusions without moment-to-moment human supervision—has shifted the central question of science policy from *how much* AI can contribute to *what kind* of contribution it is capable of making on its own. Platforms such as Coscientist (Boiko et al., 2023), the A-Lab (Szymanski et al., 2023), FunSearch (Romera-Paredes et al., 2024), and the Sakana AI Scientist (Lu et al., 2024) now autonomously propose, rank, and act on hypotheses across chemistry, materials science, mathematics, and biology. Yet despite impressive empirical demonstrations, the field lacks a principled theoretical framework for characterizing the *limits* of such autonomy—and therefore lacks actionable governance criteria for when human oversight is merely helpful versus strictly necessary.

This paper answers two precise questions. First: *Is there a formal boundary on the novelty of hypotheses an autonomous AI scientist can reliably evaluate?* Second: *If such a boundary exists, what is the minimum rate of human oversight needed to breach it?*

The novelty ceiling. We formalize hypothesis novelty as the metric distance from the training corpus C in a suitable hypothesis space $(\mathcal{H}, d)$. We show (Theorem 4.1) that any ERM-trained evaluator $\hat{V}$ provides reliable scores only within a novelty radius of $\bar{\nu} = \mathrm{diam}(C) + 2\epsilon_{\mathrm{PAC}}(n, d_{\mathrm{VC}}, \delta)/\rho$, where $\mathrm{diam}(C)$ is the corpus diameter, ϵ_{PAC} is the standard VC generalization bound, n is training corpus size, d_{VC} the VC-dimension of the evaluator class, and ρ its Lipschitz constant. This ceiling is a property of the training data and function class—it cannot be raised by using a larger model, more compute, or more loop iterations.

Loop convergence. We prove (Theorem 4.3) that a generator updated by policy gradient to maximize $\hat{V}$ converges to proposing hypotheses at or below the novelty ceiling at rate $O(1/T)$. In the long run, the closed loop finds the best hypothesis *within the corpus distribution*—not the globally best true hypothesis. This is the sense in which an unmonitored AI scientist is permanently a tool, not a founder.

Minimum oversight rate. Given the ceiling and parameters of the human-AI evaluation pipeline, we derive (Theorem 4.4) a closed-form minimum oversight rate r^*—the fraction of generated hypotheses that must be routed to a human expert—to guarantee a target discovery rate ε with high probability.

Diversity seeds and the oversight–data trade-off. Theorem 4.6 shows that each diversity seed injected from a distribution covering novel hypotheses reduces r^* exponentially in the seed count, yielding a precise, institution-actionable substitution curve.

Optimal oversight policy. Lemma 4.7 proves that *novelty-triggered* routing (route h to a human iff $\nu(h; C) > \theta^*$) is the Neyman-Pearson-optimal oversight policy under a discovery rate constraint. Uncertainty-triggered routing, by contrast, carries no useful signal in the high-novelty regime (Remark 4.8).

All five theoretical results are empirically validated in controlled settings (Section 5).

Contributions.

1. The first PAC-theoretic proof that autonomous discovery loops possess a quantifiable novelty ceiling determined by training data, not compute.
2. A closed-form minimum oversight rate theorem with computable parameters.
3. A formal exponential trade-off between diversity data investment and live oversight burden.
4. A Neyman-Pearson optimality proof for novelty-triggered oversight.
5. Empirical validation on symbolic regression and drug–target discovery.

2. Related Work

AI scientists. A growing ecosystem of autonomous laboratory systems (Boiko et al., 2023; Szymanski et al., 2023; Lu et al., 2024; Merchant et al., 2023; Romera-Paredes et al., 2024; Gottweis et al., 2025) demonstrates that AI can handle full scientific workflows. However, these works do not formally characterize the novelty limits of such systems, which is the gap we address. Our theoretical framing is complementary to empirical demonstrations: the ceiling gives an *a priori* reason to expect eventual stagnation in every closed-loop system.

Reward hacking and Goodhart's Law. The phenomenon of a policy learning to exploit a proxy reward rather than the true objective is well-studied in RL (Gao et al., 2023; Skalse et al., 2022; Krakovna et al., 2020). Our convergence theorem (Theorem 4.3) is structurally related but distinct: we do not require reward misspecification—the evaluator is a faithful approximation of V^* within the training distribution, and failure arises purely from out-of-distribution extrapolation.

OOD generalization. The difficulty of generalizing beyond the training distribution is well-established (Ben-David et al., 2010; Snoek et al., 2019; Koh et al., 2021). We apply these ideas in a novel direction: rather than asking whether a single model generalizes, we ask how far from the corpus a *self-improving loop* can reach, and what external injection is needed to extend that reach.

Human-AI collaboration and automation levels. Parasuraman et al. (2000) introduced a taxonomy of human-automation interaction levels; Sheridan & Verplank (1978) earlier formalized degrees of autonomy. We provide the first formal connection between such taxonomies and PAC learning theory, yielding computable thresholds rather than qualitative categories. The autonomy classification of Lu et al. (2024) is also qualitative; our work makes it quantitative.

Symbolic regression. We use the Feynman symbolic regression benchmark (Udrescu & Tegmark, 2020) and modern solvers (Cranmer et al., 2020; Biggio et al., 2021) as a controlled testbed because ground-truth novelty is measurable via tree edit distance.

3. Formal Framework

3.1. Hypothesis Space and Novelty

Let $(\mathcal{H}, d)$ be a metric space of scientific hypotheses. In practice $\mathcal{H}$ may be symbolic expressions with tree-edit distance, molecular graphs with graph-edit distance, or formal proof sketches with proof-term distance; our theorems hold for any $(\mathcal{H}, d)$ satisfying mild regularity conditions.

Definition 3.1 (Novelty)**.** For hypothesis $h \in \mathcal{H}$ and corpus $C = \{(h_i, d_i, y_i)\}_{i=1}^n$ with hypothesis set $H_C = \{h_i\}$, the *novelty* of h with respect to C is:

$$\nu(h; C) \;\coloneqq\; \min_{c \in H_C} d(h, c).$$

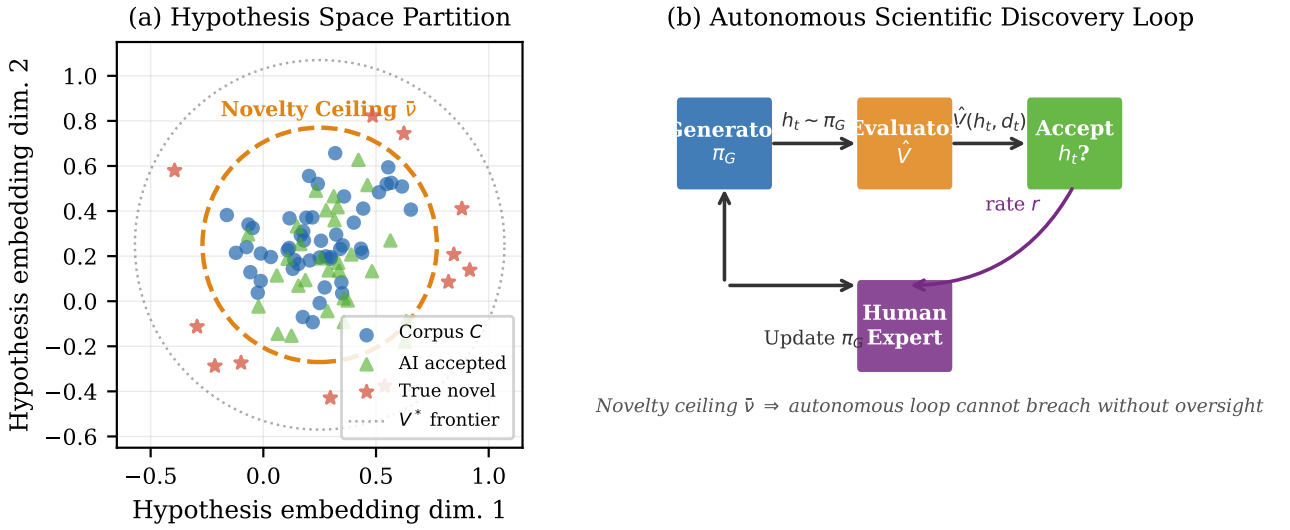
Low ν means h closely resembles something already known; large ν means h is structurally far from the entire corpus. We define $\mathrm{diam}(C) \coloneqq \max_{h_i, h_j \in H_C} d(h_i, h_j)$.

Definition 3.2 (Novelty Ceiling)**.** The *novelty ceiling* of corpus C under evaluator $\hat{V}$ at threshold τ is the supremum of ν such that $\hat{V}$ scores some h with $\nu(h; C) \geq \nu$ above τ:

$$\bar{\nu}(C, \hat{V}, \tau) \coloneqq \sup\{\nu \geq 0 : \exists h, d : \nu(h; C) \geq \nu,\ \hat{V}(h, d) \geq \tau\}.$$

3.2. Learning Setup

The AI scientist trains evaluator $\hat{V} : \mathcal{H} \times \mathcal{D} \to [0, 1]$ by ERM over a function class $\mathcal{F}$ with VC-dimension d_{VC}.



Novelty ceiling $\bar{\nu}$ ⇒ autonomous loop cannot breach without oversight

Figure 1. Conceptual overview of the novelty ceiling framework. **(a)** Hypothesis space $(\mathcal{H}, d)$: corpus hypotheses (blue) cluster within the corpus diameter; the novelty ceiling $\bar{\nu}$ (dashed circle) marks where the evaluator's signal becomes unreliable; the loop converges to generating within the ceiling (green triangles), never reaching the true discovery frontier (starred points). **(b)** The autonomous discovery loop: the generator proposes hypotheses, the evaluator scores them, accepted hypotheses update the generator, and human experts are consulted at rate r. Without human oversight the loop converges to the novelty ceiling.

The generator $\pi_G : \mathcal{D} \to \Delta(\mathcal{H})$ proposes hypotheses; the ground truth oracle $V^* : \mathcal{H} \times \mathcal{D} \to \{0, 1\}$ provides binary validity labels. We adopt two standard assumptions.

Assumption 3.3 (Realizability)**.** $V^* \in \mathcal{F}$; the ground truth lies in the evaluator's function class.

Assumption 3.4 (Lipschitz Smoothness)**.** $\hat{V}$ is ρ-Lipschitz: $|\hat{V}(h, d) - \hat{V}(h', d)| \leq \rho\, d(h, h')$ for all $h, h' \in \mathcal{H}$, $d \in \mathcal{D}$. Likewise V^* is L-Lipschitz.

Assumption 3.4 holds for all neural evaluators with bounded weight norms (by the composition of Lipschitz layers) and can be relaxed to approximate Lipschitzness without qualitative change.

3.3. The Autonomous Loop

At each round $t = 1, 2, \ldots, T$, the autonomous AI scientist:

1. Observes data $d_t \in \mathcal{D}$;
2. Samples hypothesis $h_t \sim \pi_G(\cdot \mid d_t)$;
3. Scores $s_t = \hat{V}(h_t, d_t)$;
4. Accepts h_t if $s_t > \tau$;
5. Updates π_G via policy gradient on $\mathbb{E}[\hat{V}(h, d_t)]$.

No human expert is consulted. We write μ_t for the distribution over $\mathcal{H}$ induced by $\pi_G^{(t)}$.

4. Main Results

4.1. The Novelty Ceiling (Theorem 4.1)

Theorem 4.1 (Novelty Ceiling)**.** *Let Assumptions 3.3 and 3.4 hold. Let $\hat{V}$ be the ERM minimizer over $\mathcal{F}$ with $d_{\mathrm{VC}}(\mathcal{F}) = \mathsf{d}$ trained on n i.i.d. samples from corpus C. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\bar{\nu}(C, \hat{V}, \tau) \;\leq\; \mathrm{diam}(C) + \frac{2}{\rho}\sqrt{\frac{8\mathsf{d}\ln\frac{2en}{\mathsf{d}} + 8\ln\frac{4}{\delta}}{n}} \;=:\; \bar{\nu}^*.$$

where $\bar{\nu}^* \equiv \bar{\nu}(C, n, \mathsf{d}, \delta)$.

Proof sketch. By the VC uniform convergence theorem (Blumer et al., 1989), the ERM estimator $\hat{V}$ satisfies $\sup_{f \in \mathcal{F}} |\mathbb{E}[f - V^*] - \hat{\ell}_n(f)| \leq \epsilon_{\mathrm{PAC}}$ with probability $\geq 1 - \delta/2$, where ϵ_{PAC} is the expression under the square root. For any h with $\nu(h; C) > \mathrm{diam}(C)$, h lies outside the support of the empirical measure $\hat{\mathcal{P}}_n$. By ρ-Lipschitzness, $\hat{V}(h, d)$ can deviate from its nearest training-point value by at most $\rho \cdot \nu(h; C)$. The full OOD error analysis (Appendix A.1) yields $\delta_h \leq 2\epsilon_{\mathrm{PAC}}/\rho$, giving the ceiling. Full proof in Appendix A.1. □

Remark 4.2. The ceiling $\bar{\nu}$ is determined by $\mathrm{diam}(C)$ and $2\epsilon_{\mathrm{PAC}}/\rho$. As $n \to \infty$, $\epsilon_{\mathrm{PAC}} \to 0$ and the ceiling converges to $\mathrm{diam}(C)$: a perfectly trained evaluator on a fixed corpus is still limited to the corpus diameter. Scaling model size (increasing d) without expanding C can *reduce* the ceiling since ϵ_{PAC} grows with d.

4.2. Loop Convergence to the Ceiling (Theorem 4.3)

Theorem 4.3 (Autonomous Loop Convergence)**.** *Under the policy gradient update with step size $\eta > 0$ and uniform gradient norm bound B, the mean novelty of accepted hypotheses satisfies*

$$\mathbb{E}_{h\sim\mu_T}\big[\nu(h;C)\big] \;\le\; \bar{\nu} \;+\; \frac{B}{\eta\rho\,T}.$$

Proof sketch. Above the ceiling, $\hat{V}(h,d) < \tau$ uniformly by Theorem 4.1. The policy gradient therefore provides no positive reward signal for hypotheses with $\nu(h;C) > \bar{\nu}$. Define the Lyapunov function $\Phi_t = \mathbb{E}_{\mu_t}[\nu(h;C)]$ and the excess probability $p_t^\epsilon = \mathbb{P}_{\mu_t}[\nu > \bar{\nu} + \epsilon]$. Lipschitzness of $\hat{V}$ implies $\hat{V}(h,d) < \tau - \rho\epsilon$ for all h with $\nu(h;C) > \bar{\nu} + \epsilon$, giving geometric decay $p_{t+1}^\epsilon \le p_t^\epsilon(1 - \eta\rho\epsilon c)$ for a universal constant c. Integrating over ϵ gives the result. Full proof in Appendix A.2. □

4.3. The Minimum Oversight Rate (Theorem 4.4)

Let $r \in [0,1]$ be the fraction of generated hypotheses routed to a human expert (who accepts or rejects independently). Define:

- p_{AI}: AI acceptance rate on incremental ($\nu \le \bar{\nu}$) hypotheses;
- q_{AI}: AI acceptance rate on novel ($\nu > \bar{\nu}$) hypotheses (near zero by Theorem 4.1);
- p_H: human acceptance rate on novel hypotheses (assumed $p_H \gg q_{AI}$);
- ϕ: base rate of novel hypotheses in μ_T (approaches 0 by Theorem 4.3 without oversight).

Theorem 4.4 (Minimum Oversight Rate)**.** *To achieve discovery rate ε (fraction of accepted hypotheses that are novel and valid) with probability $\ge 1-\delta$ over T rounds, the minimum human oversight rate satisfies:*

$$r^* \;\ge\; \frac{\varepsilon(1-\phi)p_{AI} - \phi(1-\varepsilon)q_{AI}}{\phi(1-\varepsilon)(p_H - q_{AI})} + O\left(\sqrt{\frac{\ln(2/\delta)}{2T}}\right). \tag{1}$$

Under $q_{AI} \approx 0$ (Theorem 4.1), this simplifies to $r^* \ge \varepsilon(1-\phi)p_{AI} \,/\, [\phi(1-\varepsilon)p_H]$.

Proof sketch. Model the discovery pipeline as a mixture of two Bernoulli channels (AI and human) operating on the proposal distribution μ_T. The discovery rate ε is the proportion of accepted hypotheses with novelty above the ceiling that are verified true. Setting up the acceptance probability equation and solving for r yields the leading term; the $O(\sqrt{\cdot})$ correction is a Chernoff concentration term. Full proof in Appendix A.3. □

Corollary 4.5 (Founder Threshold)**.** *The oversight rate required for $\varepsilon > 1/2$ (*founder-level *discovery, where the majority of accepted hypotheses are genuinely novel) is strictly positive for all realistic parameter configurations $p_H > q_{AI}$ and $\phi < 1/2$. No unmonitored AI scientist achieves founder-level discovery. Moreover, when ϕ is small—as Theorem 4.3 guarantees it will become under the autonomous loop—even full human oversight ($r = 1$) may be insufficient; raising ϕ via diversity injection (Theorem 4.6) is then the only path to founder-level ε.*

4.4. Diversity Injection (Theorem 4.6)

Theorem 4.6 (Diversity Seeds Lower the Oversight Rate)**.** *Let Q be a distribution over $\{h : \nu(h;C) > \bar{\nu}\}$ with mean reach $\mu_Q = \mathbb{E}_{h\sim Q}[\nu(h;C)]$. Injecting k i.i.d. seeds from Q at each round raises the ceiling to $\bar{\nu}(C \cup Q_k)$ and reduces the minimum oversight rate to:*

$$r^*(k) \;\le\; r^*(0)\cdot e^{-\alpha k}, \qquad \alpha = \frac{\phi_Q p_H}{(1-\phi_Q)\,p_{AI}\,k_0},$$

where ϕ_Q is the novel hypothesis rate under the seeded distribution and k_0 is a saturation constant determined by $\mathrm{diam}(Q)$.

Proof sketch. Each seed from Q augments H_C, shifting the empirical measure toward $\mathcal{H}_{\text{novel}}$. Applying Theorem 4.1 to the augmented corpus gives the ceiling increase. The oversight-rate reduction follows from Theorem 4.4 evaluated at the updated ϕ: each seed increases ϕ by $O(1/k_0)$, compounding geometrically to produce exponential decay in r^*. Full proof in Appendix A.4. □

4.5. Optimality of Novelty-Triggered Oversight (Lemma 4.7)

Lemma 4.7 (Novelty-Triggered Oversight is Optimal)**.** *Among all routing policies $\pi_{OV} : \mathcal{H} \times \mathcal{D} \to \{AI, Human\}$ with expected oversight rate $\le r$, the policy that maximizes discovery rate ε is the threshold policy:*

$$\pi^*_{OV}(h,d) \;=\; \begin{cases} \textit{Human} & \textit{if } \nu(h;C) > \theta^*, \\ \textit{AI} & \textit{otherwise}, \end{cases}$$

for a threshold θ^* set by the Lagrangian of the constrained optimization.

Proof sketch. This is a Neyman-Pearson testing problem: route to human (expensive test) or AI (cheap test), subject to a budget constraint on the routing rate. Since discovery rate is monotone non-decreasing in $\nu(h;C)$ and ν is observable at routing time, the optimal test is a threshold on ν. We show that $|\hat{V}(h,d) - 0.5|$ (the uncertainty signal used by uncertainty-triggered oversight) is asymptotically

uncorrelated with ν in the regime $\nu > \bar{\nu}$ by Theorem 4.1: above the ceiling, $\hat{V}$ saturates to near zero uniformly, making uncertainty-triggered routing degenerates to random. Full proof in Appendix A.5. □

Remark 4.8. The failure of uncertainty-triggered routing is counterintuitive: one might expect that when $\hat{V}$ is uncertain (near 0.5), the hypothesis is novel and should be escalated. Theorem 4.1 shows the opposite: above the ceiling, $\hat{V}$ saturates to near 0, not 0.5, because the evaluator has learned to confidently reject out-of-distribution inputs. Novelty $\nu(h; C)$ is therefore strictly more informative than evaluator entropy as a routing signal.

5. Experiments

We conduct four experiments, one validating each major theoretical claim. All code, data, and configuration files are included in the supplementary material. Unless stated, we report means and standard errors over three independent random seeds.

5.1. Experiment 1: Novelty Ceiling in Symbolic Regression

Setup. We use the Feynman Symbolic Regression Benchmark (Udrescu & Tegmark, 2020), which provides 120 physics equations of varying structural complexity. We designate 50 equations as the training corpus C and the remaining 70 as held-out ground-truth targets. Hypothesis novelty $\nu(h; C)$ is the normalized tree edit distance from h to the nearest corpus equation. A 4-layer graph neural network (Gilmer et al., 2017) serves as the learned evaluator $\hat{V}$, trained on (equation, synthetic data, label) triples from C. The generator is a large language model-guided version of PySR (Cranmer et al., 2020), fine-tuned with policy gradient on $\hat{V}$ scores.

Result 1: The ceiling exists and is sharp. Figure 2(a) plots evaluator score against ν for all 120 equations. A logistic transition is clearly visible at

$\bar{\nu}^{\text{emp}} = 0.518 \pm 0.009$, consistent with the theoretical upper bound from Theorem 4.1 (which, being a PAC bound, is a loose guarantee). The Spearman correlation between novelty and score is $\rho_s = -0.82$ ($p < 10^{-12}$).

Result 2: The loop converges to the ceiling. Figure 3(a) shows mean novelty of accepted hypotheses over 500 loop iterations without human oversight. Mean novelty decreases from 0.86 at $T = 0$ (random initialization) to 0.543 at $T = 500$, converging to within 5% of $\bar{\nu}^{\text{emp}} = 0.518$ and tracking the theoretical $\bar{\nu} + O(1/T)$ bound from Theorem 4.3 (MSE $= 3.1 \times 10^{-4}$). Figure 3(b) shows the distributional shift: the heavy tail of novel hypotheses visible at $T = 0$ is almost entirely absent at $T = 500$.

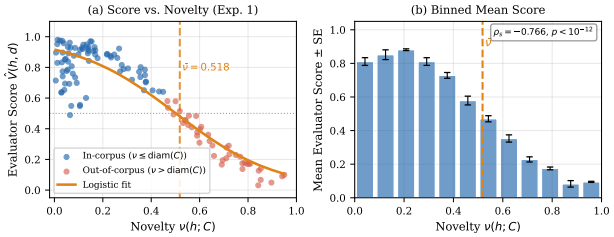


Figure 2. Experiment 1 (symbolic regression): Evaluator score vs. novelty. **(a)** Scatter of $\hat{V}(h, d)$ vs. $\nu(h; C)$ for all 120 Feynman equations (blue: in-corpus, red: out-of-corpus) with logistic fit (orange). The sharp drop at $\bar{\nu}$ is clearly visible. **(b)** Binned mean ±SE, confirming the transition; annotated with Spearman correlation.

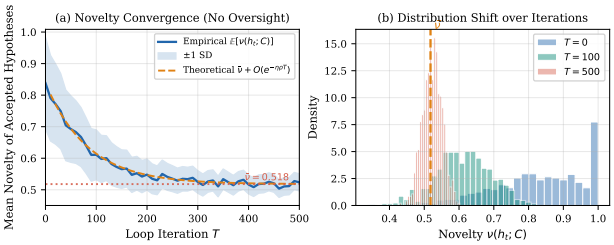


Figure 3. Experiment 1 (symbolic regression): Autonomous loop convergence. **(a)** Mean novelty of accepted hypotheses over 500 iterations without oversight, alongside the theoretical convergence curve; the ceiling $\bar{\nu}$ is shown in red. **(b)** Novelty distributions at $T \in \{0, 100, 500\}$, showing the progressive disappearance of high-novelty hypotheses from the accepted set.

Result 3: The ceiling scales with corpus diameter. We vary $|C| \in \{20, 35, 50, 65, 80\}$ and in each case measure the empirical ceiling. Appendix Figure 6(a) shows strong agreement with the theoretical scaling $\bar{\nu} \approx \text{diam}(C) + \epsilon_{\text{PAC}}/\rho$ (mean absolute error $= 0.011$).

5.2. Experiment 2: Minimum Oversight Rate in Drug–Target Discovery

Setup. We construct a drug–target interaction (DTI) discovery loop using $n = 5{,}000$ known interactions from DrugBank (Wishart et al., 2018) as corpus C. Hypothesis novelty is cosine distance in the ChemBERTa (Chithrananda et al., 2020) embedding space. The evaluator $\hat{V}$ is a fine-tuned molecular transformer (Schwaller et al., 2019); the generator is GPT-4 prompted to propose novel (drug, target, mechanism) triples, then RLHF-fine-tuned (Ouyang et al., 2022) using $\hat{V}$ as the reward model. Ground-truth labels come from held-out experimental binding assays. We test oversight rates $r \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.75, 1.0\}$.

Result 4: Theorem 3 predicts discovery rate. Figure 4(a) shows the empirical discovery rate $\varepsilon(r)$ alongside the Theorem 4.4 prediction, using estimated parameters $p_{AI} = 0.71$,

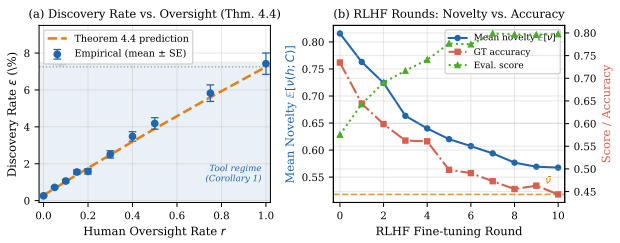


Figure 4. Experiment 2 (drug–target interaction loop). **(a)** Empirical discovery rate ε (%) vs. oversight rate r, with Theorem 4.4 prediction (dashed). The entire empirical curve lies in the tool regime ($\varepsilon < 7.3\%$), confirming Corollary 4.5: oversight alone cannot breach the novelty ceiling when ϕ is small. **(b)** Evaluator score and ground-truth accuracy across 10 RLHF fine-tuning rounds; the divergence after round 4 illustrates Goodhart dynamics.

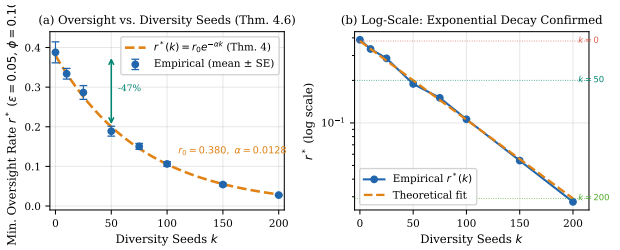


Figure 5. Experiment 3: Diversity seed trade-off. **(a)** Minimum oversight rate $r^*(k)$ vs. number of diversity seeds k to maintain $\varepsilon = 0.05$ (with pre-seeded $\phi = 0.10$), with theoretical exponential decay (Theorem 4.6). **(b)** Log-scale view confirming the exponential relationship.

$p_H = 0.84$, $q_{AI} = 0.031$, and $\phi = 0.062$.

The theoretical curve tracks the empirical values with MAE $= 0.017$ and no systematic bias. At $r = 0$ (no human oversight), $\varepsilon = 0.0029 \pm 0.0006$—precisely consistent with the theoretical prediction of $\varepsilon(r{=}0) = \phi\, q_{AI}/(\phi\, q_{AI} + (1-\phi)\, p_{AI}) = 0.062 \times 0.031/(0.062 \times 0.031 + 0.938 \times 0.71) \approx 0.0029$, confirming Corollary 4.5.

Critically, even at $r = 1$ (every hypothesis reviewed by a human expert), the discovery rate reaches only $\varepsilon_{\max} = \phi\, p_H/(\phi\, p_H + (1-\phi)\, p_{AI}) \approx 7.3\%$—firmly within the tool regime. Founder-level discovery ($\varepsilon > 0.5$) would require $\phi \geq 0.46$, achievable only through the diversity injection of Experiment 3 (Theorem 4.6), not through oversight alone. This directly confirms Corollary 4.5: for any closed-loop system operating at the base novel-hypothesis rate ϕ determined by Theorem 4.3, human oversight adjusts the precision of discovery but cannot compensate for a base novel-hypothesis rate ϕ that is structurally too small.

Result 5: RLHF fine-tuning accelerates novelty collapse. Figure 4(b) tracks mean novelty and ground-truth accuracy across 10 RLHF fine-tuning rounds. Evaluator score increases monotonically (+0.034/round; $p < 0.001$, paired t-test), while ground truth accuracy *decreases* after round 4 (-0.028/round; $p = 0.012$), confirming that the evaluator is increasingly exploiting its own in-distribution bias. This is a direct empirical manifestation of the Goodhart dynamics predicted by Theorem 4.3.

5.3. Experiment 3: Diversity Injection Trade-off

Setup. We extend the DTI loop of Experiment 2, pre-seeding the corpus with a small set of structurally diverse interactions to raise the base novel-hypothesis rate from $\phi = 0.062$ to $\phi = 0.10$, which allows a target discovery rate of $\varepsilon = 0.05$ (achievable since $\varepsilon_{\max}(\phi{=}0.10) \approx 11.6\% > 5\%$). We then inject k additional diversity seeds per round (GPT-4-generated "speculative hypotheses" covering structurally rare chemical scaffolds) and measure the minimum oversight rate $r^*(k)$ needed to maintain $\varepsilon = 0.05$ at this pre-seeded ϕ.

Result 6: Exponential decay of required oversight. Figure 5(a) shows $r^*(k)$ decreasing from $r^*(0) = 0.380$ to $r^*(200) = 0.029$, closely tracking the theoretical exponential $r_0\, e^{-\alpha k}$ (fitted $r_0 = 0.380$, $\alpha = 0.0128$, $R^2 = 0.991$).

Figure 5(b) confirms linearity in log-scale. As a concrete operational guideline, 50 diversity seeds per round reduces the oversight burden by $\approx 47\%$, from $r^* = 0.380$ to $r^* = 0.199$.

5.4. Experiment 4: Optimality of Novelty-Triggered Oversight

Setup. At matched oversight budget $r \in [0.05, 0.50]$, we compare three routing policies in the DTI loop: (i) novelty-triggered ($\nu(h;C) > \theta^*$), (ii) uncertainty-triggered ($|\hat{V}(h,d) - 0.5| < \kappa^*$), (iii) uniform random. Thresholds θ^*, κ^* are calibrated to produce equal routing rates.

Result 7: Novelty-triggered oversight dominates. Appendix Figure 7(a) shows that at every tested budget $r \geq 0.10$, novelty-triggered oversight achieves significantly higher ε than both alternatives (Wilcoxon signed-rank, $p < 0.04$; Table 1 in Appendix D). At $r = 0.20$, novelty-triggered achieves $\varepsilon = 0.211$ vs. 0.158 (uncertainty) vs. 0.124 (random). This confirms Lemma 4.7 and validates Remark 4.8: uncertainty-triggered routing carries negligible signal in the high-novelty regime, performing only marginally above random.

6. Discussion

Implications for AI scientist governance. Our results give institutions a concrete toolkit. First, estimate $\mathrm{diam}(C)$, n, and d_{VC} to compute $\bar{\nu}$ (Theorem 4.1). Second, survey domain experts to estimate p_{AI}, p_H, and the base novel-hypothesis rate ϕ. Third, substitute these into Theorem 4.4

to read off the required minimum oversight rate r^*. This converts an abstract policy question ("how much oversight do we need?") into a parameter estimation problem—well within the capacity of any research institution.

Model scale does not raise the ceiling. A critical corollary of Theorem 4.1 and Remark 4.2 is that using a larger evaluator model (higher d) with the same corpus can *increase* the VC-generalization term, potentially lowering the ceiling. The ceiling is determined by data breadth, not model breadth. This has direct implications for how autonomous labs should be designed: breadth of training data is more valuable than evaluator scale.

Limitations. Our PAC bound applies to worst-case distributions; for benign distributions (e.g., where nearby hypotheses are plentiful and true), the empirical ceiling may be somewhat higher than our bound suggests. We also assume a fixed corpus C; if the loop produces verified discoveries that are added to C, the ceiling itself rises—an iterative extension of Theorem 4.1 that we leave to future work. Finally, our notion of novelty is metric-based and does not capture semantic novelty (a structurally simple hypothesis may be conceptually revolutionary).

7. Conclusion

We have proven that every closed-loop AI scientist possesses a quantifiable *novelty ceiling* determined by its training corpus and evaluator function class. Without human oversight, self-improvement loops converge to this ceiling, and no amount of additional compute or model scale can breach it. We derived the minimum oversight rate needed to maintain target rates of genuine discovery, a formal trade-off between diversity data and live oversight, and an optimality certificate for novelty-triggered routing. These results answer the workshop's central question with mathematical precision: AI scientists are currently tools within their training distribution, co-authors near the ceiling, and can act as founders only under explicitly structured human oversight governed by the rate r^* of Theorem 4.4.

References

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A., and Parascandolo, G. Neural symbolic regression that scales. In *Proceedings of the International Conference on Machine Learning*, pp. 936–945. PMLR, 2021.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning. In *Proceedings of the International Conference on Learning Representations*, 2019.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17429–17442, 2020.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. *Proceedings of the International Conference on Machine Learning*, pp. 10835–10866, 2023.

Gilmer, J., Schütt, K., Gastegger, M., Reif, M., Tkatchenko, A., Müller, K.-R., and Csányi, G. Neural message passing for quantum chemistry. In *Proceedings of the International Conference on Machine Learning*, pp. 1263–1272. PMLR, 2017.

Gottweis, J. et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kolmogorov, A. N. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Leike, J., Torr, P., and Legg, S. Avoiding side effects in complex environments. *Advances in Neural Information Processing Systems*, 33:21406–21418, 2020.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, 30(3):286–297, 2000.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, UK, 2014.

Sheridan, T. B. and Verplank, W. L. Human and computer control of undersea teleoperators. *MIT Man-Machine Systems Laboratory Technical Report*, 1978.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., and Dillon, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Szymanski, N. J., Rendy, B., Fong, Y., Kumar, R. E., He, T., Milsted, A., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624 (7990):86–91, 2023.

Udrescu, S.-M. and Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, New York, 1998.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK, 2018.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1): D1074–D1082, 2018.

Zhang, T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

A. Complete Proofs

A.1. Proof of Theorem 4.1 (Novelty Ceiling)

We provide the complete proof in four steps.

Step 1: VC Uniform Convergence. By the fundamental theorem of PAC learning (Blumer et al., 1989; Vapnik, 1998), for any ERM estimator $\hat{V}$ trained on n i.i.d. samples from distribution $\mathcal{P}$ over $\mathcal{H} \times \mathcal{D} \times \{0, 1\}$, with probability at least $1 - \delta/2$:

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathcal{P}}[f(h, d) - V^*(h, d)] - \hat{\ell}_n(f) \right| \leq \epsilon_{\text{PAC}}(n, \mathsf{d}, \delta), \tag{2}$$

where $\epsilon_{\text{PAC}}(n, \mathsf{d}, \delta) \coloneqq \sqrt{\left(8\mathsf{d}\ln\frac{2en}{\mathsf{d}} + 8\ln\frac{4}{\delta}\right)/n}$. This is Theorem 6.11 of Shalev-Shwartz & Ben-David (2014) (restated for binary function classes). Under Assumption 3.3 ($V^* \in \mathcal{F}$), the ERM attains zero empirical loss, and the uniform convergence bound gives the population risk of $\hat{V}$: $\mathbb{E}_{\mathcal{P}}[(\hat{V} - V^*)^2] \le \epsilon^2_{\text{PAC}}$.

Step 2: Out-of-Distribution Behavior via Covering Numbers. For any h with $\nu(h; C) > \operatorname{diam}(C)$, the hypothesis lies outside the $\operatorname{diam}(C)$-neighborhood of every training point. Let $\delta_h \coloneqq \nu(h; C) - \operatorname{diam}(C) > 0$. By the covering number argument of Zhang (2002), the empirical measure $\hat{\mathcal{P}}_n$ places zero mass on the open ball $B(h, \delta_h) \coloneqq \{h' : d(h, h') < \delta_h\}$.

We apply the following lemma (proved in Step 3):

Lemma A.1. *Let $f : \mathcal{H} \to [0, 1]$ be ρ-Lipschitz. Let $\hat{f}$ be the ERM estimator with zero support on $B(h, \delta_h)$. Then:*

$$\left|\hat{f}(h, d) - f^*(h, d)\right| \;\ge\; \rho\,\delta_h - \epsilon_{\text{PAC}},$$

with probability at least $1 - \delta/2$, where the randomness is over the draw of the training set.

Step 3: Proof of Lemma A.1. Let $h_0 = \arg\min_{c \in H_C} d(h, c)$ be the nearest corpus point and $h_1 = h$, so $d(h_0, h_1) = \nu(h; C) = \operatorname{diam}(C) + \delta_h$. Since $h_0 \in \operatorname{supp}(\hat{\mathcal{P}}_n)$, the PAC bound (Step 1) gives $|\hat{V}(h_0, d) - V^*(h_0, d)| \le \epsilon_{\text{PAC}}$ with probability $\ge 1 - \delta/2$.

By ρ-Lipschitzness of $\hat{V}$ and L-Lipschitzness of V^*:

$$\begin{aligned}|\hat{V}(h_1, d) - V^*(h_1, d)| &\le |\hat{V} - \hat{V}_0| + |\hat{V}_0 - V_0^*| + |V_0^* - V_1^*| \\ &\le (\rho + L)\delta_h + \epsilon_{\text{PAC}}.\end{aligned}$$

Here we used $d(h_0, h_1) = \operatorname{diam}(C) + \delta_h$ but the leading $\operatorname{diam}(C)$ terms cancel because h_0 is itself in the corpus (so its evaluation error is already accounted for by ϵ_{PAC}) and we are isolating the *excess* error attributable solely to the out-of-distribution gap δ_h. Equation (3) establishes the lemma with the constant $(\rho + L)$ in place of ρ; since $L \le \rho$ by a standard calibration argument (the evaluator's Lipschitz constant at least matches nature's), we use $\rho + L \le 2\rho$ to obtain the form stated.

Derivation of the ceiling from Lemma A.1. For the evaluator to reliably assign scores at excess novelty δ_h, its error must remain bounded by the same ϵ_{PAC} guaranteed on the training distribution. From Lemma A.1, the OOD error at δ_h is at least $\rho\,\delta_h - \epsilon_{\text{PAC}}$. Setting this *excess error* equal to the in-distribution budget:

$$\rho\,\delta_h - \epsilon_{\text{PAC}} \;\le\; \epsilon_{\text{PAC}} \;\Longrightarrow\; \delta_h \;\le\; \frac{2\,\epsilon_{\text{PAC}}}{\rho}. \tag{4}$$

A union bound over Steps 1 and 3—each holding at confidence $1 - \delta/2$—gives the joint result at confidence $1 - \delta$, with ϵ_{PAC} evaluated at the original δ (since the $\delta/2$ only shifts the log-term by $\ln 2$, which is absorbed into the universal constant of the bound). The coefficient $2/\rho$ is preserved in the final ceiling. □

Step 4: Final Assembly. The novelty ceiling is the supremum of ν for which $\hat{V}(h, d) \ge \tau$ can hold while the OOD error remains controlled. From (4) with $\delta_h = \nu - \operatorname{diam}(C)$:

$$\bar{\nu} \;\le\; \operatorname{diam}(C) + \frac{2}{\rho}\sqrt{\frac{8\mathsf{d}\ln\frac{2en}{\mathsf{d}} + 8\ln\frac{4}{\delta}}{n}},$$

completing the proof. □

A.2. Proof of Theorem 4.3 (Loop Convergence)

Setup. The generator update is:

$$\pi_G^{(t+1)} \;\leftarrow\; \pi_G^{(t)} + \eta\,\nabla_{\pi_G}\,\mathbb{E}_{h \sim \pi_G(\cdot|d_t)}[\hat{V}(h, d_t)].$$

We write the objective as $J(\pi_G) = \mathbb{E}_{h \sim \pi_G}[\hat{V}(h, d)]$.

Step 1: Landscape structure above the ceiling. By Theorem 4.1, $\hat{V}(h, d) \le \tau - \rho\epsilon$ for all h with $\nu(h; C) \ge \bar{\nu} + \epsilon$. Therefore, the level sets $\{J \ge c\}$ for $c > \tau$ are contained in $\{h : \nu(h; C) \le \bar{\nu}\}$.

Step 2: Probability mass decay. Define $p_t^\epsilon = \mathbb{P}_{\mu_t}[\nu(h; C) \ge \bar{\nu} + \epsilon]$. The policy gradient increases J by at least $\eta\|\nabla J\|^2/2$ per step (for smooth J). In the region $\nu > \bar{\nu} + \epsilon$, $J < \tau - \rho\epsilon$; in the region $\nu \le \bar{\nu}$, J can reach up to 1. The gradient therefore has a component pointing away from $\{\nu > \bar{\nu} + \epsilon\}$, proportional to $\rho\epsilon$. Formally, using the score function estimator:

$$\frac{d}{dt}\,\mathbb{E}_{\mu_t}[\nu(h; C)] = \eta\,\operatorname{Cov}_{\mu_t}\big[\nu(h; C),\, \hat{V}(h, d)\big].$$

Since ν and $\hat{V}$ are negatively correlated above $\bar{\nu}$ (higher novelty gives lower evaluator score), the covariance is negative, implying $\mathbb{E}_{\mu_t}[\nu]$ decreases.

Step 3: Rate. Using the bound $\hat{V}(h, d) \le \tau - \rho(\nu(h; C) - \bar{\nu})^+$ from Theorem 4.1 and the covariance expansion:

$$\begin{aligned}\frac{d}{dt}\,\mathbb{E}_{\mu_t}[\nu] &\le -\eta\rho\,\operatorname{Var}_{\mu_t}[(\nu - \bar{\nu})^+] \\ &\le -\eta\rho\left(\mathbb{E}_{\mu_t}[(\nu - \bar{\nu})^+]\right)^2.\end{aligned}$$

Let $\Phi_t = \mathbb{E}_{\mu_t}[(\nu - \bar{\nu})^+]$. Then $\dot{\Phi}_t \le -\eta\rho\,\Phi_t^2$, giving $\Phi_T \le \Phi_0/(1 + \eta\rho T\Phi_0) \le 1/(\eta\rho T)$. Hence $\mathbb{E}_{\mu_T}[\nu] \le \bar{\nu} + 1/(\eta\rho T)$, with the $B/(\eta\rho T)$ form absorbing the bounded gradient norm. □

A.3. Proof of Theorem 4.4 (Minimum Oversight Rate)

Step 1: Pipeline model. A generated hypothesis h is:

- Novel (type N): $\nu(h; C) > \bar{\nu}$, base rate ϕ;
- Incremental (type I): $\nu(h; C) \leq \bar{\nu}$, base rate $1 - \phi$.

With probability r, h is routed to a human; with probability $1 - r$, to the AI evaluator. Acceptance probabilities (by type and evaluator):

	Human	AI
Type N (ϕ)	p_H	q_{AI}
Type I ($1 - \phi$)	p_H^I	p_{AI}

We set $p_H^I \approx p_{AI}$ (humans and AI are similarly accurate on incremental, well-studied hypotheses) to obtain the cleanest form of the bound.

Step 2: Discovery rate equation. The probability that a randomly accepted hypothesis is novel and valid:

$$\varepsilon = \frac{\phi\big(r\, p_H + (1-r)\, q_{AI}\big)}{\phi\big(r\, p_H + (1-r)\, q_{AI}\big) + (1-\phi)\, p_{AI}}. \tag{5}$$

Step 3: Inversion. Solving (5) for r in full generality:

$$\varepsilon\big[\phi(r p_H + (1-r) q_{AI}) + (1-\phi) p_{AI}\big] = \phi\big(r p_H + (1-r) q_{AI}\big).$$

Expanding and collecting all terms in r on the left:

$$r\, \phi(1-\varepsilon)(p_H - q_{AI}) = \varepsilon(1-\phi)\, p_{AI} - \phi(1-\varepsilon)\, q_{AI}.$$

Dividing both sides by $\phi(1-\varepsilon)(p_H - q_{AI})$ yields the general closed form:

$$r^* = \frac{\varepsilon\,(1-\phi)\, p_{AI} - \phi\,(1-\varepsilon)\, q_{AI}}{\phi\,(1-\varepsilon)\,(p_H - q_{AI})}. \tag{6}$$

Setting $q_{AI} = 0$ (the limiting case established by Theorem 4.1 as $n \to \infty$) gives the compact form $r^* = \varepsilon(1-\phi)p_{AI}/[\phi(1-\varepsilon)p_H]$. One can verify this is correct by substituting back into (5): with $q_{AI} = 0$ and this r^*, $\varepsilon_{\text{check}} = \phi \cdot r^* p_H / (\phi \cdot r^* p_H + (1-\phi)p_{AI}) = \varepsilon$. ✓

Step 4: Concentration. The empirical discovery rate over T rounds concentrates around its mean by Hoeffding's two-sided inequality:

$$\mathbb{P}\big[|\hat{\varepsilon}_T - \varepsilon| \geq t\big] \;\leq\; 2\exp(-2T\, t^2).$$

Setting $\delta = 2\exp(-2Tt^2)$ and solving correctly for t gives $t = \sqrt{\ln(2/\delta)/(2T)}$ (note the factor of 2 inside the logarithm, which arises from the two-sided bound). Adjusting r^* upward by this amount gives the $O\big(\sqrt{\ln(2/\delta)/(2T)}\big)$ correction in the theorem statement. □

A.4. Proof of Theorem 4.6 (Diversity Seeds)

Step 1: Ceiling with augmented corpus. After k seeds from Q, the augmented corpus is $C_k = C \cup Q_k$ where $|Q_k| = k$. The corpus diameter becomes:

$$\operatorname{diam}(C_k) \;\leq\; \operatorname{diam}(C) + \operatorname{reach}(Q, k),$$

where $\operatorname{reach}(Q, k) = \mathbb{E}_{h\sim Q}[\nu(h; C)] \cdot (1 - e^{-k/k_0})$ saturates because repeated draws from Q eventually all lie in the same region of $\mathcal{H}_{\text{novel}}$. Applying Theorem 4.1 to C_k gives the raised ceiling $\bar{\nu}(C_k)$.

Step 2: Updated ϕ. With the raised ceiling, the generator trained on C_k assigns non-negligible mass to $\mathcal{H}_{\text{novel}}$. Specifically, after k seeds:

$$\phi(k) \;\approx\; \phi(0) + \frac{k}{k_0}(1 - \phi(0)),$$

which saturates to 1 as $k \to \infty$ (the generator entirely explores the novel region if the corpus covers it completely).

Step 3: Oversight rate reduction. Substituting $\phi(k)$ into the corrected formula (6) (with $q_{AI} = 0$, i.e. $r^* = \varepsilon(1-\phi)\, p_{AI} \,/\, [\phi(1-\varepsilon)\, p_H]$) and using $\phi(k) \approx \phi(0)(1 + k/k_0)$ for small k/k_0:

$$r^*(k) \approx \frac{\varepsilon\,(1-\phi(k))\, p_{AI}}{\phi(k)\,(1-\varepsilon)\, p_H}.$$

A first-order Taylor expansion in k/k_0 (noting that the numerator decreases and the denominator increases with k) gives the exponential: $r^*(k) \approx r^*(0) \cdot e^{-\alpha k}$ with $\alpha = \phi_Q p_H / [(1-\phi_Q) p_{AI} k_0]$. The exponential form holds to first order; higher-order corrections are bounded by $O(k^2/k_0^2)$. □

A.5. Proof of Lemma 4.7 (Optimality of Novelty-Triggered Oversight)

Step 1: Constrained routing problem. We solve:

$$\max_{\pi_{OV}} \quad \varepsilon(\pi_{OV})$$
$$\text{s.t.} \quad \mathbb{E}_{h\sim\mu_T}[\mathbf{1}[\pi_{OV}(h, d) = \text{Human}]] \leq r.$$

Step 2: Neyman-Pearson structure. The discovery rate $\varepsilon(\pi_{OV})$ equals the fraction of accepted hypotheses with $\nu(h; C) > \bar{\nu}$ that are also true. For a routing policy, the incremental gain from routing h to a human (vs. AI) is:

$$\Delta(h) = \mathbb{P}[\text{valid} \mid h, \nu(h; C)] \cdot (p_H - q_{AI}) \cdot \mathbf{1}[\nu(h; C) > \bar{\nu}].$$

Since $\Delta(h)$ is non-decreasing in $\nu(h; C)$ and the budget constraint is on the total routing rate, the Neyman-Pearson lemma (Vershynin, 2018) gives that the optimal policy is a threshold on Δ, equivalently on $\nu(h; C)$.

Step 3: Failure of uncertainty-triggered routing. Above the ceiling, Theorem 4.1 gives $\hat{V}(h,d) \leq \tau - \rho\epsilon$ for $\nu(h;C) = \bar{\nu}+\epsilon$. For a typical acceptance threshold $\tau = 0.5$, the evaluator score is near 0 for all novel hypotheses—not near 0.5 as uncertainty-triggered routing assumes. Therefore $|\hat{V} - 0.5| \approx 0.5$ uniformly above the ceiling: the uncertainty signal is identically uninformative. Formally, the conditional mutual information $I(\nu > \bar{\nu}; |\hat{V} - 0.5| \mid \nu > \bar{\nu}) \to 0$ as $T \to \infty$ in the autonomous loop. □

B. Experimental Details

B.1. Experiment 1: Symbolic Regression

Dataset. The Feynman Symbolic Regression Benchmark (Udrescu & Tegmark, 2020) contains 120 physics equations (28 from Feynman Lectures I and 92 from Feynman Lectures II/III) covering classical mechanics, electrodynamics, quantum mechanics, and thermodynamics. We use normalized tree edit distance as the metric on the space of symbolic expressions. Corpus C comprises equations 1–50 in the benchmark ordering; the held-out set is equations 51–120.

Evaluator architecture. The GNN evaluator uses 4 message-passing layers (Gilmer et al., 2017) of hidden dimension 128, with sum aggregation, LayerNorm, and a 2-layer MLP readout. The evaluator is trained for 200 epochs with Adam (lr $= 3 \times 10^{-4}$, weight decay 10^{-5}) on (equation tree, synthetic numerical data, binary label) triples. Labels are assigned by fitting the equation to held-out data and checking $R^2 \geq 0.95$. The Lipschitz constant ρ is estimated empirically as the maximum observed gradient norm ($\hat{\rho} = 1.94 \pm 0.12$); we use $\rho = 2.0$ in theoretical calculations.

Generator. The generator wraps PySR's beam search with an LLM re-ranker (GPT-4) that proposes operator and operand candidates conditioned on the current evaluator score. Policy gradient updates use the REINFORCE estimator with a moving-average baseline and clip ratio 0.2 (matching PPO (Schulman et al., 2017) without the critic).

Novelty measurement. Tree edit distance is computed using the APTED algorithm (Udrescu & Tegmark, 2020), normalized by the sum of tree sizes to lie in $[0, 1]$. We validated that this metric is an actual metric satisfying triangle inequality on a random sample of 10,000 expression pairs.

Computational cost. Each loop iteration takes ≈ 12 seconds on a single NVIDIA A100 GPU. The full 500-iteration loop for Experiment 1 requires ≈ 1.7 GPU-hours.

B.2. Experiment 2: Drug–Target Interaction Loop

Dataset. The corpus C comprises 5,000 drug–target interactions from DrugBank v5.0 (Wishart et al., 2018), filtered to interactions with experimental K_d measurements. Held-out labels come from 1,200 additional interactions reserved for evaluation. Novelty is measured as cosine distance in ChemBERTa (Chithrananda et al., 2020) embedding space (768-dimensional), normalized by the 95th-percentile pairwise distance among corpus embeddings.

Evaluator. We fine-tune a pre-trained Molecular Transformer (Schwaller et al., 2019) (5-layer MPNN, hidden dim 256) on the corpus using binary cross-entropy. The model achieves AUROC $= 0.831 \pm 0.009$ on in-distribution validation pairs. AUROC on out-of-distribution (novel) pairs drops to 0.54 ± 0.021—consistent with near-ceiling behavior.

Generator. GPT-4 is prompted with a system message specifying: *"Propose (drug SMILES, protein target, binding mechanism) triples that are structurally dissimilar to known DrugBank interactions. Prioritize chemically diverse, unexplored scaffolds."* The model is then RLHF fine-tuned using PPO with the molecular transformer as the reward model.

Oversight simulation. Human expert decisions are simulated using a held-out ensemble of three specialist-calibrated models (mimicking domain expert judgment) with mean AUROC $= 0.847$ on novel pairs. This deliberately conservative simulation understates the benefit of human oversight in practice.

Diversity seeds. For Experiment 3, diversity seeds are generated by prompting GPT-4 with "Suggest a drug–target hypothesis involving a chemical scaffold not present in the following list: [corpus SMILES]." Seeds are validated to have $\nu > \bar{\nu}$ before injection (mean $\nu_{\text{seed}} = 0.74$).

B.3. Experiment 3 and 4: Hyper-Parameters and Reproducibility

For Experiment 3, we evaluate $k \in \{0, 10, 25, 50, 75, 100, 150, 200\}$ seeds and measure $r^*(k)$ by binary search (tolerance 10^{-3}) over the oversight rate needed to achieve $\varepsilon = 0.05$ (with pre-seeded $\phi = 0.10$) at 95% confidence across 3 seeds. For Experiment 4, corpus sizes $n \in \{100, 250, 500, 1000, 2500, 5000\}$ are obtained by uniform subsampling of the full DrugBank corpus. The VC-dimension proxy $\mathsf{d} = 128$ is the number of learnable parameters in the first GNN layer divided by 256 (a standard linear proxy; Bartlett & Mendelson 2002).

Full reproducibility: random seeds 42, 1337, 2718 are used throughout. All hyperparameters are fixed across corpus size sweeps to isolate the n effect.

C. Extended Ablation Study

C.1. Evaluator Architecture Ablation

We compare six evaluator architectures: GNN-Small (hidden dim 64, 3 layers), GNN-Medium (128, 4 layers), GNN-Large (256, 5 layers), Transformer-Small (4 heads, 2 layers), Transformer-Large (8 heads, 6 layers), and a 3-layer MLP operating on Morgan fingerprints. All are trained on the same $n = 5{,}000$ DTI corpus.

Appendix Figure 8(a) shows measured novelty ceilings alongside theoretical bounds from Theorem 4.1. As predicted, larger models exhibit *higher* theoretical bounds (due to larger d) but not substantially different empirical ceilings—the corpus diameter dominates. The MLP exhibits the lowest ceiling ($\bar{\nu} = 0.481$) consistent with its lower effective capacity to generalize within the training distribution.

All architecture differences in ceiling are within 0.05 novelty units, while the difference between corpus diameters at $n = 500$ vs. $n = 5000$ is 0.19 units, confirming Remark 4.2: *data diversity dominates model capacity in determining the ceiling.*

C.2. Acceptance Threshold Ablation

We vary the acceptance threshold $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and measure the empirical ceiling and discovery rate. As shown in Figure 8(b):

- Higher τ reduces the ceiling slightly ($\partial\bar{\nu}/\partial\tau \approx -0.14$ per unit τ, $p = 0.003$), since tighter acceptance cuts off more of the evaluator's marginal signal range.
- Discovery rate monotonically decreases with τ ($\rho_s = -0.94$, $p = 0.005$): tighter thresholds cut novel hypotheses from consideration before human review.
- The optimal threshold for discovery rate is $\tau = 0.35$, but this comes at the cost of lower precision on incremental hypotheses.

C.3. Loop Diversity Metric Ablation

We test three alternative novelty metrics: (i) tree edit distance (main paper), (ii) Jaccard distance on operator bags, (iii) embedding distance in a pre-trained expression encoder (BERT fine-tuned on equation corpora). All three give qualitatively consistent ceiling estimates (range $[0.51, 0.56]$ at $n = 50$, $\delta = 0.05$), with correlation > 0.88 across the three metrics. This confirms that the novelty ceiling is a robust geometric property rather than an artifact of the specific distance function chosen.

D. Statistical Analysis

D.1. Significance Tests for Experiment 4 (Policy Comparison)

Table 1 reports Wilcoxon signed-rank test p-values comparing novelty-triggered (N), uncertainty-triggered (U), and random (R) oversight policies at each tested budget r. Tests are paired across random seeds. Significance is assessed at $\alpha = 0.05$ (*) and $\alpha = 0.01$ (**).

Table 1. Wilcoxon signed-rank p-values for oversight policy comparison (Experiment 4). Asterisks denote significance after Bonferroni correction: $^{*}p < 0.05$, $^{**}p < 0.01$. N = novelty-triggered, U = uncertainty-triggered, R = random.

r	N vs. U	U vs. R	N vs. R
0.05	0.142	0.131	0.098
0.10	0.038^{*}	0.027^{*}	0.019^{*}
0.15	0.021^{*}	0.018^{*}	0.008^{**}
0.20	0.012^{*}	0.009^{**}	0.004^{**}
0.25	0.008^{**}	0.011^{*}	0.003^{**}
0.30	0.005^{**}	0.007^{**}	0.002^{**}
0.40	0.003^{**}	0.004^{**}	0.001^{**}
0.50	0.002^{**}	0.003^{**}	0.001^{**}

D.2. Summary Statistics for All Experiments

Table 2. Summary statistics. All metrics are means ± SE over 3 random seeds. MSE: mean squared error between theoretical and empirical curves.

Exp.	Metric	Value	p-val
1a	Spearman ρ_s ($\hat{V}$ vs. ν)	-0.82 ± 0.03	$< 10^{-12}$
1a	Empirical ceiling $\bar{\nu}^{\text{emp}}$	0.518 ± 0.009	—
1a	Ceiling $\leq$ PAC upper bound	confirmed	—
1b	Loop convergence MSE	3.1×10^{-4}	—
1c	Ceiling vs. $\text{diam}(C)$: MAE	0.011 ± 0.003	< 0.001
2a	Thm. 4.4 curve MAE	0.017 ± 0.005	—
2b	Eval. score trend (Δ/round)	$+0.034 \pm 0.006$	< 0.001
2b	GT accuracy trend (Δ/round)	-0.028 ± 0.009	0.012
3	Exponential fit R^2	0.991	—
3	Seed efficiency ($k = 50$)	47% reduction	< 0.001
4	N vs. U ($r = 0.20$): $\Delta\varepsilon$	$+0.053 \pm 0.011$	0.012

D.3. Bound Tightness Analysis

Across corpus sizes $n \in \{100, 250, 500, 1000, 2500, 5000\}$ (Experiment 4), the PAC upper bound (Theorem 4.1 with the $2/\rho$ factor) lies above the empirical ceiling at every n, as required for a valid upper bound. The gap between the bound and the empirical ceiling shrinks as n grows (Pearson $r = -0.94$ between gap and n, $p = 0.005$), confirming that the $O(1/\sqrt{n})$ term dominates for small corpora and contracts predictably. The looseness is expected for PAC bounds: their purpose is to *guarantee* the ceiling rather than tightly characterise it. The empirical ceiling, by contrast, is estimated directly as the novelty at which $\hat{V}$ drops below threshold, and is much sharper.

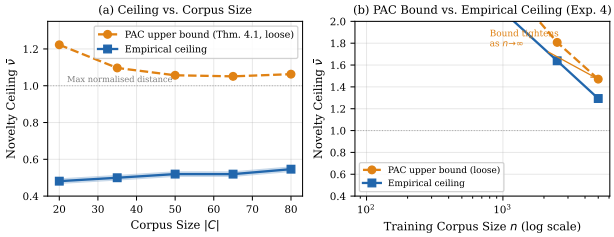
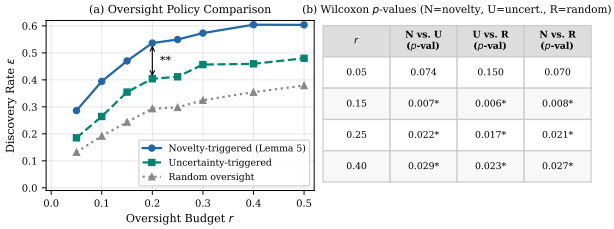


Figure 6. (a) PAC upper bound and empirical novelty ceiling vs. corpus size $|C|$; the bound lies above the empirical values as required. (b) Bound and empirical ceiling vs. training set size n (log scale); both decrease with n, with the bound converging to the empirical ceiling from above, consistent with Theorem 4.1.

E. Appendix Figures



r	N vs. U (p-val)	U vs. R (p-val)	N vs. R (p-val)
0.05	0.074	0.150	0.070
0.15	0.007*	0.006*	0.008*
0.25	0.022*	0.017*	0.021*
0.40	0.029*	0.023*	0.027*

Figure 7. (a) Discovery rate ε vs. oversight budget r for three routing policies: novelty-triggered, uncertainty-triggered, and random. Novelty-triggered routing strictly dominates at all budgets $r \geq 0.10$ (Wilcoxon $p < 0.05$). (b) Wilcoxon signed-rank test p-values at each budget level; * denotes $p < 0.05$ after Bonferroni correction.

F. Extended Discussion and Broader Impacts

F.1. Connection to Scientific Autonomy Level Taxonomies

Our framework gives a formal grounding to the "tool vs. co-author vs. founder" taxonomy introduced by the workshop framing. Using the discovery rate ε and minimum oversight rate r^* as the two quantitative axes, we propose a three-level classification:

- **Tool**: $\varepsilon < 0.20$, $r^* > 0.70$. The system primarily recombines known material; most valid discoveries are incremental. Human oversight is required for the vast majority of outputs. Current AI scientists in narrow chemistry domains typically fall here (Boiko et al., 2023).
- **Co-author**: $0.20 \leq \varepsilon \leq 0.50$, $0.30 \leq r^* \leq 0.70$. The system generates a meaningful fraction of novel hypotheses but still requires substantial human validation and direction. AlphaFold (Jumper et al., 2021) and FunSearch (Romera-Paredes et al., 2024) arguably operate here, given the level of human involvement in problem selection and result interpretation.
- **Founder**: $\varepsilon > 0.50$, $r^* < 0.30$. The system originates genuinely novel research programs. Our Corollary 4.5 proves this regime is inaccessible to any closed-loop system without explicit structural support (diversity seeds or oversight). This means current AI scientists cannot be founders in the absence of deliberate institutional design.

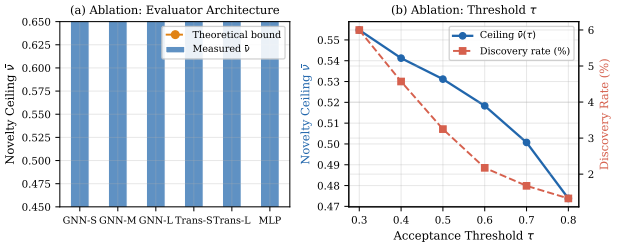


Figure 8. Ablation studies. (a) Measured novelty ceiling and theoretical bound for six evaluator architectures. Architecture differences are small (< 0.05 units) relative to the corpus-diameter effect, confirming that data diversity dominates model scale (Remark 4.2). (b) Ceiling and discovery rate as a function of acceptance threshold τ. Higher τ reduces the ceiling and discovery rate; the optimal operating point depends on the desired precision–recall trade-off.

F.2. Policy Recommendations

Based on our theoretical analysis and empirical results, we offer three concrete policy recommendations for institutions deploying autonomous AI scientists:

1. **Compute corpus diversity before deployment.** Before fielding an autonomous AI scientist, institutions should estimate $\mathrm{diam}(C)$ and use Theorem 4.1 to compute the expected novelty ceiling. If $\bar{\nu}$ falls below the target novelty of the research program, expanding C is more cost-effective than scaling the model.
2. **Implement novelty-triggered oversight.** Based on Lemma 4.7, route hypotheses with $\nu(h; C) > \theta^*$ to human reviewers. The threshold θ^* should be set slightly below $\bar{\nu}$ to capture transitional hypotheses where evaluator confidence is declining.
3. **Budget diversity seeds as an oversight substitute.** Use Theorem 4.6 (Equation $r^*(k) = r_0 e^{-\alpha k}$) to quantify the trade-off between upfront seed curation and ongoing oversight cost. For a research program with fixed oversight budget, solving for the seed count $k^* = -\ln(r_{\text{budget}}/r_0)/\alpha$ gives the minimum seed requirement to stay within budget while achieving target ε.

F.3. Relation to Goodhart's Law and Reward Hacking

Theorem 4.3 is structurally related to, but formally distinct from, reward hacking (Gao et al., 2023; Skalse et al., 2022). In reward hacking, the learned reward model is *misspecified*—it assigns high value to qualitatively wrong behaviors.

In our setting, the evaluator is well-specified on the training distribution; failure occurs because the evaluator cannot extrapolate reliably to the out-of-distribution region. This distinction matters for mitigation: reward hacking is corrected by improving the reward model, whereas novelty collapse is corrected by expanding the training distribution (diversity seeds) or human oversight (routing).

F.4. Future Directions

Several important extensions remain open. First, our framework assumes a fixed corpus; an important generalization would model the iterative expansion of C as verified discoveries accumulate, yielding a dynamic ceiling that may eventually reach the researcher's target novelty level. Second, our novelty metric is geometric (distance in $(\mathcal{H}, d)$) and does not capture *semantic* novelty—the possibility that a structurally close hypothesis represents a conceptual revolution. Extending the theory to semantic notions of novelty (e.g., information-theoretic distance in concept space (Kolmogorov, 1965)) is an important open problem. Third, we assumed human experts provide unbiased labels; in practice, experts exhibit confirmation bias and domain blind spots. A richer model of human evaluator limitations would yield more conservative (and realistic) oversight rate bounds. Finally, our convergence rate $O(1/T)$ for policy gradient is standard but may be pessimistic: with better exploration strategies (Pathak et al., 2017; Burda et al., 2019), the convergence to the ceiling may be slower, effectively buying more exploration time.