MECHVAR: VARIANCE-GUIDED MECHANISM DISCRIMINATION FOR AUTONOMOUS ML EXPERIMENT SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A reproduced benchmark gain does not by itself identify the mechanism that caused it. We study finite-library mechanism discrimination: given posterior-weighted candidate mechanisms, executable probes, and a limited budget, which experiment should be run next to separate the remaining explanations? MECHVAR scores each probe by the posterior-weighted variance of its predicted responses. Under a shared-Gaussian predictive model, this quantity is exactly proportional to the classical Box–Hill posterior-weighted pairwise-KL criterion, while admitting an $O(KE)$ vectorized evaluation and an additive mechanism-pair audit. A local expansion further relates the score to expected information gain (EIG) for small response separations. In a 25-block stress audit, MECHVAR outperforms confirmation-first in several moderate misspecification regimes; its primary comparisons with EIG are statistically unresolved. In a held-out Digits loop, normalized mechanism-identification AUC is 0.8975 for MECHVAR, 0.7825 for a score-greedy policy, and 0.9092 for EIG. At $K = 100, E = 200$, median single-thread full-library scoring is $10.36\,\mu$s for MECHVAR and 57.69 ms for six-node quadrature EIG in the recorded environment. MECHVAR is therefore a lightweight, auditable rule for finite-library experiment selection when a shared predictive scale is a defensible approximation.

1 INTRODUCTION

A benchmark gain can have several explanations. A graph-aware model may exploit relational structure, benefit from a change in effective capacity, or use the graph as an identifier-like shortcut. A semi-supervised learner may improve through unlabeled data, augmentation, or an optimization side effect. Reproducing the gain alone does not distinguish among these mechanisms.

Autonomous ML workflows need to decide how to investigate such ambiguity. Research agents can modify code, launch training jobs, inspect metrics, and propose follow-up experiments (Huang et al., 2024; Chan et al., 2025; Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025). Given a finite set of plausible mechanisms and executable interventions, the next task is to allocate a limited budget to experiments that distinguish explanations with substantial posterior support.

We formulate this task as sequential model discrimination. Mechanisms are candidate explanations for an observed effect; probes are ablations, perturbations, or controlled retraining runs. The objective is to distinguish mechanisms that remain plausible under the posterior. Maximizing benchmark accuracy answers a different question. In the held-out Digits experiment (Section 4.5), score-greedy uses the same posterior and calibration templates as the discrimination policies but identifies the hidden mechanism less efficiently.

MECHVAR scores a probe by the posterior-weighted variance of its predicted responses. The weighted-variance identity expresses this score as a sum of pairwise squared response differences. Under the belief model's shared predictive variance, the classical ordered-pair Gaussian Box–Hill criterion is exactly this variance multiplied by $1/\sigma_b^2$ (Box and Hill, 1967). The reduction preserves the selected probe, makes the contribution of each mechanism pair explicit, and evaluates the complete library in $O(KE)$ arithmetic.

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

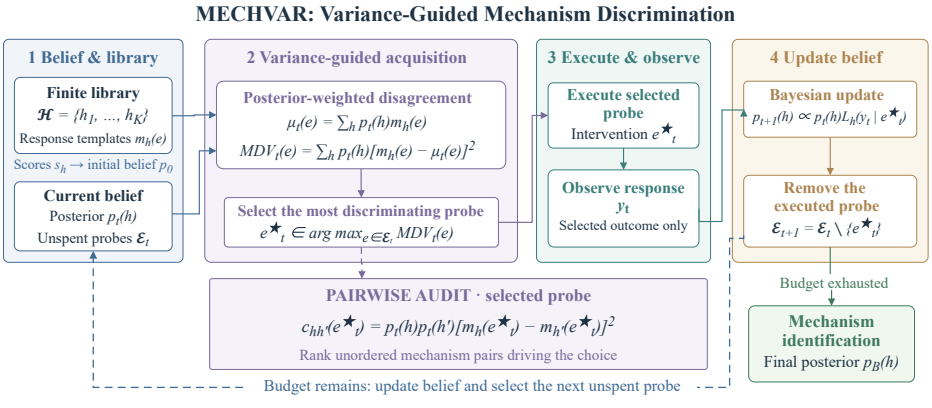


Figure 1: Finite-library mechanism-discrimination loop. The posterior and response templates determine the MDV score; the pairwise decomposition records which mechanism pairs contribute most to the selected probe. Only the executed outcome is revealed before the posterior update.

The method is deliberately narrow: it is not a new divergence and it does not generate hypotheses or probes. Its role is to make a classical model-discrimination objective cheap to rescore and easy to audit inside an iterative ML workflow. We test that role under model mismatch, small budgets, larger mechanism libraries, and a held-out loop that executes one selected intervention before updating the posterior.

The main contributions are:

- We cast follow-up ablations and interventions as finite-library mechanism identification and separate this objective empirically from maximizing the next experiment's benchmark score.
- We show that, under shared Gaussian predictive variance, the Box–Hill pairwise-KL objective reduces exactly to posterior-weighted response variance; this yields $O(KE)$ full-library rescoring, mechanism-pair attribution, and a local connection to EIG.
- We evaluate the resulting acquisition rule under forecast and likelihood mismatch, budget and mechanism-set scaling, structured errors, and disjoint held-out execution, with paired inference, quadrature diagnostics, and a documented score-only runtime benchmark.

2 Related Work

Bayesian experimental design chooses measurements or interventions to maximize expected utility under uncertainty (Lindley, 1956; Chaloner and Verdinelli, 1995). Model discrimination uses this framework to separate competing explanations. The Box–Hill criterion combines posterior model probabilities with predictive separation (Box and Hill, 1967). Our contribution is not to replace that criterion: under shared-Gaussian predictive uncertainty, we expose an exact variance form of the posterior-weighted pairwise-KL objective. The form matters operationally because it supports vectorized rescoring of a finite probe library and an additive attribution to the mechanism pairs driving each choice.

Information gain is another common objective for query selection. BALD uses mutual information between predictions and model uncertainty (Houlsby et al., 2011). In our finite-mechanism setting, EIG is $I(H; Y \mid e)$ and serves as the information-theoretic comparator. Computational approaches to sequential design include Deep Adaptive Design, which learns a policy mapping experiment history to the next design (Foster et al., 2021); Rainforth et al. (2024) review variational, amortized, and policy-based methods. MECHVAR assumes that mechanisms and probes have already been enumerated. It rescans this library as the posterior changes and records why each probe is selected. Our EIG implementation uses deterministic Gauss–Hermite quadrature; Appendix E compares the six-node rule with higher orders.

Research agents automate several stages surrounding experiment selection. MLAgentBench and MLE-bench evaluate agents that modify code, train models, and assess results (Huang et al., 2024; Chan et al., 2025); ScienceAgentBench covers scientific data-analysis tasks (Chen et al., 2025). Other systems combine hypothesis generation, experimentation, and iterative refinement (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Gottweis et al., 2025), while FIRE-Bench evaluates executable rediscovery of scientific findings (Wang et al., 2026). MECHVAR addresses the selection step once candidate mechanisms and probes are available. It does not generate new hypotheses or experiments.

3 Mechanism-Discriminating Experiment Planning

3.1 Finite-Hypothesis Setting

Let $\mathcal{H} = \{h_1, \dots, h_K\}$ denote a finite set of candidate mechanisms and $\mathcal{E} = \{e_1, \dots, e_E\}$ a finite library of candidate probes. Each mechanism h predicts a scalar response $m_h(e)$ under probe e. Observational evidence is summarized by mechanism scores s_h, which are converted into an initial distribution

$$p_0(h) \propto \exp\left[-\frac{s_h - \min_j s_j}{\tau_s}\right], \tag{1}$$

where τ_s sets the influence of the observational ranking on the prior. The stress audit allows an incorrect decoy mechanism to appear favorable in these scores, so the initial ranking need not identify the true mechanism.

After selecting probe e_t and observing response y_t, the planner updates its posterior according to

$$y_t \mid h, e_t \sim \mathcal{N}\big(m_h(e_t), \sigma_b^2\big), \qquad p_{t+1}(h) \propto p_t(h) \exp\left[-\frac{(y_t - m_h(e_t))^2}{2\sigma_b^2}\right]. \tag{2}$$

The experimental budget restricts execution to $B < E$ probes.

3.2 MECHVAR and the Box–Hill Equivalence

Define the posterior-weighted predictive mean and the corresponding mechanism-disagreement variance as

$$\mu_t(e) = \sum_h p_t(h) m_h(e), \tag{3}$$

$$\mathrm{MDV}_t(e) = \sum_h p_t(h)\big(m_h(e) - \mu_t(e)\big)^2. \tag{4}$$

MECHVAR selects

$$e_t^{\textsc{MechVar}} \in \arg\max_{e \in \mathcal{E}_t} \mathrm{MDV}_t(e). \tag{5}$$

Proposition 1 (Shared-Gaussian equivalence)**.** *Assume*

$$Y \mid h, e \sim \mathcal{N}\big(m_h(e), \sigma_b^2\big),$$

where the predictive variance $\sigma_b^2 > 0$ is shared across mechanisms and probes. Then

$$\mathrm{MDV}_t(e) = \sum_{h<h'} p_t(h) p_t(h')\big(m_h(e) - m_{h'}(e)\big)^2, \tag{6}$$

and the posterior-weighted ordered-pair KL criterion

$$\mathrm{BH}_t(e) = \sum_{h,h'} p_t(h) p_t(h') D_{\mathrm{KL}}\big(\mathcal{N}(m_h(e), \sigma_b^2) \,\|\, \mathcal{N}(m_{h'}(e), \sigma_b^2)\big) \tag{7}$$

satisfies

$$\mathrm{BH}_t(e) = \frac{\mathrm{MDV}_t(e)}{\sigma_b^2}. \tag{8}$$

Consequently, MDV and the Box–Hill criterion induce the same argmax set over candidate probes.

Proof. For brevity, write $p_h = p_t(h)$, $m_h = m_h(e)$, and $\mu = \sum_h p_h m_h$. Expanding the weighted variance gives

$$\mathrm{MDV}_t(e) = \sum_h p_h (m_h - \mu)^2 \tag{9}$$

$$= \sum_h p_h m_h^2 - \mu^2 \tag{10}$$

$$= \sum_h p_h m_h^2 - \sum_{h,h'} p_h p_{h'} m_h m_{h'} \tag{11}$$

$$= \frac{1}{2} \sum_{h,h'} p_h p_{h'} (m_h - m_{h'})^2 \tag{12}$$

$$= \sum_{h<h'} p_h p_{h'} (m_h - m_{h'})^2. \tag{13}$$

For equal-variance Gaussians,

$$D_{\mathrm{KL}}\big(\mathcal{N}(m_h, \sigma_b^2) \,\|\, \mathcal{N}(m_{h'}, \sigma_b^2)\big) = \frac{(m_h - m_{h'})^2}{2\sigma_b^2}. \tag{14}$$

Substituting into the ordered-pair criterion yields

$$\mathrm{BH}_t(e) = \frac{1}{2\sigma_b^2} \sum_{h,h'} p_h p_{h'} (m_h - m_{h'})^2 = \frac{\mathrm{MDV}_t(e)}{\sigma_b^2}. \tag{15}$$

Because the multiplier is positive and independent of the probe, it does not affect the maximizing set. □

Equation (6) further yields the pairwise attribution

$$c_{hh'}(e) = p_t(h) p_t(h') \big(m_h(e) - m_{h'}(e)\big)^2, \qquad h < h', \tag{16}$$

with

$$\mathrm{MDV}_t(e) = \sum_{h<h'} c_{hh'}(e). \tag{17}$$

For a selected probe, ranking $c_{hh'}(e)$ identifies the mechanism pairs that contribute most to the score, accounting for both posterior weight and predicted response separation.

Corollary 1 (Vectorized scoring cost)**.** *Let $R \in \mathbb{R}^{K \times E}$ collect the response templates, with $R_{he} = m_h(e)$, and let $p \in \mathbb{R}^K$ denote the current posterior. The scores of all E probes can be computed jointly as*

$$\boldsymbol{\mu} = p^\top R, \qquad \boldsymbol{s} = p^\top (R \odot R) - \boldsymbol{\mu} \odot \boldsymbol{\mu}, \tag{18}$$

in $O(KE)$ arithmetic. The $O(K^2)$ pairwise expansion is needed only when an attribution is requested for a selected probe.

Proposition 2 (Local EIG agreement)**.** *Fix a probe e and write $a_h(e) = m_h(e) - \mu_t(e)$, so that $\sum_h p_t(h) a_h(e) = 0$. Consider the locally scaled predictive family*

$$Y_\lambda \mid h, e \sim \mathcal{N}\big(\mu_t(e) + \lambda a_h(e), \sigma_b^2\big)\,.$$

With mutual information measured in nats, as $\lambda \to 0$,

$$I(H; Y_\lambda \mid e) = \frac{\lambda^2}{2\sigma_b^2}\, \mathrm{MDV}_t(e) + O(\lambda^4). \tag{19}$$

Thus, for any two probes with a nonzero MDV gap, sufficiently small response separations induce the same pairwise ordering under EIG and MECHVAR. The proof, based on the Gaussian-channel I–MMSE identity (Guo et al., 2005), is given in Appendix A. The statement is local: higher-order terms can change the ordering once response separations are no longer small compared with σ_b.

The worked example in Appendix B illustrates this attribution: one posterior-plausible pair accounts for 79.9% of the selected probe's MDV score.

Remark. *The equivalence requires predictive variance to be shared across mechanisms and probes. Under heteroscedastic predictive uncertainty, KL-based and variance-based acquisition can rank probes differently. The prediction-misspecification experiments in Section 4.3 perturb the response-template matrix available to the planner while retaining a shared σ_b. Conditional on any realized perturbed matrix, Proposition 1 therefore continues to hold internally: misspecification changes how accurately the planner's templates represent the hidden task, but not the algebraic relationship between MDV and the Box–Hill score under the planner's assumed belief model.*

4 Experimental Evaluation

4.1 Baselines and Statistical Protocol

Random selects uniformly from the remaining probes. For confirmation-first, let $\hat{h}_t \in \arg\max_h p_t(h)$ and renormalize the remaining posterior as $\tilde{p}_t(h) = p_t(h)/(1 - p_t(\hat{h}_t))$ for $h \neq \hat{h}_t$. Its acquisition score is

$$C_t(e) = \left[m_{\hat{h}_t}(e) - \sum_{h \neq \hat{h}_t} \tilde{p}_t(h) m_h(e) \right]^2, \tag{20}$$

and it selects the remaining probe with maximum $C_t(e)$. EIG maximizes $I(H; Y \mid e)$ under the same Gaussian belief model and is evaluated using deterministic six-node Gauss–Hermite quadrature. The held-out Digits benchmark additionally includes score-greedy,

$$G_t(e) = \sum_h p_t(h) q_h(e), \tag{21}$$

where $q_h(e)$ is the calibration-template accuracy for mechanism h under intervention e.

The primary stress audit defines a nine-condition family for each comparator: five prediction-misspecification levels and four non-duplicate likelihood-misspecification levels. The paired block, not an individual hidden task, is the inferential unit. We apply paired Wilcoxon tests with Holm adjustment within each family (Holm, 1979), supplemented by exact sign tests after discarding zero differences in near-ceiling settings. Budget, mechanism-count, and structured-error sweeps are exploratory. For Digits, each held-out seed forms one paired block; the four AUC comparisons involving MECHVAR constitute a separate Holm family.

4.2 Motivating Diagnostics

The motivating studies show how similar benchmark observations can support different mechanism-level interpretations. A ColoredMNIST-style color swap raises OOD accuracy by 0.6028 (95% CI $[0.5588, 0.6469]$); a 50-node NOTEARS study changes from mean SHD 79.7 to 160.0 under permutation; and a global class permutation in CIFAR-10→SVHN collapses classification accuracy while leaving maximum-softmax AUROC effectively unchanged. These studies motivate the discrimination objective but do not evaluate the planner. Figure 5 and the full six-study summary appear in Appendix C.

4.3 Robustness Under Misspecification

The audit contains 25 paired blocks of 50 hidden tasks, for 1,250 tasks per condition. Each task has five candidate mechanisms, ten probes, and a budget of four. Three probes have reduced separation between mechanisms. Observational scores are noisy, and an incorrect decoy receives an apparent advantage with probability 0.58. Policies share the underlying tasks, prediction matrices, realized observations, and budgets.

The planner assumes $\sigma_b = 0.22$. Prediction misspecification perturbs each mechanism–probe response forecast before selection, with

$$\sigma_{\text{pred}} \in \{0, 0.10, 0.25, 0.50, 1.00\}.$$

Likelihood misspecification instead varies the true-to-assumed noise ratio as

$$\frac{\sigma_{\text{true}}}{\sigma_b} \in \{0.5, 1.5, 2, 3\},$$

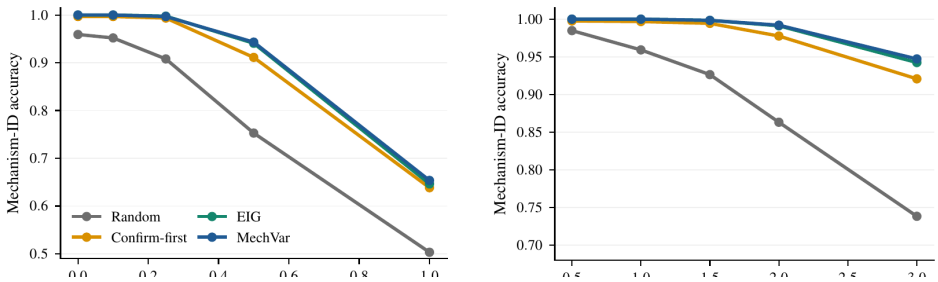


Figure 2: Full misspecification sweeps. Left: prediction misspecification. Right: likelihood-noise mismatch. The clearest resolved differences from confirmation-first occur in the moderate misspecification regimes highlighted in the text, while the primary MECHVAR–EIG comparisons remain statistically unresolved.

Table 1: Selected stress-audit conditions. Accuracy columns report mechanism-identification percentages. Parentheses show family-wise Holm-adjusted Wilcoxon p-values for the indicated comparator.

Condition	Random	Confirm-first	EIG	MECHVAR	Δ_{M-C} (Holm p)	Δ_{M-E} (Holm p)
Well specified	95.92	99.68	100.00	100.00	+0.32 (0.228)[†]	0.00 (–)
Pred. $\sigma = .5$	75.28	91.12	94.08	94.32	+3.20 (0.00196)	+0.24 (1.000)
Pred. $\sigma = 1$	50.32	63.84	64.64	65.36	+1.52 (0.497)	+0.72 (1.000)
Noise 2×	86.32	97.76	99.12	99.20	+1.44 (0.0276)	+0.08 (1.000)
Noise 3×	73.84	92.08	94.24	94.72	+2.64 (0.00800)	+0.48 (1.000)

[†]Only 4/25 paired blocks have nonzero differences; the exact sign-test gives $p = 0.125$. The near-ceiling design therefore has low power for small between-policy differences.

while keeping the planner's update rule fixed at σ_b. Latent block seeds are shared across policies wherever applicable.

Accuracy is near ceiling in the well-specified condition: random reaches 95.92%, confirmation-first 99.68%, and MECHVAR and EIG both reach 100%. MECHVAR and confirmation-first differ in only four of the 25 paired blocks. The exact sign-test value is $p = 0.125$, and the Holm-adjusted Wilcoxon value is 0.228. With so few nonzero differences, this condition has limited power to detect small policy differences and cannot establish equivalence.

At $\sigma_{\text{pred}} = 0.5$, mechanism-identification accuracy is 94.32% for MECHVAR, 91.12% for confirmation-first, 94.08% for EIG, and 75.28% for random. MECHVAR's 3.20-point advantage over confirmation-first has a 95% CI of $[1.81, 4.59]$ (raw Wilcoxon $p = 2.18 \times 10^{-4}$; Holm-adjusted $p = 0.00196$; exact sign-test $p = 0.00434$). Its 0.24-point difference from EIG remains statistically unresolved (raw $p = 0.714$; Holm-adjusted $p = 1.0$).

When the true likelihood noise is tripled, MECHVAR reaches 94.72% accuracy and confirmation-first reaches 92.08%. The 2.64-point difference has a 95% CI of $[1.26, 4.02]$ (raw Wilcoxon $p = 9.99 \times 10^{-4}$; Holm-adjusted $p = 0.00800$; exact sign-test $p = 0.00149$). EIG reaches 94.24%, leaving an unresolved 0.48-point difference from MECHVAR (raw $p = 0.154$; Holm-adjusted $p = 1.0$). Table 1 summarizes these comparisons.

The structured-error sweep examines forecast errors beyond IID prediction noise. MECHVAR has the highest mean accuracy under probe-shared bias and low-rank correlated error, while EIG is slightly higher under mechanism-shared bias and heavy-tailed errors. These are exploratory numerical comparisons; Appendix D.1 reports the full results.

4.4 Budget, Scale, and Acquisition Cost

With $\sigma_{\text{pred}} = 0.5$, MECHVAR's mechanism-identification accuracy rises from 61.0% at $B = 1$ to 94.1% at $B = 4$ and 96.7% at $B = 8$. At $B = 4$, its 94.1% accuracy compares with 90.8% for confirmation-first (paired Wilcoxon $p = 0.00134$) and 94.4% for EIG. Under 3× likelihood noise,

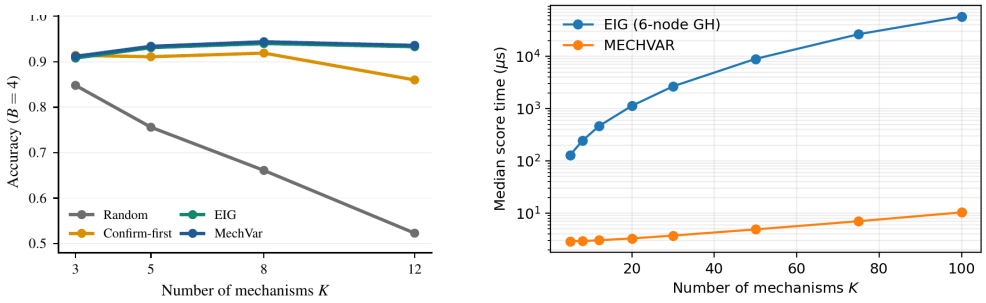


Figure 3: Left: mechanism-set scaling at fixed $B = 4$. Right: freshly rerun single-thread acquisition-score timing with cached six-node Gauss–Hermite nodes. Runtime is environment-specific; the widening gap reflects the different online scoring costs as K grows.

MECHVAR reaches 58.3%, 91.9%, and 95.9% at the same budgets. In these sweeps, informed policies become more similar as the budget grows; the disadvantage of confirmation-first is more apparent at low-to-moderate budgets.

We next vary the number of mechanisms with $E = 2K$ and $B = 4$. At $K = 12$, $E = 24$, and $\sigma_{\text{pred}} = 0.5$, accuracy is 93.6% for MECHVAR, 93.3% for EIG, 86.0% for confirmation-first, and 52.3% for random. The exploratory comparison gives a 7.6-point MECHVAR advantage over confirmation-first ($p = 1.84 \times 10^{-4}$); its difference from EIG remains unresolved.

We measure acquisition cost with one Python process and one BLAS thread on an AMD EPYC 9V74 virtual CPU, using Python 3.13.5 and NumPy 2.3.5 (Appendix F). At $K = 5$, $E = 10$, median full-library scoring takes 2.90 μs for MECHVAR and 129.41 μs for six-node EIG. At $K = 100$, $E = 200$, the medians are 10.36 μs and 57.69 ms, respectively, a 5,568$\times$ ratio on this machine. These timings cover score computation only and exclude model fitting and probe execution.

4.5 Held-Out Digits-Shortcut Validation

To complement the synthetic audit, we evaluate an executable closed loop on scikit-learn Digits (Pedregosa et al., 2011). Six hidden mechanism classes vary the strength of the digit signal and two planted binary shortcuts. Seven diagnostic interventions flip, randomize, or remove either shortcut, or permute the original image features at test time. We estimate response templates from 24 calibration seeds and evaluate on 25 disjoint held-out seeds, each containing all six mechanism classes. An earlier pilot seed range is excluded from inferential analyses.

Each task starts with a native logistic-regression fit. The policy then selects an unspent intervention, and the executor performs or retrieves its measured fit. Only the selected intervention's held-out accuracy is revealed before updating the posterior. Score-greedy uses the same posterior and calibration templates as the discrimination policies but chooses the intervention with the highest posterior-expected benchmark accuracy.

Over budgets 0–4, normalized mechanism-identification AUC is 0.8975 for MECHVAR and 0.7825 for score-greedy. The difference of $+0.1150$ has Wilcoxon $p = 8.66 \times 10^{-5}$ and Holm-adjusted $p = 2.60 \times 10^{-4}$. MECHVAR exceeds confirmation-first by $+0.0217$ (Holm-adjusted $p = 0.0162$) and random by $+0.0325$ (Holm-adjusted $p = 8.00 \times 10^{-5}$). EIG is numerically higher than MECHVAR by 0.0117, with the paired difference unresolved ($p = 0.0701$). At $B = 1$, mechanism-identification accuracy is 0.8867 for MECHVAR and 0.6200 for score-greedy; MECHVAR and EIG both reach 1.0000 by $B = 2$.

We also run a leave-one-seed-out replay on five measured intervention-response tables. Macro normalized AUC is 0.6604 for MECHVAR, 0.6601 for EIG, 0.6586 for confirmation-first, and 0.6260 for random. Because the profile labels and intervention library are fixed before replay, this experiment is a diagnostic of identification within that library. Appendix H gives the protocol and per-study results.

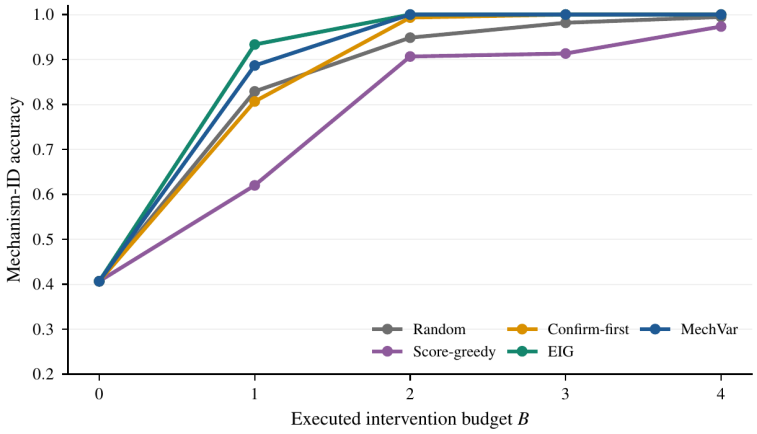


Figure 4: Held-out Digits-Shortcut loop. Score-greedy optimizes expected benchmark accuracy yet identifies the hidden mechanism more slowly; MECHVAR and EIG both reach perfect mechanism identification by budget $B = 2$.

5 Discussion

In Digits, score-greedy uses the same posterior and templates as the discrimination policies but selects for expected accuracy, identifying the planted mechanism more slowly. Confirmation-first is less effective in the stress audit for another reason: concentrating probes on the posterior favorite can waste a small budget when forecasts or likelihood assumptions are inaccurate.

MECHVAR and EIG are numerically close in the tested settings, which is compatible with the shared-Gaussian Box–Hill identity (Proposition 1) and the local EIG expansion (Proposition 2). The primary audit comparisons do not statistically distinguish them, and EIG has a small numerical AUC advantage on Digits; these observations do not establish equivalence or superiority in either direction. The practical distinction evaluated here is computational and diagnostic: MECHVAR uses matrix operations and exposes pairwise score contributions, whereas the EIG baseline evaluates a Gaussian-mixture integral for every candidate probe.

5.1 Limitations

The exact Box–Hill reduction depends on a shared predictive variance. In the unequal-scale check (Appendix E), MECHVAR and heteroscedastic Box–Hill select the same top probe in 81.2% of states at log-scale dispersion $\delta = 0.10$, falling to 25.7% at $\delta = 0.50$. The scalar response model is another restriction: multi-metric outcomes require an explicit multivariate likelihood or scalarization.

The planner also treats the finite mechanism set and response templates as given. It does not represent uncertainty in template estimation, recover from a missing true mechanism, generate new hypotheses, or optimize a non-myopic multi-step design. These are substantive boundaries for deployment in open-ended research agents.

The EIG comparator uses a numerical approximation. In a 1,000-state convergence diagnostic, six-node Gauss–Hermite quadrature matches the 50-node top probe 99.7% of the time, with a 95th-percentile per-state maximum score error of 7.99×10^{-4} nats. This synthetic check does not cover every posterior state encountered in the main experiments.

Finally, the stress audit is synthetic, Digits uses planted mechanisms, and the replay uses an existing intervention library. The evidence therefore supports finite-library discrimination rather than open-world scientific discovery. Runtime measurements cover score computation in one environment (Appendix F); end-to-end savings depend on the cost of fitting models and executing probes.

6 Conclusion

MECHVAR addresses a specific question: which available experiment best separates the mechanisms that remain plausible? Under a shared Gaussian predictive scale, the classical posterior-weighted Box–Hill objective reduces to a response-variance score with the same probe ordering, an additive mechanism-pair audit, and $O(KE)$ full-library evaluation. Across the stress audit and held-out Digits loop, MECHVAR is competitive with quadrature EIG while requiring much cheaper score computation in the measured implementation; score-greedy, by contrast, optimizes benchmark performance rather than identification. The result is a simple acquisition rule for settings in which the candidate mechanisms, probes, and shared-scale approximation are credible.

Reproducibility Statement

Section 4 specifies the evaluation design and inferential unit; Appendix G records the disjoint Digits calibration and held-out seed ranges; and Section 3 with Appendix A states the assumptions and proofs. The accompanying archive contains the manuscript source, figures, the numerical-check script used for quadrature convergence, heteroscedastic sensitivity, and score-only timing, the corresponding CSV outputs, and the recorded software environment. These archived numerical diagnostics are directly reproducible from `scripts/numerical_checks.py`.

AI Use Statement

Generative AI tools were used for language editing, code organization, and implementation support. The authors reviewed AI-assisted material and verified the experimental definitions, numerical results, statistical analyses, and manuscript claims against the archived scripts and outputs. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with generative-AI assistance.

References

Box, G. E. P. and Hill, W. J. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Chan, J. S., Chowdhury, N., Jaffe, O., et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *ICLR*, 2025.

Chen, Z., Chen, S., Ning, Y., et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *ICLR*, 2025.

Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:3384–3395, 2021.

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

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *ICLR*, 2021.

Guo, D., Shamai, S., and Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Houlsby, N., Huszar, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. arXiv:1112.5745, 2011.

Huang, Y., Song, Z., Li, K., and Arora, S. InstaHide: Instance-hiding schemes for private distributed learning. *ICML*, 2020.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. MLAgentBench: Evaluating language agents on machine learning experimentation. *ICML*, 2024.

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

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:2408.06292, 2024.

Pedregosa, F., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Schmidgall, S., et al. Agent Laboratory: Using LLM agents as research assistants. *Findings of EMNLP*, 2025.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. *ICML*, 2017.

Shi, C., Blei, D. M., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *NeurIPS*, 2019.

Sohn, K., et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020.

Wang, Z., Bai, F., Luo, Z., et al. FIRE-Bench: Evaluating agents on the rediscovery of scientific insights. *ICML*, 2026.

Yamada, Y., Lange, R. T., Lu, C., et al. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv:2504.08066, 2025.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. *NeurIPS*, 2018.

A Local EIG Expansion

Proof of Proposition 2. For fixed e, let the discrete random variable X take value $a_h(e)$ with probability $p_t(h)$. By construction $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = \mathrm{MDV}_t(e)$. After centering by $\mu_t(e)$ and scaling by σ_b, the observation is equivalent to the scalar Gaussian channel

$$Y' = \sqrt{\gamma}\, X + Z, \qquad \gamma = \frac{\lambda^2}{\sigma_b^2}, \qquad Z \sim \mathcal{N}(0,1).$$

Because Y' depends on H only through X, $I(H; Y_\lambda \mid e) = I(X; Y')$. The I–MMSE identity (Guo et al., 2005) gives

$$\frac{d}{d\gamma} I(X; \sqrt{\gamma}X + Z) = \frac{1}{2}\,\mathrm{mmse}(X \mid \sqrt{\gamma}X + Z).$$

At $\gamma = 0$, the observation is independent of X and the MMSE equals $\mathrm{Var}(X) = \mathrm{MDV}_t(e)$. Since X has finite support, the MMSE is smooth in a neighborhood of zero, so integration yields

$$I(X; Y') = \frac{\gamma}{2}\mathrm{MDV}_t(e) + O(\gamma^2) = \frac{\lambda^2}{2\sigma_b^2}\mathrm{MDV}_t(e) + O(\lambda^4).$$

This proves Equation (19). If two probes have a nonzero MDV gap, the $O(\lambda^4)$ remainder cannot reverse the leading $O(\lambda^2)$ ordering for all sufficiently small λ. □

B Worked Pairwise Audit Example

The following calculation is an illustrative audit example, separate from the empirical evaluation. Let the current posterior over three mechanisms be $p = (0.50, 0.30, 0.20)$ and let a selected probe have predicted responses $m = (0.15, 0.75, 0.40)$. Then $\mu = 0.38$ and $\mathrm{MDV} = 0.0676$. The pairwise decomposition is:

Table 2: Illustrative audit of one probe. The three pair contributions sum exactly to the MDV score.

Mechanism pair	$c_{hh'}(e)$	Share of MDV
(h_1, h_2)	0.05400	79.9%
(h_1, h_3)	0.00625	9.2%
(h_2, h_3)	0.00735	10.9%

Thus the audit identifies (h_1, h_2) as the pair responsible for roughly four fifths of the score, making the reason for selecting the probe easy to inspect.

C Complete Diagnostic Case Summary

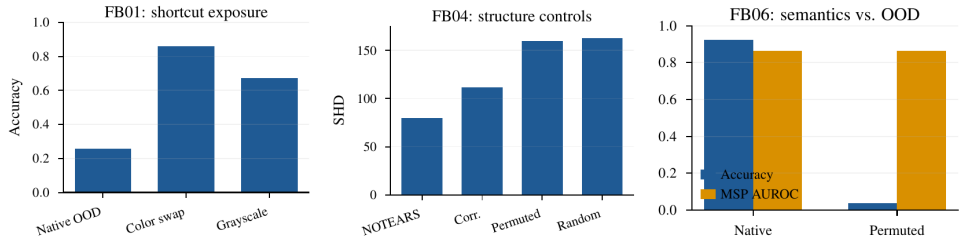


Figure 5: Representative diagnostics. Left: shortcut exposure. Middle: structural controls. Right: semantic class permutation leaves MSP OOD discrimination nearly unchanged.

The diagnostic implementations draw on the ColoredMNIST/DomainBed setup (Gulrajani and Lopez-Paz, 2021), FixMatch (Sohn et al., 2020), DragonNet- and TARNet-style treatment-effect models (Shi et al., 2019; Shalit et al., 2017), NOTEARS (Zheng et al., 2018), and maximum-softmax OOD scoring (Hendrycks and Gimpel, 2017). FB02 is a local protocol-matched negative result and differs in protocol from the corresponding external setup. FB03 distinguishes component-level evidence from aggregate model ranking. FB05 is an InstaHide-inspired local proxy (Huang et al., 2020); its conclusion is restricted to the attack families evaluated here.

Table 3: Six diagnostic studies used to motivate the mechanism-identification objective. Planner performance is evaluated separately in the main experiments.

ID	Domain	Key diagnostic result	Interpretation
FB01	ColoredMNIST	OOD accuracy $0.2575 \to 0.8612$ under color swap; $\Delta = +0.6028$ (95% CI $[0.5588, 0.6469]$).	Shortcut dependence supported
FB02	CIFAR-100 SSL	Supervised-200 0.0493 vs. FixMatch 0.0225; difference -0.0268 (95% CI $[-0.0677, 0.0140]$).	No support under this protocol
FB03	IHDP causal inference	Targeted regularization improves DragonNet PEHE by 0.1346, while mean PEHE still favors TARNet (1.2581 vs. 1.3852).	Mixed component-level evidence
FB04	50-node causal discovery	SHD: NOTEARS 79.7; correlation 111.7; permuted 160.0; random 162.7.	Structural sensitivity supported
FB05	Poisoned MNIST	Harmonic clean/robust gain: noise-mix $+0.0122$ (95% CI $[0.0027, 0.0217]$); FGSM $+0.0475$; trigger gain approximately zero.	Partial robustness support
FB06	CIFAR-10→SVHN OOD	Accuracy $0.9227 \to 0.0385$ under class permutation while $\lvert\Delta\text{MSP-AUROC}\rvert \leq 8.62 \times 10^{-7}$.	Semantic/OOD dissociation

D Robustness and Statistical Details

D.1 Structured Forecast Misspecification

To complement the IID prediction-noise sweep in the main text, we evaluate five structured forecast-error models at scale 0.5 with $K = 5$, $E = 10$, and $B = 4$.

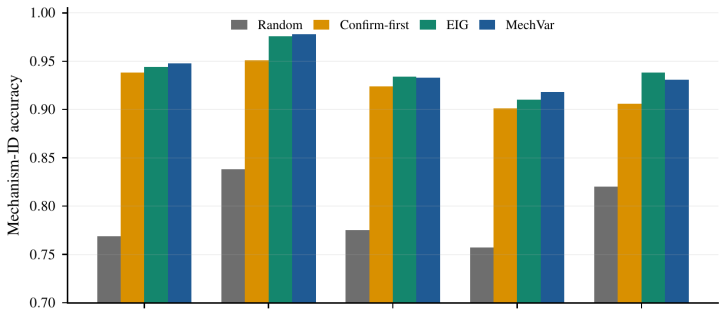


Figure 6: Mechanism-identification accuracy under five structured forecast-error models at scale 0.5, with $K = 5$, $E = 10$, and $B = 4$.

Table 4: Mean mechanism-identification accuracy across 20 paired blocks under five structured forecast-misspecification models. This sweep is exploratory; no confirmatory pairwise inference is attached to the numerical ordering.

Error structure	Random	Confirm-first	EIG	MECHVAR
IID Gaussian	.769	.938	.944	.948
Probe-shared bias	.838	.951	.976	.978
Mechanism-shared bias	.775	.924	.934	.933
Low-rank correlated	.757	.901	.910	.918
Heavy-tail t_3	.820	.906	.938	.931

The ordering is not uniform across error geometries: MECHVAR is highest under IID Gaussian, probe-shared bias, and low-rank correlated error, whereas EIG is slightly higher under mechanism-shared bias and heavy-tailed error.

D.2 Multiplicity Control

For the robust audit, the five prediction-misspecification levels and four non-duplicate likelihood-misspecification levels form a nine-condition family for each comparator. Holm-adjusted Wilcoxon p-values for MECHVAR versus confirmation-first are 0.228, 0.228, 0.497, 0.00196, and 0.497 for $\sigma_{\text{pred}} = 0, 0.1, 0.25, 0.5, 1.0$, respectively, and 0.497, 0.152, 0.0276, and 0.00800 for true-noise multipliers $0.5, 1.5, 2, 3$, respectively. All corresponding Holm-adjusted p-values for the MECHVAR–EIG family equal 1.0.

E Numerical Diagnostics for EIG and Model Assumptions

E.1 Gauss–Hermite Convergence

We evaluated quadrature order on 1,000 independently generated planner states with $K = 5$, $E = 10$, $\sigma_b = 0.22$, posterior draws $p \sim \text{Dirichlet}(2, \ldots, 2)$, and response templates sampled uniformly from $[0.15, 0.85]$. For every state we computed EIG with 6, 20, and 50 Gauss–Hermite nodes; the 50-node result serves as the numerical reference. Table 5 reports probe-ranking agreement and the maximum absolute score error within each state.

Table 5: Gauss–Hermite convergence over 1,000 synthetic planner states. Score errors are in nats and are measured against 50-node quadrature.

Nodes	Top-probe agreement	Median Spearman	Median max error	95th-pct. max error
6	0.997	1.000	3.84×10^{-4}	7.99×10^{-4}
20	1.000	1.000	3.62×10^{-7}	3.16×10^{-6}

The largest six-node score discrepancy observed in this diagnostic was 1.99×10^{-3} nats. The 20-node rule matched the 50-node top probe in all 1,000 states.

E.2 Heteroscedastic Sensitivity

To quantify the shared-scale boundary, we generated 2,000 planner states for each dispersion level using the same K, E, posterior, and response-template sampling scheme. Mechanism–probe scales were drawn as

$$\sigma_h(e) = 0.22 \exp(\delta z_{he}), \qquad z_{he} \sim \mathcal{N}(0, 1),$$

with values clipped to $[0.05, 0.80]$. We compare the MECHVAR ranking, which uses the response means, with the exact posterior-weighted ordered-pair KL score for heteroscedastic Gaussians. At $\delta = 0$ the rankings coincide by Proposition 1; disagreement grows as scale variation increases.

Table 6: Synthetic sensitivity to mechanism–probe-dependent predictive scales. Relative BH loss is the fractional gap between the optimal heteroscedastic Box–Hill score and the score attained by the MECHVAR-selected probe.

δ	Top-probe agreement	Median Spearman	Median BH loss	95th-pct. BH loss
0.00	1.000	1.000	0.000	0.000
0.10	0.812	0.976	0.000	0.134
0.25	0.560	0.855	0.000	0.425
0.50	0.257	0.539	0.359	0.796

These numbers make the model boundary concrete: mild scale variation often preserves the broad ranking, while strong heteroscedasticity can change the selected experiment materially.

F Extended Score-Runtime Benchmark

We reran the score-only benchmark on an AMD EPYC 9V74 80-Core Processor exposed through a KVM virtual machine (five single-thread vCPUs visible to the container), using Python 3.13.5, NumPy 2.3.5, and SciPy 1.17.0. One Python process was used and `OMP_NUM_THREADS`, `OPENBLAS_NUM_THREADS`, `MKL_NUM_THREADS`, and `NUMEXPR_NUM_THREADS` were each fixed to one. Six-node Gauss–Hermite nodes and weights were cached outside the timed region. The EIG routine loops over the latent mechanism while vectorizing over probes, quadrature nodes, and mixture components; no JIT compiler or GPU is used. For each K, we generated five independent posterior–template instances with $E = 2K$; reported values are medians of the per-instance median call times after warm-up.

Table 7: Single-thread acquisition-score timing. EIG uses six-node Gauss–Hermite quadrature; values exclude model fitting and probe execution.

K	E	MECHVAR (μs)	EIG (μs)	EIG/MECHVAR
5	10	2.90	129.41	44.6×
8	16	2.92	243.28	83.2×
12	24	3.03	463.51	152.8×
20	40	3.29	1134.53	345.4×
30	60	3.73	2674.41	718.0×
50	100	4.89	8958.05	1831.0×
75	150	7.01	26527.21	3784.2×
100	200	10.36	57687.03	5568.0×

The accompanying script records the same environment and reproduces the CSV used for this table and Figure 3. The ratio is a wall-clock measurement of these implementations on this machine, not a hardware-independent complexity constant.

G Online Digits-Shortcut Details

The final six-mechanism evaluation uses held-out seeds 3100–3124, while response templates are estimated exclusively from calibration seeds 1100–1123. The earlier four-mechanism pilot using seeds 2100–2124 is excluded from all inferential results reported in the paper. A shared Gaussian belief scale of $\sigma_b = 0.04$ is estimated from calibration residuals with a conservative floor. Each evaluation block contains all six mechanism classes, and random-policy results are averaged over 100 trajectories.

Table 8: Held-out Digits-Shortcut mechanism-identification accuracy across budgets. Budget 0 denotes the posterior after the native run and before any diagnostic intervention is executed.

Policy	$B = 0$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
Random	.4067	.8290	.9486	.9817	.9946
Score-greedy	.4067	.6200	.9067	.9133	.9733
Confirm-first	.4067	.8067	.9933	1.0000	1.0000
EIG	.4067	.9333	1.0000	1.0000	1.0000
MECHVAR	.4067	.8867	1.0000	1.0000	1.0000

Normalized budget-AUC is 0.8650 for random, 0.7825 for score-greedy, 0.8758 for confirmation-first, 0.9092 for EIG, and 0.8975 for MECHVAR. Across the 25 paired held-out blocks, the MECHVAR AUC difference is $+0.0325$ relative to random (Wilcoxon $p = 2.0 \times 10^{-5}$; Holm-adjusted $p = 8.0 \times 10^{-5}$), $+0.1150$ relative to score-greedy ($p = 8.66 \times 10^{-5}$; Holm-adjusted $p = 2.60 \times 10^{-4}$), $+0.0217$ relative to confirmation-first ($p = 0.00811$; Holm-adjusted $p = 0.0162$), and -0.0117 relative to EIG ($p = 0.0701$; statistically unresolved).

H RETROSPECTIVE EMPIRICAL REPLAY

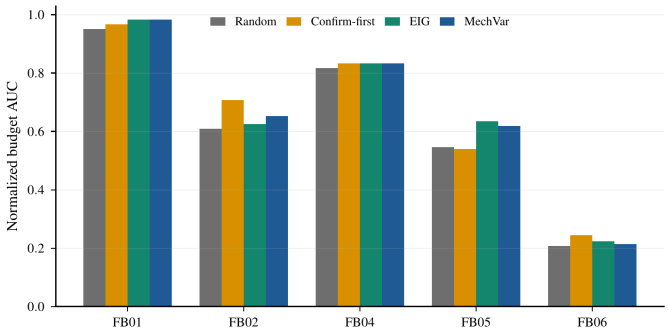


Figure 7: Normalized budget-AUC for leave-one-seed-out identification of empirical intervention-response profiles using context-derived priors.

Response profiles and standardization statistics are estimated in a leave-one-seed-out manner, after which held-out response dimensions are revealed sequentially. Macro normalized budget-AUC across FB01, FB02, FB04, FB05, and FB06 is 0.6604 for MECHVAR, 0.6601 for EIG, 0.6586 for confirmation-first, and 0.6260 for random. The intervention library and profile labels are fixed before replay; the experiment therefore measures sequential identification within an existing library.