ONLINE SURROGATE REPAIR: DECOUPLING HIGH-FIDELITY FEEDBACK FROM SEARCH LENGTH IN CLOSED-LOOP DISCOVERY

Anonymous authors
Paper under double-blind review

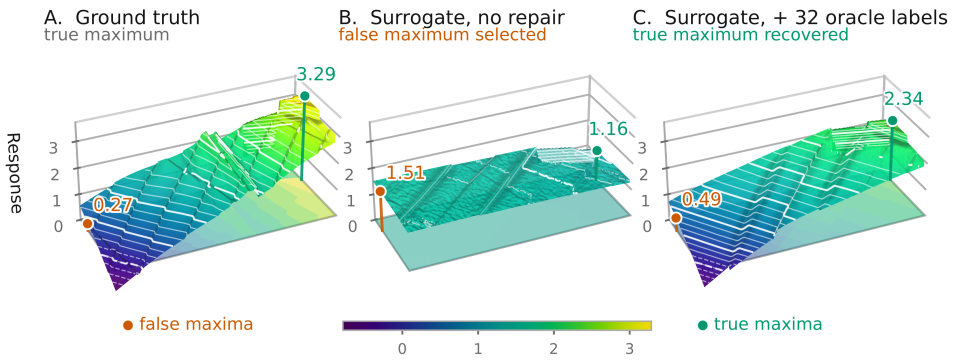


Figure 1: OSR corrects optimizer-relevant error. (A) Ground truth; (B) an unrepaired surrogate selects a false maximum; (C) 32 EI-selected oracle labels restore the true maximum.

ABSTRACT

Closed-loop AI scientists can generate candidate designs at low marginal computational cost, whereas reliable feedback may require wet-lab synthesis, characterization, or high-fidelity computation. Addressing this imbalance through custom laboratory automation remains infrastructure-intensive and costly, while replacing new experiments with a fixed surrogate leaves persistent model errors that can be amplified by optimization. We propose *online surrogate repair* (OSR), a closed-loop algorithm that uses sparse high-fidelity evaluations to update the surrogate throughout a longer agent search conducted primarily with inexpensive surrogate feedback. An acquisition rule selects which designs from the agent's accumulated proposals receive high-fidelity evaluation, and the resulting labels update the surrogate used in subsequent episodes. Across controlled synthetic environments, we demonstrate that improving global surrogate fit does not necessarily reduce maximum regret, whereas Q90-UCB and expected improvement (EI) substantially reduce regret by directing evaluations toward regions that determine the optimizer's decisions. On MADE, controls receiving high-fidelity feedback after every episode require 6.36–$7.23\times$ more oracle queries to match Online EI under two LLM orchestrators and $10.27\times$ more under the non-LLM Chemeleon+MLIP workflow. Online surrogate repair introduces a novel third feedback regime between fixed-surrogate operation and high-fidelity feedback after every episode, separating the frequency of high-fidelity evaluation from the duration of the agent's search.

1 INTRODUCTION

AI scientists built on large language models increasingly automate hypothesis generation, tool use, computational experimentation, and iterative planning (Lu et al., 2024; Boiko et al., 2023). Related systems now span materials design, quantum experiments, and automated theory formation (Jia et al., 2024; Cao et al., 2025; Jagadish et al., 2026). Their reasoning and proposal generation can scale

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

at comparatively low computational cost, whereas reliable scientific feedback may require wet-lab synthesis, characterization, or high-fidelity computation. Discovery can therefore be limited by validated outcomes rather than by the supply of candidate designs.

This imbalance is especially visible in computational materials discovery. Crystal generators can propose structures at scale (Xie et al., 2022; Zeni et al., 2025; Park et al., 2025), and learned pipelines have expanded the number of candidate stable materials (Merchant et al., 2023). Those candidates must still be ranked or validated by interatomic potentials, electronic-structure calculations, or experiments. Static benchmarks evaluate property prediction, stability ranking, or one-shot generation (Dunn et al., 2020; Riebesell et al., 2025; Betala et al., 2026), but do not reveal how approximation errors alter a policy that repeatedly acts on model feedback. MADE instead evaluates complete discovery policies under a sequential oracle budget (Malik et al., 2026).

Self-driving laboratories address the feedback bottleneck by automating synthesis, handling, and characterization (Burger et al., 2020; Kusne et al., 2020; Szymanski et al., 2023; Rapp et al., 2024; Sheng et al., 2024; Seifrid et al., 2022). They reduce human intervention and experimental latency, but require specialized infrastructure, substantial integration effort, and continued expenditure on each physical evaluation (Abolhasani & Kumacheva, 2023; Pilon et al., 2026; Zhang et al., 2025). At the other limit, a fixed surrogate permits a long search without further high-fidelity evaluation, but its errors persist and are sampled non-uniformly by optimization. Maximization preferentially selects overestimated regions, so a localized false maximum can determine the search trajectory even when global predictive error is low (Trabucco et al., 2021; Riebesell et al., 2025; Malik et al., 2026).

We introduce *online surrogate repair* (OSR), a third feedback regime in which most interactions use inexpensive surrogate feedback while sparse high-fidelity evaluations update it during the optimization process. The AI scientist retains control of proposal generation and design selection, while a separate acquisition rule chooses archived proposals for oracle evaluation and appends the returned labels to the surrogate context used in subsequent episodes (Hollmann et al., 2025; Grinsztajn et al., 2026). Because the purpose of the surrogate is to guide search, repair is evaluated by downstream optimization rather than by global fit alone.

Our contributions are:

- We formulate online surrogate repair, in which a scientific agent operates under inexpensive surrogate feedback while a separate acquisition rule allocates a limited number of oracle queries to update the surrogate during the search.
- We show that global prediction error and maximum regret are distinct repair objectives, and identify acquisition rules that substantially improve the surrogate's optimization performance.
- We demonstrate that online surrogate repair reduces the oracle feedback required for closed-loop optimization in controlled synthetic environments and for open-ended materials discovery on MADE under both LLM and non-LLM proposal policies.

These results show that a limited oracle budget can improve the surrogate throughout a closed-loop search, rather than requiring an oracle evaluation after every agent episode.

2 Related Work

2.1 Scientific agents and closed-loop discovery

Scientific agents couple language-model planning with external tools and feedback (Lu et al., 2024; Boiko et al., 2023). In materials science, LLMatDesign uses tool-mediated self-reflection (Jia et al., 2024); MOFGPT and MOFGen combine language models with property predictors, generators, and simulators (Badrinarayanan et al., 2025; Inizan et al., 2025); and LLEMA uses LLM-guided evolutionary search (Abhyankar et al., 2026). Existing benchmarks test domain reasoning and scenario- or project-level discovery (Zhang et al., 2026; Song et al., 2026), while MADE focuses on budgeted, closed-loop materials discovery (Malik et al., 2026). OSR is proposal-policy agnostic: it changes how costly evidence updates feedback and can accompany either an LLM agent or a fixed generator.

2.2 Active learning, Bayesian optimization, and multi-fidelity search

Active learning and Bayesian optimization allocate expensive evaluations with surrogates and acquisition functions (Lookman et al., 2019; Jones et al., 1998; Garnett, 2023). Their use in materials has motivated benchmarks of sequential and active-learning policies (Rohr et al., 2020; Wang et al., 2022) and closed-loop laboratory demonstrations (Kusne et al., 2020). In standard Bayesian optimization, acquisition is itself the proposal policy. Multi-fidelity optimization also chooses among information sources (Poloczek et al., 2017), while Gemini learns a cross-fidelity correction that drives a Bayesian optimizer (Hickman et al., 2021). OSR leaves search to the scientific agent: acquisition only chooses which archived proposal receives a high-fidelity label to repair future feedback. Agent episodes and oracle evaluations can therefore vary independently.

2.3 Surrogate verifiers and tabular foundation models

Predictive accuracy can be misaligned with discovery performance. Matbench tests property prediction on fixed datasets (Dunn et al., 2020), while Matbench Discovery ranks a fixed stability pool and identifies consequential false positives near the stability boundary (Riebesell et al., 2025). Learned interatomic potentials make large-scale screening practical (Batatia et al., 2025), but their errors become consequential under distribution shift. Ensemble and Bayesian interatomic potentials seek calibrated epistemic uncertainty (Busk et al., 2023; Coscia et al., 2026), while probabilistic stability models incorporate convex-hull uncertainty into active learning (Novick et al., 2024). These methods improve or quantify the model, while OSR asks how sparse labels should repair a surrogate already embedded in a longer discovery loop.

Tabular foundation models permit new measurements to be incorporated as additional context rows without task-specific training (Hollmann et al., 2025; Grinsztajn et al., 2026; QU et al., 2026). This makes repair immediate, rather than requiring a retraining job after every oracle query. Active in-context learning selects labels from a fixed pool to improve global prediction (Treerath & Pittorino, 2026). We acquire labels from an evolving archive of agent proposals and measure their effect on subsequent closed-loop optimization.

3 Methods

Closed-loop AI scientists generate candidates far more cheaply than they obtain reliable feedback. Fixed surrogates enable long searches but preserve exploitable errors while high-fidelity evaluation after every episode is reliable but costly. OSR separates search from verification: the agent searches with inexpensive surrogate feedback, while a sparse outer loop evaluates decision-relevant archived proposals and uses the labels to repair future feedback. We first identify acquisition rules for optimization-critical errors, then place the strongest in the full agent loop and evaluate OSR on synthetic and real-world benchmarks.

3.1 Problem Definition

Let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ be the initial context of expensive measurements, where x is a mixed-type tabular design and $y = f(x)$ is the oracle response for a hidden world f. An oracle query invokes a physical measurement or high-fidelity computation in deployment, and an exact or higher-fidelity simulated evaluator in our benchmarks. Q is the additional query budget beyond $\mathcal{D}_N$.

A tabular foundation model conditioned on a context $\mathcal{D}$ returns a predictive distribution $p_{\mathcal{D}}(y \mid x)$ in a single forward pass, with mean $\mu_{\mathcal{D}}(x)$ (Hollmann et al., 2025; Grinsztajn et al., 2026). We write $q_{\alpha,\mathcal{D}}(x)$ for the value below which this distribution places probability α, so $q_{0.90,\mathcal{D}}(x)$ is its optimistic upper end. Appending the returned measurement as a labelled row to the context is the entire model update, and no parameters are trained. Despite low global error, a limited context can retain false maxima that an optimizer will exploit. We define *surrogate repair* as selecting designs for the Q additional oracle queries and incorporating their returned labels into $\mathcal{D}$ to improve subsequent surrogate predictions.

At each repair step t, an acquisition rule selects a design x_{t+1} using only the current context and the surrogate's own predictions; the oracle returns $y_{t+1} = f(x_{t+1})$, and the context becomes

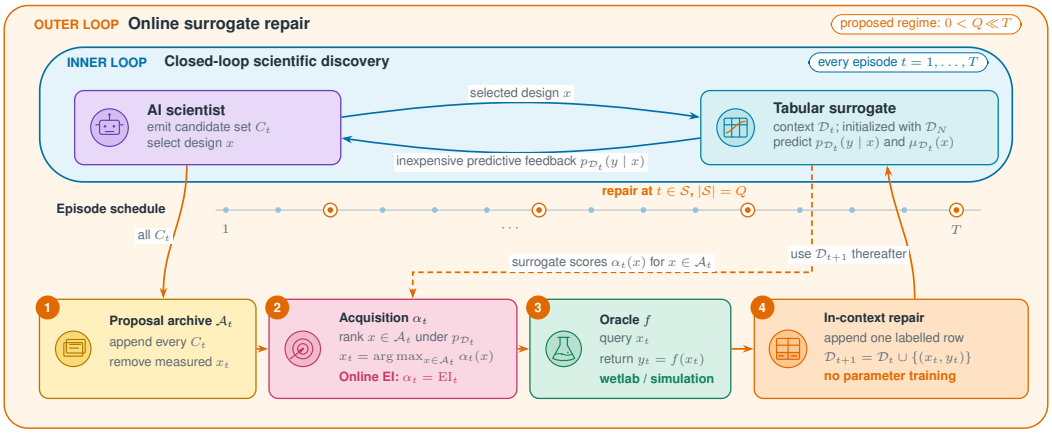


Figure 2: Online surrogate repair separates a frequent discovery loop from sparse surrogate updates. Candidate sets accumulate in $\mathcal{A}_t$; at $t \in \mathcal{S}$, acquisition selects x_t for oracle evaluation, and the returned label is appended to $\mathcal{D}_t$ for subsequent episodes. Online EI uses $\alpha_t = \mathrm{EI}_t$.

$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_{t+1}, y_{t+1})\}$. We measure the repaired surrogate on two axes: global fit, as the normalized error E (NRMSE) against a hidden audit sample of the world, and optimizer-facing quality, as the normalized maximum regret

$$R_{\max}(\mathcal{D}) = \frac{f(x^\star) - f(\hat{x}_{\mathcal{D}})}{R_w}, \qquad \hat{x}_{\mathcal{D}} = \arg\max_x \mu_{\mathcal{D}}(x), \tag{1}$$

where $x^\star$ is the best design found by a high-budget reference search on the world itself, $\hat{x}_{\mathcal{D}}$ is the best design found by the same search procedure on the surrogate, and R_w is a robust response range that makes regret comparable across worlds. The improvements ΔR and ΔE report the initial value minus the final value, so positive is better. The acquisition rules below differ only in how they choose x_{t+1}, and all other components are held fixed in our experiments.

3.2 Online Surrogate Repair for Closed-loop AI Scientists

Across T episodes, full feedback has $Q = T$, a fixed surrogate $Q = 0$, and OSR $0 < Q \ll T$, with surrogate feedback each episode and Q repairs scheduled online.

Figure 2 illustrates the dual-loop structure, and Algorithm 1 gives the complete procedure. Every emitted candidate is recorded in an append-only archive $\mathcal{A}_t$, and a design leaves the archive only by being measured. The oracle is queried on a fixed schedule $\mathcal{S} \subset \{1, \ldots, T\}$ of Q evenly spaced episodes. At $t \in \mathcal{S}$ an acquisition rule α from Section 3.3 selects

$$x_t = \arg\max_{x \in \mathcal{A}_t} \alpha_t(x), \tag{2}$$

the oracle returns $y_t = f(x_t)$, and the labelled pair is appended to the surrogate's context, so every subsequent episode is scored by the repaired surrogate. The measured design need not be the episode's own proposal. Instead, the label goes to whichever archived design the acquisition method rates highest under the current surrogate. The framework accepts any repair rule as α. Our experiments identify expected improvement as the strongest choice, and we call this instantiation Online EI.

Algorithm 1 Online surrogate repair under sparse experimental feedback

Require: initial labelled context $\mathcal{D}_N$, agent, oracle f, episodes T, schedule $\mathcal{S}$, $|\mathcal{S}| = Q$

$\mathcal{D} \leftarrow \mathcal{D}_N$ ▷ the surrogate starts from the initial context
$\mathcal{A} \leftarrow \emptyset$ ▷ the archive fills with the agent's emissions
for $t = 1$ **to** T **do**
  agent emits candidates C_t and selects x using $p_{\mathcal{D}}$
  $\mathcal{A} \leftarrow \mathcal{A} \cup C_t$
  return $p_{\mathcal{D}}(y \mid x)$ to the agent ▷ surrogate feedback, every episode
  if $t \in \mathcal{S}$ **then**
    $x_t \leftarrow \arg\max_{x \in \mathcal{A}} \alpha_t(x)$ ▷ Eq. (2)
    $y_t \leftarrow f(x_t)$; $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x_t, y_t)\}$ ▷ one query repairs future feedback
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{x_t\}$
  end if
end for

A static repair variant spends the same Q queries before the campaign and then freezes the surrogate, isolating the value of allocating measurements online.

3.3 Acquisition Rules

Below we introduce the acquisition rules used in our surrogate repair experiments.

3.3.1 Target agnostic methods

These acquisition rules select labels without reference to the surrogate's estimate of the optimum, they expand the context by improving coverage and global error.

Random. Distribution-matched random sampling draws a design from the world's own design distribution, labels it, and updates the surrogate's context, $x_{t+1} \sim p(x)$. The surrogate plays no role in selection, making this the control for every other method at the same query budget.

Global-UQ. Classical active learning allocates labels to reduce predictive uncertainty globally (MacKay, 1992; Cohn et al., 1996). A bounded candidate set $\mathcal{C}_t$ is constructed from the geometry of the current context (midpoints of the longest edges of a nearest-neighbour graph over the labelled designs), and the rule selects

$$x_{t+1} \;=\; \arg\max_{x \in \mathcal{C}_t} \big[\, q_{0.90,t}(x) - q_{0.10,t}(x) \,\big], \tag{3}$$

the design with the largest 80% predictive interval. This is the natural policy when the target is global fit.

3.3.2 Optimum seeking methods

These rules aim the query budget directly at the optimum. They share one procedure: run a fixed surrogate-only search over the design space and query the design that maximizes a pointwise surrogate score. They differ in how that score reads the predictive distribution. PEAK validates the design the surrogate itself would select, while Q90-UCB validates the design the surrogate considers most plausibly excellent.

Peak. This rule naively validates the predicted optimum. It runs the surrogate-only search to approximately maximize the predictive mean $\mu_t(x)$, labels the incumbent $\hat{x}_{\mathcal{D}_t}$, and re-optimizes after every update. The rule aims directly at the maximum but uses only the surrogate's point estimate. It can therefore repair false maxima only one at a time by spending an oracle query at each location.

Q90-UCB. The same procedure read at the model's optimistic upper quantile instead of its point prediction:

$$x_{t+1} \;=\; \arg\max_{x} \; q_{0.90,t}(x). \tag{4}$$

Table 1: Static repair after Q=32 oracle queries. The four central columns report reductions in maximum regret ΔR for each surrogate and context size, the final columns give averages of the reduction in maximum regret and the NRMSE ΔE over the four settings. Higher is better. Full $R_{\max}$ and NRMSE results are given in Appendix D. **Bold** best, underlined second best per column.

Method	TabPFN-100	TabPFN-1000	TabICL-100	TabICL-1000	Mean ΔR	Mean ΔE
Q90-UCB	**0.4978**	0.2393	**0.5909**	**0.3664**	**0.4236**	−0.0002
EI	0.4516	**0.2862**	0.5686	0.3309	0.4093	0.0011
Peak	0.4057	0.2615	0.5593	0.3216	0.3870	−0.0017
G2P(16)	0.3094	0.1672	0.4891	0.2617	0.3069	0.0013
Global-UQ	0.0694	−0.0018	0.1256	0.0593	0.0631	**0.0046**
Random	0.0565	0.0188	0.1456	−0.0129	0.0520	0.0045

This is an upper-confidence rule expressed directly in the model's own quantiles (Srinivas et al., 2010). It queries designs whose plausible upside is largest. A large upper quantile highlights two kinds of design that oracle queries can distinguish: designs the surrogate overestimates and designs that are genuine optima. The label is useful in both cases, since it collapses the quantile of an overestimated design and confirms a genuine one, and the search then moves to the next candidate.

G2P(k). A hybrid of GLOBAL-UQ and PEAK: spend the first k queries with GLOBAL-UQ and the remaining $Q - k$ with PEAK. We set $k = Q/2$, yielding G2P(16) at $Q = 32$, so half of the budget improves global coverage before peak validation begins.

3.3.3 IMPROVEMENT AWARE ACQUISITION

EI. Expected improvement (Mockus et al., 1978; Jones et al., 1998) references the incumbent measurement rather than the surrogate's landscape alone. With incumbent $y_t^{\max} = \max_{(x,y)\in\mathcal{D}_t} y$, the score of a design is its expected margin over the best label seen so far,

$$\mathrm{EI}_t(x) \;=\; \mathbb{E}_{p_t(y|x)}[\,(y - y_t^{\max})_+\,], \tag{5}$$

computed as the mean of $\max(q - y_t^{\max}, 0)$ over the model's full predicted distribution, so the tail that drives improvement is integrated rather than truncated. Selection follows the same surrogate-only search as the optimum seeking methods: at each step, $x_{t+1} = \arg\max_x \mathrm{EI}_t(x)$, the oracle label is appended to the context, and the incumbent is updated before the next step.

Unlike Q90-UCB, EI integrates the full upper tail and references the verified incumbent, so it stops prioritizing regions that cannot improve the best observed response.

4 EXPERIMENTS

Our experiments proceed in three stages. The static study isolates the acquisition hypothesis of Section 3.3 on a large matrix of worlds. The synthetic world study places the repair methods inside the full loop of Section 3.2 and evaluates a single-maximum regression task. The MADE experiment runs the same loop on a multi-minima materials discovery target.

4.1 STATIC SURROGATE REPAIR

Each acquisition rule repairs the surrogate with $Q = 32$ oracle queries on 40 synthetically generated tabular worlds, with two independent context draws per world, at both context sizes $N = 100$ and $N = 1000$, and under two surrogates, TabPFN-3 (Grinsztajn et al., 2026) and TabICLv2 (QU et al., 2026). Different surrogate campaigns share their worlds, contexts and audit samples exactly, so the models are directly comparable. Table 1 reports the final value and the improvement of both measures of Section 3.1.

Across the four settings, Q90-UCB and EI achieve mean maximum-regret reductions of 0.424 and 0.409, compared with 0.063 and 0.052 for Global-UQ and Random. Conversely, Global-UQ and Random produce the largest mean NRMSE improvements, while the optimizer-targeting methods

Table 2: Synthetic world results over 20 paired worlds under noise-free and noisy feedback. Values are final normalized regret with 95% CIs (lower is better). Noise affects both the initial context and subsequent oracle observations, while regret is evaluated using the selected design's noise-free response. **Bold** best, underlined second best within each block. The Online EI query-budget sweep is given in Appendix C.1.

Method	Oracle queries	Mean regret (95% CI)	Median
Noise-free feedback			
Online EI Control (Q=T=200)	200	**0.0278** (0.0100, 0.0499)	**0.0014**
Oracle Control (Q=T=200)	200	0.0755 (0.0468, 0.1069)	0.0592
Online EI (Q=32), TabPFN-3	32	0.0552 (0.0298, 0.0843)	0.0393
Online Q90-UCB (Q=32), TabPFN-3	32	0.0600 (0.0338, 0.0887)	0.0455
Online Q90-UCB (Q=32), TabICLv2	32	0.0681 (0.0406, 0.0967)	0.0412
Archive-BO (Q=32), GP + EI	32	0.0807 (0.0490, 0.1144)	0.0592
Online Peak (Q=32), TabPFN-3	32	0.0976 (0.0702, 0.1259)	0.0928
Static EI (Q=32)	32	0.0984 (0.0686, 0.1291)	0.0983
Tab-AICL (Hybrid Method) (Q=32)	32	0.1169 (0.0894, 0.1466)	0.0968
Static Global-UQ (Q=32)	32	0.1220 (0.0921, 0.1520)	0.1105
No repair	0	0.1432 (0.1163, 0.1725)	0.1231
Noisy feedback (σ=0.05R_w)			
Online EI Control (Q=T=200)	200	**0.0639** (0.0384, 0.0914)	**0.0488**
Oracle Control (Q=T=200)	200	0.1128 (0.0780, 0.1505)	0.0898
Online EI (Q=32), TabPFN-3	32	0.0905 (0.0587, 0.1252)	0.0732
Online Q90-UCB (Q=32), TabPFN-3	32	0.0981 (0.0672, 0.1292)	0.0892
Static EI (Q=32)	32	0.1124 (0.0811, 0.1442)	0.1069
No repair	0	0.1546 (0.1264, 0.1830)	0.1447

yield little global error improvement. Improving average prediction therefore does not remove the false maxima exploited during optimization. This is a crucial result because TabArena evaluates global predictive accuracy and TabPFN is trained for global prediction, not maximum regret, creating an objective mismatch with closed-loop discovery (Erickson et al., 2025; Hollmann et al., 2025; Grinsztajn et al., 2026). Figure 1 visualizes this failure and its correction in a representative world. We carry Q90-UCB and EI into the online study and retain Peak and Global-UQ as ablations.

4.2 Synthetic worlds

Our main comparison uses an exact oracle $f(x)$, providing a controlled analogue of experimental ground truth. Each of the 20 held-out worlds is a frozen, seeded structural causal model (SCM) (Pearl, 2009) inspired by TabPFN's sampled causal-graph prior (Hollmann et al., 2025), with a deterministic response and a single global maximum. These tasks complement MADE's multi-minima discovery setting. Appendix B gives the construction and reference protocol. A ReAct agent (Yao et al., 2023) using DeepSeek-V4-Flash proposes designs for $T = 200$ episodes, and a TabPFN surrogate scores the selected design as in Algorithm 1.

At the same episode budget, Oracle Control returns oracle feedback on the agent-selected design each episode ($Q = T = 200$). Online EI Control uses surrogate feedback and EI-selected archive repairs every episode ($Q = T = 200$). At $Q = 32$, online repair acquires labels at evenly spaced episodes, while static repair acquires them before search and then freezes the surrogate. No repair ($Q = 0$) provides the reference. Noise-free runs report normalized regret of the best oracle-verified design.

At matched $Q = 32$, Online EI reduces mean regret by 44% versus Static EI and closes 76% of the gap from no repair to Online EI Control ($Q = 200$), using 16% of its queries (Table 2). All methods use their current surrogate within episodes, Oracle Control also appends labels to its context after each episode. Both EI settings yield lower mean regret, consistent with more useful archive repairs.

We adapt two baselines to test whether the gains depend on the surrogate or acquisition objective. Archive-BO retains the archive, schedule, and EI acquisition but replaces TabPFN with a Gaussian process. Its mean regret of 0.0807 lies between Static EI and Online EI, showing that online repair's

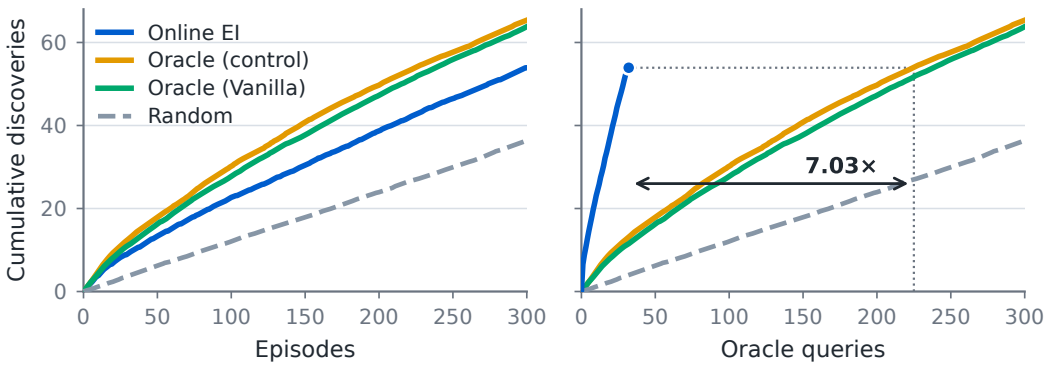


Figure 3: MADE convergence over the DeepSeek runs. Online EI ends after 32 oracle queries. On the aggregate query curve, the full-feedback control required $7.03\times$ as many oracle queries to reach the same discovery count.

benefit is not specific to in-context updating. Tab-AICL adapts active in-context learning (Treerath & Pittorino, 2026) to sequential regression on the agent archive (Appendix F). It reaches 0.1169, last among the online methods and only marginally ahead of Static Global-UQ. As in Section 4.1, selection for global coverage is less effective than targeting the regions that determine the optimizer's decisions.

The online setting also reverses Q90-UCB's slight aggregate lead over EI in the static study. Appendix C.2 relates acquisition geometry to final regret, where mean distances between successive acquisitions are 0.08, 0.36, and 0.92 times the initial-context distance scale for Peak, EI, and Tab-AICL, respectively. EI lies between these extremes and achieves the lowest regret, so broader dispersion alone does not explain performance. This is consistent with the acquisition objectives: EI combines uncertainty with improvement over a verified incumbent, while Peak uses point predictions and Tab-AICL targets coverage. The budget sweep in Appendix C.1 shows diminishing returns when $Q = 32$ is doubled to 64.

We add independent Gaussian noise ($\sigma = 0.05R_w$) to all oracle observations on the same 20 worlds, including the shared 1,000-row initial context. Each run selects the measured design with the highest noisy label and regret uses its noise-free response. Online EI at $Q = 32$ retains 19% lower mean regret than Static EI and again has lower mean regret than Oracle Control. Online EI Control retains the lowest mean regret (Table 2).

4.3 MADE BENCHMARK

We further evaluate OSR on MADE (Malik et al., 2026), a closed-loop materials discovery benchmark. Each environment specifies a chemical system and its known phases, and the agent proposes crystal structures over an iterative campaign. The weaker MACE model (Batatia et al., 2025) provides surrogate feedback, while the stronger Orb-v3 model (Rhodes et al., 2025) serves as the oracle. A novel structure within $\tau = 0.1$ eV/atom of the convex hull counts as a discovery. To measure multi-minima discovery, we audit all episode proposals from every method on MADE with the oracle post hoc, without returning audit labels to the search or including audit calls in the reported feedback budget Q, this is distinct from the synthetic worlds experiments above, where we score the incumbent selected from the initial context and budgeted oracle measurements ($Q = 32$ for Online EI and $Q = 200$ for both synthetic controls, Table 2).

We evaluate 30 chemical systems with three independent $T = 300$ runs per system under two LLM orchestrators, DeepSeek-V4-Flash and GPT-5.6-Luna. Online EI receives $Q = 32$ oracle queries on an evenly spaced schedule. Two full feedback baselines receive an oracle query after every episode ($Q = 300$). Vanilla follows the original MADE agent with MACE candidate scoring, whereas Control adds the TabPFN surrogate to the same stack.

Table 3 shows that Control requires 7.23(28) and 6.36(20) times as many oracle queries to match Online EI under DeepSeek and Luna, respectively, equivalent to 86% and 84% fewer queries for

Table 3: MADE benchmark with 3 repeats per method. Online EI uses Q=32 oracle queries; Vanilla and the TabPFN+MACE Control use Q=300. Control receives the same per-oracle surrogate repair as Online EI. Unmatched counts runs (out of 90) in which Control does not reach a method's 32-query discovery count within 300 queries. **Bold** best, underlined second best per column within each block. AF averages per-run ratios, whereas the 7.03× in Figure 3 is read from the run-averaged discovery curve, the two differ because ratios and means do not commute. Full table is given in Appendix G
*AUDC can exceed one here because our setting can yield multiple discoveries per oracle query.

Method	EF	AF	Disc.		Unmatched	AUDC		U. comps.		U. SGs	
	@32	@k	@32	@300	@full budget	@32	@300	@32	@300	@32	@300
DeepSeek-V4-Flash											
Online EI	**4.13(21)**	**7.23(28)**	**53.9(28)**	—	**21/90**	**1.96(11)***	—	**18.9(12)**	—	**22.03(77)**	—
Vanilla	0.882(51)	1.077(74)	11.52(67)	63.8(31)	0/90	0.401(21)	0.243(12)	9.50(58)	21.2(14)	8.16(32)	**23.08(59)**
Control	1.00 (ref)	1.00 (ref)	13.07(72)	**65.4(34)**	—	0.452(23)	**0.256(13)**	10.88(70)	20.9(15)	8.57(38)	23.03(58)
Random	0.303(55)	0.363(34)	3.96(71)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	**35.3(62)**	1.36(14)	3.61(31)
GPT-5.6-Luna											
Online EI	**5.91(29)**	**6.36(20)**	**56.9(29)**	—	**7/90**	**2.07(11)***	—	**19.0(11)**	—	**23.57(58)**	—
Vanilla	1.035(72)	1.280(77)	9.97(69)	**86.3(49)**	0/90	0.291(22)	**0.307(17)**	5.52(50)	20.3(14)	7.19(46)	27.16(94)
Control	1.00 (ref)	1.00 (ref)	9.63(67)	84.6(47)	—	0.286(22)	0.296(15)	5.24(41)	20.8(14)	7.23(47)	**27.30(84)**
Random	0.411(74)	0.879(89)	3.96(70)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	**35.3(61)**	1.36(14)	3.61(31)
Chemeleon+MLIP											
Online EI	**4.55(48)**	**10.27(92)**	**130(14)**	—	**31/90**	**5.26(43)***	—	**119(12)**	—	**23.1(18)**	—
Vanilla	0.914(39)	1.02(12)	26.2(11)	134(15)	1/90	0.828(36)	0.561(48)	24.9(10)	124(13)	11.66(64)	23.5(17)
Control	1.00 (ref)	1.00 (ref)	28.68(53)	**141(16)**	—	0.900(15)	**0.595(50)**	27.32(48)	**128(14)**	13.46(58)	**24.3(19)**
Random	0.138(25)	0.209(23)	3.96(71)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	35.3(62)	1.36(14)	3.61(31)

Online EI. Control fails to catch up within 300 queries in 21 and 7 of the 90 runs, so these acceleration factors are lower bounds. Figure 3 shows the discovery curves.

With Q=32, Online EI reaches 82% and 67% of the corresponding Controls' Q=300 discovery counts and leads both full-feedback baselines in unique compositions and space groups at matched budget. With the non-LLM Chemeleon+MLIP workflow (Park et al., 2025; Malik et al., 2026), it reaches 130 discoveries, 93% of Control's 141, with an AF of 10.27.

5 Discussions and Conclusion

We demonstrated that global prediction error and maximum regret are misaligned surrogate objectives. In light of this we introduced *online surrogate repair*, a novel algorithm that updates the surrogate during a longer agent search with sparse oracle queries. Controlled synthetic experiments show that global predictive accuracy can improve without reducing optimizer-facing regret, while Online EI requires substantially fewer oracle queries than full feedback on both synthetic worlds and MADE.

These findings suggest that feedback models should be evaluated through the searches they guide, alongside their predictive accuracy. The Chemeleon results also show that proposal quality and surrogate repair can contribute together: a specialized generator produces more discoveries than either LLM under both feedback regimes while benefiting from the same repair procedure. At matched discovery counts, Online EI incurs modest additional LLM cost (Appendix E), making this approach attractive when high-fidelity evaluation is expensive relative to continued search.

The current framework requires a tabular representation of the discovery problem and a tabular foundation model that can incorporate new labels in context. Most closed-loop discovery problems can be expressed in this form, as our MADE experiment illustrates, but some environments may fall out-of-distribution of the model's capabilities which we did not encounter or test in our experiments.

We have not tested OSR or Online EI with agents using test-time training or reinforcement learning, with other surrogate classes, or with more advanced and optimization-aware scheduling.

AI USE STATEMENT

In addition to the LLM-based experimental workflows described in this paper, we used generative AI tools to polish the manuscript and correct grammatical errors, assist with writing code, assist with creating an editable diagram in TikZ, and support literature search and identification of related work. The authors reviewed the AI-assisted text, tested the AI-assisted code, and checked suggested references against the original papers. We take responsibility for the final content of this work, including its text, figures, code, results, and scientific claims.

REPRODUCIBILITY STATEMENT

The online surrogate repair procedure is described in Algorithm 1, with the problem formulation and acquisition rules given in Sections 3.1 and 3.3. Section 4 describes the experimental settings, surrogate and oracle models, baselines, query budgets, and evaluation protocols. Appendix B details the synthetic-world construction, initial contexts, audit samples, and reference-search protocol. Appendix A defines the MADE evaluation metrics. Additional query-budget comparisons and details of the Tab-AICL baseline adaptation are provided in Appendices C.1 and F, respectively. Source code is provided in the supplementary material.

REFERENCES

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K. Reddy. LLEMA: Evolutionary search with LLMs for multi-objective materials discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TIqzhBvCNB.

Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2(6):483–492, Jun 2023. ISSN 2731-0582. doi: 10.1038/s44160-022-00231-0. URL https://doi.org/10.1038/s44160-022-00231-0.

Srivathsan Badrinarayanan, Rishikesh Magar, Akshay Antony, Radheesh Sharma Meda, and Amir Barati Farimani. Mofgpt: Generative design of metal–organic frameworks using language models. *Journal of Chemical Information and Modeling*, 65(17):9049–9060, 2025. doi: 10.1021/acs.jcim.5c01625. URL https://doi.org/10.1021/acs.jcim.5c01625. PMID: 40875331.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 11 2025. ISSN 0021-9606. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Siddharth Betala, Samuel P. Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K. Joshi, Sékou-Oumar Kaba, Félix Therrien, Alex Hernandez-Garcia, Rocío Mercado, N. M. Anoop Krishnan, and Alexandre Duval. Lematgenbench: A unified evaluation framework for crystal generative models, 2026. URL https://arxiv.org/abs/2512.04562v2.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583 (7815):237–241, Jul 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2442-2. URL https://doi.org/10.1038/s41586-020-2442-2.

Jonas Busk, Mikkel N. Schmidt, Ole Winther, Tejs Vegge, and Peter Bjørn Jørgensen. Graph neural network interatomic potential ensembles with calibrated aleatoric and epistemic uncertainty on energy and forces. *Physical Chemistry Chemical Physics*, 25(37):25828–25837, 10 2023. ISSN 1463-9076. doi: 10.1039/d3cp02143b. URL https://doi.org/10.1039/d3cp02143b.

Shuxiang Cao, Zijian Zhang, Mohammed Alghadeer, Simone D. Fasciati, Michele Piscitelli, Mustafa Bakr, Peter Leek, and Alán Aspuru-Guzik. Automating quantum computing laboratory experiments with an agent-based ai framework. *Patterns*, 6(10):101372, October 2025. ISSN 2666-3899. doi: 10.1016/j.patter.2025.101372. URL http://dx.doi.org/10.1016/j.patter.2025.101372.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, March 1996. ISSN 1076-9757. doi: 10.1613/jair.295. URL http://dx.doi.org/10.1613/jair.295.

Dario Coscia, Pim De Haan, and Max Welling. BLIPs: Bayesian learned interatomic potentials. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZSZW2diTs5.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, Sep 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0519. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1697e3fb412da11dc9488249f9e7bbc9-Paper-Datasets_and_Benchmarks_Track.pdf.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986.

Riley J. Hickman, Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. Gemini: Dynamic bias correction for autonomous experimentation and molecular simulation, 2021. URL https://arxiv.org/abs/2103.03391.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, Jan 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Theo Jaffrelot Inizan, Sherry Yang, Aaron Kaplan, Yen hsu Lin, Jian Yin, Saber Mirzaei, Mona Abdelgaid, Ali H. Alawadhi, KwangHwan Cho, Zhiling Zheng, Ekin Dogus Cubuk, Christian Borgs, Jennifer T. Chayes, Kristin A. Persson, and Omar M. Yaghi. System of agentic ai for the discovery of metal-organic frameworks, 2025. URL https://arxiv.org/abs/2504.14110.

Akshay K. Jagadish, Younes Strittmatter, Nori Jacoby, George Kachergis, Eric Schulz, Nathaniel Daw, Suyog H. Chandramouli, and Thomas L. Griffiths. Closing the loop to discover psychological theories with an automated cognitive scientist, 2026. URL https://arxiv.org/abs/2606.26448.

Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models, 2024. URL https://arxiv.org/abs/2406.13163.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, Dec 1998. ISSN 1573-2916. doi: 10.1023/A:1008306431147. URL https://doi.org/10.1023/A:1008306431147.

A. Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, Albert V. Davydov, Ritesh Agarwal, Leonid A. Bendersky, Mo Li, Apurva Mehta, and Ichiro Takeuchi. On-the-fly closed-loop materials discovery via bayesian active learning. *Nature Communications*, 11(1):5966, Nov 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19597-w. URL https://doi.org/10.1038/s41467-020-19597-w.

Turab Lookman, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, Feb 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0153-8. URL https://doi.org/10.1038/s41524-019-0153-8.

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

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, July 1992. ISSN 1530-888X. doi: 10.1162/neco.1992.4.4.590. URL http://dx.doi.org/10.1162/neco.1992.4.4.590.

Shreshth A Malik, Tiarnan Doherty, Panagiotis Tigas, Muhammed Razzak, S Roberts, Aron Walsh, and Yarin Gal. MADE: Benchmark environments for closed-loop materials discovery. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nrXxVDYMMF.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL https://doi.org/10.1038/s41586-023-06735-9.

Jonas Mockus, Vytautas Tiešis, and Antanas Žilinskas. The application of Bayesian methods for seeking the extremum. In L. C. W. Dixon and G. P. Szegő (eds.), *Towards Global Optimisation 2*, pp. 117–129. North-Holland Publishing Company, Amsterdam, 1978. ISBN 0-444-85171-2.

Andrew Novick, Diana Cai, Quan Nguyen, Roman Garnett, Ryan Adams, and Eric Toberer. Probabilistic prediction of material stability: integrating convex hulls into active learning. *Materials Horizons*, 11(21):5381–5393, 11 2024. ISSN 2051-6347. doi: 10.1039/d4mh00432a. URL https://doi.org/10.1039/d4mh00432a.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16(1):4379, May 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-59636-y. URL https://doi.org/10.1038/s41467-025-59636-y.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Simone Pilon, Elia Savino, Oliver M. Bayley, Michael Vanzella, Miguel Claros, Petros Siasiaridis, Junsong Liu, Florian Lukas, Matteo Damian, Vasilis Tseliou, Niccolò Intini, Aidan Slattery, Jesus SanJosé-Orduna, Tim den Hartog, Ron A. H. Peters, Andrea F. G. Gargano, Francesco G. Mutti, and Timothy Noël. A flexible and affordable self-driving laboratory for automated reaction optimization. *Nature Synthesis*, Apr 2026. ISSN 2731-0582. doi: 10.1038/s44160-026-01053-0. URL https://doi.org/10.1038/s44160-026-01053-0.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1):97–107, Jan 2024. ISSN 2948-1198. doi: 10.1038/s44286-023-00002-4. URL https://doi.org/10.1038/s44286-023-00002-4.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale, 2025. URL https://arxiv.org/abs/2504.06231.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, Jun 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01055-1. URL https://doi.org/10.1038/s42256-025-01055-1.

Brian Rohr, Helge S. Stein, Dan Guevarra, Yu Wang, Joel A. Haber, Muratahan Aykol, Santosh K. Suram, and John M. Gregoire. Benchmarking the acceleration of materials discovery by sequential learning. *Chemical Science*, 11(10):2696–2706, 2020. ISSN 2041-6539. doi: 10.1039/c9sc05999g. URL http://dx.doi.org/10.1039/C9SC05999G.

Martin Seifrid, Robert Pollice, Andrés Aguilar-Granda, Zamyla Morgan Chan, Kazuhiro Hotta, Cher Tian Ser, Jenya Vestfrid, Tony C. Wu, and Alán Aspuru-Guzik. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466, 08 2022. ISSN 0001-4842. doi: 10.1021/acs.accounts.2c00220. URL https://doi.org/10.1021/acs.accounts.2c00220.

Hongyuan Sheng, Jingwen Sun, Oliver Rodríguez, Benjamin B. Hoar, Weitong Zhang, Danlei Xiang, Tianhua Tang, Avijit Hazra, Daniel S. Min, Abigail G. Doyle, Matthew S. Sigman, Cyrille Costentin, Quanquan Gu, Joaquín Rodríguez-López, and Chong Liu. Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation. *Nature Communications*, 15(1):2781, Mar 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-47210-x. URL https://doi.org/10.1038/s41467-024-47210-x.

Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M. Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, Yinkai Wang, Haorui Wang, Jeff Guo, Jingru Gan, Parshin Shojaee, Di Luo, Andres M Bran, Gen Li, Qiyuan Zhao, Shao-Xiong Lennon Luo, Yuxuan Zhang, Xiang Zou, Wanru Zhao, Yifan F. Zhang, Wucheng Zhang, Shunan Zheng, Saiyang Zhang, Sartaaj Takrim Khan, Mahyar Rajabi-Kochi, Samantha Paradi-Maropakis, Tony Baltoiu, Fengyu Xie, Tianyang Chen, Kexin Huang, Weiliang Luo, Meijing Fang, Xin Yang, Lixue Cheng, Jiajun He, Soha Hassoun, Xiangliang Zhang, Wei Wang, Chandan K. Reddy, Chao Zhang, Zhiling Zheng, Mengdi Wang, Le Cong, Carla P. Gomes, Chang-Yu Hsieh, Aditya Nandy, Philippe Schwaller, Heather J. Kulik, Haojun Jia, Huan Sun, Seyed Mohamad Moosavi, and Chenru Duan. Evaluating large language models in scientific discovery, 2026. URL https://arxiv.org/abs/2512.15567v2.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624 (7990):86–91, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06734-w. URL https://doi.org/10.1038/s41586-023-06734-w.

Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10358–10368. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/trabucco21a.html.

Wilailuck Treerath and Fabrizio Pittorino. Active in-context learning for tabular foundation models, 2026. URL https://arxiv.org/abs/2603.27385.

Alex Wang, Haotong Liang, Austin McDannald, Ichiro Takeuchi, and Aaron Gilad Kusne. Benchmarking active learning strategies for materials optimization and discovery. *Oxford Open Materials Science*, 2(1):itac006, 01 2022. ISSN 2633-6979. doi: 10.1093/oxfmat/itac006. URL https://doi.org/10.1093/oxfmat/itac006.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=03RLpj-tc_.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, Mar 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08628-5. URL https://doi.org/10.1038/s41586-025-08628-5.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. Matscibench: Benchmarking the reasoning ability of large language models in materials science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 12868–12879, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818888. URL https://doi.org/10.1145/3770855.3818888.

Wenyu Zhang, Lucy Hao, Veronica Lai, Ryan Corkery, Jacob Jessiman, Jiayu Zhang, Junliang Liu, Yusuke Sato, Maria Politi, Matthew E. Reish, Rebekah Greenwood, Noah Depner, Jiyoon Min, Rama El-khawaldeh, Paloma Prieto, Ekaterina Trushina, and Jason E. Hein. Ivoryos: an interoperable web interface for orchestrating python-based self-driving laboratories. *Nature Communications*, 16(1):5182, Jun 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60514-w. URL https://doi.org/10.1038/s41467-025-60514-w.

A MADE EVALUATION METRICS

We retain MADE's discovery and diversity metrics (Malik et al., 2026). Let $D_\pi(q)$ be the cumulative number of discoveries made by policy π after q oracle queries. Higher is better for every metric in Table 3.

- **Discoveries (Disc.).** The number of proposed structures that are stable within the specified hull threshold, absent from the initial reference set, and distinct from earlier proposals under MADE's structure matching criterion.
- **Area under the discovery curve (AUDC).** $\mathrm{AUDC}_\pi(q) = \frac{2}{q^2}\int_0^q D_\pi(u)\,du$, which rewards discoveries made earlier in the query sequence.
- **Unique compositions (U. comps.).** The number of distinct chemical compositions among the discovered structures.
- **Unique space groups (U. SGs).** The number of distinct crystal space groups among the discovered structures.
- **Enhancement factor (EF).** At a matched query budget q, $\mathrm{EF}_\pi(q) = D_\pi(q)/D_{\mathrm{Control}}(q)$. We report EF at $q = 32$.
- **Acceleration factor (AF).** For each paired run, let j and k be the query counts at which policy π and Control first reach $D_\pi(32)$, respectively. We report the mean of k/j over runs with $D_\pi(32) > 0$. If Control does not reach the target within 300 queries, we set $k = 300$, giving a lower bound.
- **Unmatched.** The number of runs (out of 90) in which Control does not reach policy π's discovery count at $q = 32$ within its full 300-query budget, i.e., $D_{\mathrm{Control}}(300) < D_\pi(32)$.

Under MADE's original full-feedback setting, AUDC is bounded by one because at most one discovery is made per oracle query. Online surrogate repair can produce multiple discoveries per oracle query, so AUDC values above one are valid here. Except for Unmatched, values in Table 3 are means over 90 runs, with parentheses giving the standard error in the final displayed digit or digits.

B SYNTHETIC WORLD CONSTRUCTION

Each world is one frozen SCM with a mixed-type tabular design space and a deterministic scalar response. Following the sampled-DAG construction of TabPFN's synthetic prior (Hollmann et al., 2025), its seed fixes the schema, root distributions, graph, and every mechanism parameter. Separate row-sampling streams ensure that a design has the same response whether evaluated alone or in a batch.

A world has 14–22 observable roots drawn from six types: continuous, integer, ordinal, nominal, Boolean, and cyclic. Continuous roots follow rescaled Beta laws; integers are uniform over finite ranges; ordinal and nominal probabilities are drawn from Dirichlet distributions; Booleans follow Bernoulli laws; and cyclic values are uniform over periods of 12, 24, or 360. Roots are encoded to $[-1, 1]$ before graph evaluation, with cyclic features represented by the sine of their phase.

Above the roots is a 5–8-layer DAG with 44–72 hidden nodes and 2–6 parents per node. Its twelve nonlinear mechanisms are affine, neural, tree, quadratic, Fourier, oscillator, radial basis, threshold, plateau, product, gated, and categorical. The scalar response combines 4–7 final-layer branches, two pairwise products, and a direct root term together with its sine. Five mechanism profiles produce smooth multiscale, tree/threshold/plateau, oscillatory, localized-basin, and categorical/quantized families. We use the third disjoint cohort, W5241–W5260, balanced at four worlds per family; these worlds are not used to train the surrogates.

For each world, preparation draws a 1,000-row initial context and a held-out 50,000-row audit sample. A guided world-level search uses 10 restarts, 20 iterations, and 500 proposals per iteration. The reference response is the larger maximum found by the audit sample and guided search, and R_w is the audit sample's $q_{0.95} - q_{0.05}$ range. Static EI and Global-UQ select their 32 queries as a single batch from a shared, locked pool of 10,000 first-occurrence rows using the initial surrogate. This well-established SCM construction provides a reproducible, controlled basis for comparing acquisition rules and schedules on identical worlds.

C QUERY-BUDGET SWEEP AND SELECTION GEOMETRY

C.1 ONLINE EI QUERY-BUDGET SWEEP

Table 4: Online EI query-budget sweep on the 20 synthetic worlds, all with TabPFN. Values are final normalized regret with 95% CIs (lower is better).

Method	Oracle queries	Mean regret (95% CI)	Median
Online EI (Q=64)	64	**0.0481** (0.0233, 0.0760)	**0.0139**
Online EI (Q=32)	32	<u>0.0552</u> (0.0298, 0.0843)	<u>0.0393</u>
Online EI (Q=16)	16	0.0775 (0.0502, 0.1060)	0.0813
Online EI (Q=8)	8	0.0918 (0.0599, 0.1251)	0.0831

Online EI improves monotonically over the eightfold range from Q=8 to Q=64, so its advantage does not depend on a narrow choice of budget. Returns diminish: the step from Q=8 to Q=32 reduces mean regret by 0.0366, whereas doubling again to Q=64 reduces it by a further 0.0071. Even at Q=8, four times fewer queries than the main setting, Online EI reaches 0.0918, still ahead of Static EI at Q=32 (0.0984) and well ahead of no repair (0.1432).

C.2 SELECTION GEOMETRY OF ONLINE ACQUISITION RULES

We characterize the 32 acquisitions made by each online method in each world. Designs use the world-specific $[-1, 1]$ representation, with wraparound distance for cyclic features. Let d_0 denote the mean pairwise RMS distance among the 1,000 initial context designs. Step is the distance between successive acquisitions, while Spread is the mean pairwise distance among all acquired designs, and both are normalized by d_0. An incumbent update occurs when an acquired label exceeds the previous verified incumbent. The Step statistics comprise 620 successive acquisition pairs per method across the 20 worlds.

Table 5: Selection geometry of the four TabPFN online methods. Step and Spread are normalized by the initial-context scale d_0; incumbent updates are reported per world.

Method	Step / d_0	Spread / d_0	Incumbent updates	Mean regret
Online Peak	0.079	0.189	6.1	0.0976
Online Q90-UCB	0.189	0.255	6.0	0.0600
Online EI	0.356	0.448	5.6	0.0552
Tab-AICL	0.915	0.916	1.1	0.1169

Step and Spread have the same ordering: Peak, Q90-UCB, EI, and Tab-AICL. EI moves $1.9\times$ farther per step than Q90-UCB and $4.5\times$ farther than Peak. Regret, however, is not monotonic in either measurement. Peak concentrates its acquisitions within a small region, whereas Tab-AICL covers nearly the full initial-context scale; their mean regrets are 0.0976 and 0.1169. EI lies between these extremes and obtains the lowest mean regret, 0.0552. Neither local concentration nor broad coverage is therefore sufficient for low regret. This distinction is consistent with the acquisition objectives: Peak uses an incumbent reference without uncertainty, Tab-AICL targets coverage without an incumbent reference, and EI combines uncertainty with improvement over a verified incumbent.

Incumbent update frequency also does not explain the performance ordering. Peak, Q90-UCB, and EI update the incumbent 6.1, 6.0, and 5.6 times per world, despite their different final regrets. The three methods also take smaller steps following an update: 0.032 rather than 0.081 for Peak, 0.119 rather than 0.183 for Q90-UCB, and 0.186 rather than 0.350 for EI. EI's advantage is therefore not associated with greater dispersion immediately after finding a new incumbent.

These measurements establish an association between acquisition geometry and final regret, rather than a causal mechanism. Archive-BO is omitted here to compare acquisition rules under the same TabPFN surrogate.

D FULL STATIC SURROGATE REPAIR RESULTS

Table 6 reports the complete final maximum-regret and NRMSE matrix summarized in Table 1.

Table 6: Static surrogate repair with Q=32 across TabPFN-3 and TabICLv2 at context sizes $N \in \{100, 1000\}$. **Bold** best, underlined second best per metric within each block.

Method	Final $R_{\max}\downarrow$	$\Delta R\uparrow$	Final NRMSE $\downarrow$	$\Delta E\uparrow$
TabPFN-3, N=100				
Q90-UCB	**0.5672**	**0.4978**	0.2248	0.0023
EI	0.6133	0.4516	0.2219	0.0052
Peak	0.6593	0.4057	0.2273	−0.0002
G2P(16)	0.7555	0.3094	0.2227	0.0044
Global-UQ	0.9955	0.0694	**0.2183**	**0.0088**
Random	1.0084	0.0565	0.2197	0.0074
TabPFN-3, N=1000				
Q90-UCB	0.5583	0.2393	0.1736	−0.0006
EI	**0.5113**	**0.2862**	0.1731	−0.0002
Peak	0.5361	0.2615	0.1741	−0.0011
G2P(16)	0.6303	0.1672	0.1730	−0.0001
Global-UQ	0.7993	−0.0018	**0.1723**	**0.0007**
Random	0.7787	0.0188	0.1724	0.0005
TabICLv2, N=100				
Q90-UCB	**0.5694**	**0.5909**	0.2313	−0.0009
EI	0.5918	0.5686	0.2301	0.0004
Peak	0.6010	0.5593	0.2330	−0.0025
G2P(16)	0.6712	0.4891	0.2284	0.0021
Global-UQ	1.0347	0.1256	0.2223	0.0082
Random	1.0147	0.1456	**0.2209**	**0.0095**
TabICLv2, N=1000				
Q90-UCB	**0.4752**	**0.3664**	0.1792	−0.0017
EI	0.5107	0.3309	0.1786	−0.0011
Peak	0.5200	0.3216	0.1804	−0.0030
G2P(16)	0.5799	0.2617	0.1786	−0.0012
Global-UQ	0.7823	0.0593	**0.1769**	0.0005
Random	0.8545	−0.0129	**0.1769**	**0.0006**

E COST ESTIMATION

All experiments ran on two H100 SXM GPUs, with runtime dominated by waiting for LLM responses and comparable GPU costs for Online EI and Control. LLM charges were estimated from records collected through OpenAI's native Responses API and OpenRouter's Chat Completions API, with uncertainty from stochastic context caching and provider routing. At matched mean discovery counts, Online EI spends an additional \$0.33 (DeepSeek) and \$0.68 (Luna) per trajectory relative to Control (Table 7), a small increase compared with a typical wet-lab synthesis and characterization cycle, as motivated in Section 1.

Table 7: Estimated costs per MADE trajectory. LLM costs are in USD; H100-s denotes GPU-seconds. Discovery counts follow Table 3. Catch-up denotes Control reaching Online EI's final discovery count. Cost per discovery includes LLM charges only.

Operating point	LLM cost ($)	H100-s	Discoveries	$/discovery
DeepSeek-V4-Flash				
Online EI, full run (Q=32)	1.35	922	53.9	0.025
Control, catch-up (ep. 224.5)	1.02	813	53.9	0.019
Control, matched (Q=32)	0.12	102	13.07	0.009
Control, full run (Q=300)	1.44	1,125	65.4	0.022
GPT-5.6-Luna				
Online EI, full run (Q=32)	2.02	342	56.9	0.036
Control, catch-up (ep. 190.6)	1.34	387	56.9	0.024
Control, matched (Q=32)	0.16	63	9.63	0.017
Control, full run (Q=300)	2.22	652	84.6	0.026

F TAB-AICL ADAPTATION

We adapt Tab-AICL's hybrid rule (Treerath & Pittorino, 2026) to regression with one label per acquisition. Classification entropy is replaced by the surrogate's native 80% predictive interval width, $q_{0.90,t}(x) - q_{0.10,t}(x)$. We retain the original shortlist size $N_{\text{cand}} = \min(M, \max(2, \lfloor M/2 \rfloor))$, where $M = |\mathcal{A}_t|$, keeping the most uncertain archived designs. For $B = 1$, we replace batch k-means with the paper's Coreset criterion: select the shortlisted design maximizing $D_{\min,t}(x) = \min_{(x',y') \in \mathcal{D}_t} \|z_t(x) - z_t(x')\|_2$.

The preprocessing map z_t standardizes numerical features and ordinal-encodes categorical features using the current labelled context. The surrogate, $Q = 32$ budget, query schedule, and candidate archive match Online EI, only the selection rule differs.

G ADDITIONAL MADE RESULTS

MADE counts distinct discoveries across a campaign, rather than only its best design. In the full-feedback setting, every episode proposal receives an oracle evaluation. For OSR, restricting evaluation to the 32 queries used at run time would exclude the remaining proposals and cap the discovery count at 32, even when those proposals contain additional discoveries. This would confound discovery yield with which proposals receive feedback. Table 3 therefore applies the same post-hoc oracle evaluation to all episode proposals from every method. Audit labels do not enter the search or surrogate context, and audit oracle calls are excluded from the feedback budget Q.

We report the main evaluation (*audited*) alongside *budgeted* Online EI, which uses only its 32 run time oracle outcomes for every metric. All runs retain 300 episodes and the settings of Table 3.

We also evaluate the best materials found. Let $h_0(x)$ denote the signed ORB energy gap to the fixed initial hull, in meV/atom, and define $s(x) = \max(0, 100 - h_0(x))$. Let $S_{k,\pi}(q)$ be the mean of the k highest scores among novel, distinct materials evaluated by policy π at query horizon q, assigning zero to missing slots.

- **Top-1 and Top-5.** $S_{1,\pi}(q)$ and $S_{5,\pi}(q)$, respectively, higher scores indicate better material quality.
- **Top-5 quality AUC.** $\frac{1}{q}\int_0^q S_{5,\pi}(u)\,du$, evaluated by trapezoidal integration, rewards earlier attainment of a good shortlist.
- **EF-top.** The ratio of mean Top-1 scores, $\overline{S_{1,\pi}(q)}/\overline{S_{1,\mathrm{Control}}(q)}$, at the same query horizon.
- **AF-top.** The paired first-passage ratio used for AF in Appendix A, with target $S_{1,\pi}(q)$. Control is searched through 300 queries at either reported horizon; zero targets are excluded and unreached targets contribute lower bounds.

Both tables report means over 30 systems $\times$ three repeats, with bootstrap standard errors in parentheses. EF and EF-top hold the pooled Control denominator fixed when estimating these errors.

Table 8: MADE material quality at matched query horizons under the settings of Table 3. Top-1, Top-5, and Top-5 AUC are in meV/atom. **Bold** best, underlined second best per metric column within each block. † marks AF-top estimates containing lower bounds; Unmatched reports unreached / positive AF-top targets.

Method	Top-1		Top-5		Top-5 AUC		AF-top		Unmatched		EF-top	
	@32	@300	@32	@300	@32	@300	@32	@300	@32	@300	@32	@300
DeepSeek-V4-Flash												
Online EI (audited)	**129.8(65)**	—	**109.9(48)**	—	**95.5(41)**	—	**49(11)**†	—	**46/90**	—	**1.136(57)**	—
Online EI (budgeted)	126.0(63)	—	104.7(43)	—	77.1(34)	—	24.8(57)†	—	38/90	—	1.102(55)	—
Vanilla	110.8(65)	**131.0(62)**	78.3(39)	110.9(40)	54.7(33)	97.0(37)	19.6(51)†	**12.0(45)**†	19/90	**51/90**	0.970(57)	**1.006(47)**
Control	114.3(60)	130.2(60)	84.1(34)	**111.8(40)**	59.8(30)	**99.1(35)**	1.00	1.00	—	—	1.00	1.00
Random	49.6(59)	89.4(35)	24.0(39)	69.6(39)	13.3(23)	49.1(40)	1.71(51)†	0.78(32)†	2/78	7/90	0.434(52)	0.687(27)
GPT-5.6-Luna												
Online EI (audited)	**137.6(61)**	—	**116.8(47)**	—	**101.4(39)**	—	**48.0(72)**†	—	**41/90**	—	**1.466(65)**	—
Online EI (budgeted)	133.7(64)	—	111.9(49)	—	87.3(41)	—	34.9(58)†	—	36/90	—	1.424(69)	—
Vanilla	94.7(48)	138.6(73)	67.2(39)	**119.4(58)**	39.2(29)	**99.1(44)**	6.4(15)†	**4.4(14)**†	9/87	**38/90**	1.009(51)	0.985(52)
Control	93.9(50)	**140.8(78)**	66.3(39)	118.6(55)	36.8(28)	96.2(38)	1.00	1.00	—	—	1.00	**1.00**
Random	49.6(59)	89.4(35)	24.0(39)	69.6(39)	13.3(23)	49.1(40)	8.1(44)†	1.77(67)†	5/78	10/90	0.529(63)	0.635(25)
Chemeleon+MLIP												
Online EI (audited)	**128.4(58)**	—	**109.0(41)**	—	**105.8(40)**	—	**117(16)**†	—	**49/90**	—	**1.028(46)**	—
Online EI (budgeted)	125.5(57)	—	106.1(38)	—	89.1(34)	—	45(10)†	—	41/90	—	1.005(45)	—
Vanilla	124.8(56)	**127.6(55)**	103.0(37)	108.6(40)	85.1(37)	105.5(38)	50.9(95)†	**50.8(95)**†	39/90	**41/90**	0.999(45)	**1.000(43)**
Control	124.9(54)	127.5(55)	106.1(36)	**108.8(39)**	90.2(32)	**106.4(37)**	1.00	1.00	—	—	1.00	**1.00**
Random	49.6(59)	89.4(35)	24.0(39)	69.6(39)	13.3(23)	49.1(40)	0.71(30)†	0.47(29)†	1/78	5/90	0.397(47)	0.701(27)

At matched $Q = 32$, Online EI improves the quality of the best materials even when evaluation uses only paid oracle outcomes (Table 8). Under DeepSeek and Luna, budgeted Top-1 scores exceed

Control by 10.2% and 42.4%, and Top-5 scores by 24.5% and 68.8%, respectively. Top-5 quality AUC increases by 28.9% and 137.2%, showing higher average shortlist quality over the first 32 queries as well as better endpoint scores. Chemeleon+MLIP is near parity with Control on these budgeted quality comparisons.

The advantage also appears when comparing the short oracle budget with full-feedback endpoints. Across the three workflows, Online EI's 32 run time queries retain 95.0–98.4% of Control's 300-query mean Top-1 score and 93.6–97.5% of its Top-5 score, using 10.7% of the feedback queries over the same 300-episode search horizon. Budgeted AF-top estimates are 24.8, 34.9, and 45 for DeepSeek, Luna, and Chemeleon+MLIP, these include lower-bound contributions when Control does not attain the quality target within 300 queries.

For Online EI, Chemeleon+MLIP leads discovery counts, whereas Luna leads mean Top-1 and Top-5 under both accountings. DeepSeek is closer to Chemeleon in endpoint quality, and Chemeleon leads Top-5 quality AUC.

Table 9 gives the corresponding discovery and diversity results. Budgeted Online EI retains discovery EF of 1.451 and 1.938 under DeepSeek and Luna, with AF of 2.04 and 2.19, Chemeleon+MLIP reahces near parity with Control.

Table 9: Additional MADE discovery results. Settings and highlighting follow Table 8. *AUDC can exceed one under audited proposal evaluation.

Method	EF	AF	Disc.		Unmatched	AUDC		U. comps.		U. SGs	
	@32	@k	@32	@300	@full budget	@32	@300	@32	@300	@32	@300
DeepSeek-V4-Flash											
Online EI (audited)	**4.13(21)**	**7.23(28)**	**53.9(28)**	—	**21/90**	**1.96(11)***	—	**18.9(12)**	—	**22.03(77)**	—
Online EI (budgeted)	1.451(43)	2.04(13)	18.96(56)	—	0/90	0.583(19)	—	10.88(48)	—	12.60(40)	—
Vanilla	0.882(51)	1.077(74)	11.52(67)	63.8(31)	0/90	0.401(21)	0.243(12)	9.50(58)	21.2(14)	8.16(32)	**23.08(59)**
Control	1.00	1.00	13.07(72)	**65.4(34)**	—	0.452(23)	**0.256(13)**	10.88(70)	20.9(15)	8.57(38)	23.03(58)
Random	0.303(55)	0.363(34)	3.96(71)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	**35.3(62)**	1.36(14)	3.61(31)
GPT-5.6-Luna											
Online EI (audited)	**5.91(29)**	**6.36(20)**	**56.9(29)**	—	**7/90**	**2.07(11)***	—	**19.0(11)**	—	**23.57(58)**	—
Online EI (budgeted)	1.938(73)	2.19(13)	18.67(70)	—	0/90	0.616(21)	—	10.10(47)	—	12.26(42)	—
Vanilla	1.035(72)	1.280(77)	9.97(69)	**86.3(49)**	0/90	0.291(22)	**0.307(17)**	5.52(50)	20.3(14)	7.19(46)	27.16(94)
Control	1.00	1.00	9.63(67)	84.6(47)	—	0.286(22)	0.296(15)	5.24(41)	20.8(14)	7.23(47)	**27.30(84)**
Random	0.411(74)	0.879(89)	3.96(70)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	**35.3(61)**	1.36(14)	3.61(31)
Chemeleon+MLIP											
Online EI (audited)	**4.55(48)**	**10.27(92)**	**130(14)**	—	**31/90**	**5.26(43)***	—	**119(12)**	—	**23.1(18)**	—
Online EI (budgeted)	0.985(19)	0.9863(70)	28.23(55)	—	0/90	0.866(18)	—	26.92(51)	—	13.10(59)	—
Vanilla	0.914(39)	1.02(12)	26.2(11)	134(15)	1/90	0.828(36)	0.561(48)	24.9(10)	124(13)	11.66(64)	23.5(17)
Control	1.00	1.00	28.68(53)	**141(16)**	—	0.900(15)	**0.595(50)**	27.32(48)	**128(14)**	13.46(58)	**24.3(19)**
Random	0.138(25)	0.209(23)	3.96(71)	36.4(63)	0/90	0.121(21)	0.120(21)	3.91(70)	35.3(62)	1.36(14)	3.61(31)