Hard to Find, Easy to Fool: Witness Rank for Budgeted Scientific Verification

Luca Mondonico* Stanford University
Albert Luo Vmax
Yi Cui Stanford University
Zhenan Bao Stanford University

Abstract

AI scientists and autonomous design systems can rank candidates faster than experiments can evaluate them, yet standard benchmark metrics do not quantify the evidence for the single candidate ultimately nominated. We ask how many reference measurements are needed to overturn that nomination. On a declared finite pool, the margin-aware *witness rank* R_Δ is the position, under a policy that cannot access unobserved measurements, of the first candidate that beats the nomination by the margin Δ. Across 230 audits on four fully measured protein and DNA landscapes, including two published offline optimizers, median witness ranks conditional on a witness existing are three to five, and structured policies expose witnesses far more often than random probing. Rank-shuffling locates much of this advantage in the ordering among the proxy's highest-ranked candidates. Oracle-controlled rankings with Spearman correlation near 0.95 among the top candidates by measured performance retain conditional medians of two to four; higher margins mainly reduce whether a qualifying witness exists. In 7,300 simulated matched-budget campaigns with exact measurements, returning the best measured candidate, including the nomination, cannot lower the nomination's score. For noisy measurements, confidence bounds limit false claims of improvement; when a noiseless shortlist finds no witness, an independent random sample bounds the fraction of witnesses in the pool. Witness rank reports post-selection reliability in experimental-budget units and gives a rule for replacing or retaining a proxy-guided recommendation.

1 Introduction

AI scientists and autonomous laboratories can generate or rank many more hypotheses and designs than can receive trusted evaluation, meaning assessment by a reference experiment or evaluator [1–3]. In scientific design, the resulting pipeline often fits a proxy to scarce measurements, optimizes it over an admissible library, and sends one top-ranked candidate to experiment [4, 5]. Proxy queries are cheap, whereas synthesis, characterization, or biological testing is not. The resulting verification bottleneck is therefore a decision problem: which few candidates should receive trusted evaluation, and what evidence is sufficient to act? It is also a statistical problem because optimization selects an extreme prediction, precisely where average held-out error is least informative [6, 7].

A poor outcome can reflect two different failures. The desired candidate may be *hard to find*: a limited optimizer fails even when it searches the trusted landscape. Or the learned objective may be *easy to fool*: the optimizer succeeds on the proxy, but the proxy ranks an inferior candidate highest. More computation can address the first failure but cannot correct the second. Existing metrics measure prediction, optimization, or end-to-end performance, but none reports the trusted cost of challenging the specific recommendation about to be acted upon.

*Correspondence: lmondo@stanford.edu

Accepted at the AI for Science workshop (NeurIPS 2026).

Full verification is usually impossible, but falsification can be cheap. To prove that a nomination is not the best candidate in a declared pool, one does not need to identify the pool optimum; one needs a single counterexample. We call such a candidate a *witness*. Once an ordering has been fixed without access to unobserved trusted labels, the rank of the first witness is the number of trusted evaluations required to overturn the nomination. A margin-aware rank requires the witness to improve on the nomination by more than Δ, separating a nominal counterexample from a scientifically meaningful one.

This paper studies the empirical geometry of those witnesses. Across four fully measured landscapes, the declared selection pools contain tens to hundreds of thousands of candidates, but meaningful witnesses usually occur within the first few positions of a proxy-guided ordering. The contrast is the central result: *in the measured landscapes, candidate-pool size is large while nomination-level audit cost is small.* This is not a theorem about finite pools. We give constructed cases where witness rank grows with pool size, and we report both witness existence and conditional rank so that high-quality nominations are not mischaracterized as difficult audits.

We formulate nomination-level falsification and define margin-aware witness rank, an operational measure in trusted-evaluation units rather than another aggregate proxy score. The empirical study audits four measured protein and DNA landscapes and two published offline optimizers. Across 230 audits and selection pools of 45,570–152,929 candidates, conditional median witness rank is three to five. We test alternative explanations with a rank-shuffle ablation, oracle-controlled quality ladder, margin sweep, and scaling study. We also use the audit to choose the delivery, the final candidate returned, retaining the nomination as the incumbent (the fallback choice). We extend the audit's certificate, a lower bound on the nomination's shortfall from the pool optimum, to noisy evaluations and bound witness prevalence when a noiseless shortlist finds none.

The protocol is to freeze the nomination, candidate pool, margin, and acquisition policy; evaluate the nomination and a budgeted shortlist; and replace the nomination only when a measured candidate is certified better. If none appears, retain the nomination and report that the available budget did not falsify it. An optional independent random sub-budget can then bound how prevalent remaining witnesses may be. The audit does not certify optimality, prescribe a new optimizer, or verify open-ended claims outside the declared pool.

2 Related work

Our contribution sits at the intersection of AI-enabled scientific discovery, offline model-based optimization, decision-focused prediction, and best-arm identification. These areas explain how scientific candidates are generated and why aggregate predictive quality may not determine the reliability of the particular candidate ultimately selected.

AI scientists and scarce verification. Agentic systems increasingly generate research ideas, experimental plans, and ranked hypotheses at a scale that makes exhaustive human or laboratory verification impossible [1, 2]. Autonomous materials platforms expose the same asymmetry between inexpensive model queries and scarce physical measurements [5]. Our scope is narrower than verification of an open-ended scientific claim: we study the commitment point at which such a system nominates one member of a declared design pool for trusted evaluation.

Offline design and proxy exploitation. When trusted evaluation is expensive, offline model-based optimization learns from a fixed table of designs and scores. Existing methods invert desired scores back to designs [8], penalize unsupported predictions [9], constrain optimized designs through forward and backward mappings [10], or replace regression targets with order information [11]. Design-Bench provides standardized scientific and engineering tasks for comparing such pipelines [4]. These approaches primarily improve the proxy or the optimizer. Witness rank instead evaluates the realized nomination after that pipeline has run and asks what trusted evidence would overturn it.

Decision quality after optimization. The optimizer's curse explains why maximizing noisy estimates produces systematic post-decision disappointment [6]. Decision-focused learning and predict-then-optimize methods train models for downstream decisions rather than average prediction [12, 13], and ranking-based offline optimization directly targets order information [11]. Related work also

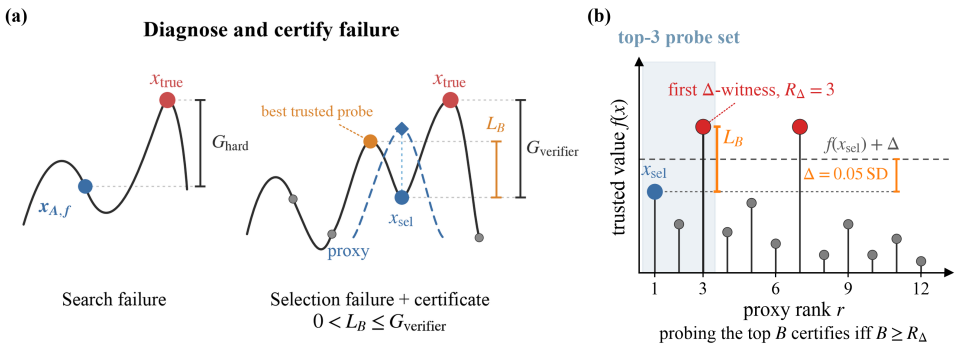


Figure 1: **Search and selection are different failure modes.** *(a)* When A searches the trusted score, stopping at $x_{A,f}$ leaves the search shortfall G_{hard}. Exact proxy optimization can instead nominate x_{sel}, leaving selection regret G_{verifier}; only this second failure is audited here. *(b)* In proxy-rank order, the first Δ-witness beats the nomination by more than Δ; its position is the trusted cost of falsification. In this schematic, SD denotes the standard deviation of trusted pool scores; Figure 2 reports the measured distribution.

shows that overoptimizing learned rewards can reduce trusted performance [7], that in-silico benchmark rankings depend on the evaluation oracle [14], and that decision learnability need not follow predictive accuracy [15]. These results characterize a method or pipeline in aggregate. Our audit conditions on its realized decision and measures the trusted cost of refuting that choice.

Falsification rather than identification. Best-arm identification seeks the optimum and its complexity depends on gaps across the candidate population [16, 17]. We ask only whether one already-selected candidate can be beaten, an asymmetry that can make a much smaller audit sufficient. Fully measured landscapes are essential for the retrospective study because they reveal both whether a witness exists and where it falls in a label-free ordering. We use measured protein and DNA landscapes from GB1, TrpB, TF-Bind-8, and avGFP [18–22]. These datasets provide exact finite-pool counterfactuals; they do not turn the pool optimum into a claim about the physical global optimum.

3 Nomination-level auditing

3.1 Setting and the two failures

Let $\mathcal{X}$ be a finite declared pool, $f : \mathcal{X} \to \mathbb{R}$ the trusted score, and $\hat{f} : \mathcal{X} \to \mathbb{R}$ a proxy fit without labels from the selection pool, with higher scores preferred. Write

$$x_{\text{true}} \in \arg\max_{x \in \mathcal{X}} f(x) \qquad x_{\text{sel}} \in \arg\max_{x \in \mathcal{X}} \hat{f}(x)$$

for the trusted pool optimum and proxy nomination. If running a budgeted algorithm A directly on the trusted score returns $x_{A,f}$, define

$$G_{\text{hard}}(A) = f(x_{\text{true}}) - f(x_{A,f}) \qquad G_{\text{verifier}} = f(x_{\text{true}}) - f(x_{\text{sel}}). \tag{1}$$

Here, G_{hard} is the trusted-space search shortfall: it measures how far a budgeted optimizer stops from the trusted pool optimum. In contrast, G_{verifier} is the trusted selection regret caused by the proxy nomination. More search can reduce G_{hard}, whereas additional proxy optimization cannot remove G_{verifier}; witness rank measures the cost of exposing this second failure (Figure 1(a)).

3.2 Witness rank and its budget interpretation

For a set $S_B \subseteq \mathcal{X}$ of at most B evaluated candidates including x_{sel}, the audit certificate is the largest measured improvement over the nomination,

$$L_B = \max_{x \in S_B} f(x) - f(x_{\text{sel}}). \tag{2}$$

Then $0 \leq L_B \leq G_{\text{verifier}}$: a positive L_B proves the nomination suboptimal, while $L_B = 0$ is inconclusive if candidates remain unmeasured. For a margin $\Delta \geq 0$ and a fixed ordering $x_{(1)}, \ldots, x_{(|\mathcal{X}|)}$ that does not use selection-pool labels, with $x_{(1)} = x_{\text{sel}}$, define

$$R_\Delta = \min\{r : f(x_{(r)}) > f(x_{\text{sel}}) + \Delta\} \qquad R_\Delta = \infty \text{ if no such candidate exists} \tag{3}$$

For the first B candidates, $L_B > \Delta$ if and only if $R_\Delta \leq B$ (Figure 1(b)). An adaptive policy instead chooses each candidate using offline data, the proxy, fixed randomization, and previously acquired trusted labels, never unobserved selection-pool labels. Its witness rank counts evaluations through the first qualifying witness, including the nomination; the certificate remains valid.

Our primary margin is $\Delta = 0.05$ standard deviations of trusted selection-pool scores, allowing comparisons across retrospective landscapes; the sensitivity analysis tests margins from 0 to 0.5 standard deviations. A prospective campaign must choose the required improvement from application requirements and freeze the margin in physical units, or derive its scale from historical or training measurements, before reading any selection-pool outcome. We report $\Pr(R_\Delta < \infty)$, the probability that a qualifying witness exists, separately from the conditional rank distribution so that a high-quality nomination with no meaningful competitor is not confused with an expensive-to-audit failure.

3.3 What rank correlation does not determine

Top-tail Spearman correlation, ρ_{tail}, measures agreement between proxy and trusted rankings among the candidates with the highest trusted scores, but does not determine which candidates the proxy places ahead of that set. Consequently, it cannot by itself determine the audit budget.

Proposition 1 (Correlation does not identify audit cost)**.** *For a pool of N candidates, an integer k with $2 \leq k < N - 1$, and any $\Delta > 0$, there exist a trusted score and two proxy rankings with the same suboptimal nomination and Spearman correlation one within the trusted top k, for which the first Δ-witness has rank 2 in one ordering and $N - k + 1$ in the other.*

Appendix A gives the construction. The proposition motivates the quality ladder in Section 5.3: we report nomination-level failure incidence and witness rank alongside correlation rather than treating correlation as a surrogate for them. Both quantities are therefore needed to characterize the reliability of a realized nomination.

3.4 Selecting the final candidate with exact or noisy evaluation

With exact evaluations, return the best measured candidate, $x_{\text{safe}} = \arg\max_{x \in S_B} f(x)$, using the same measurements. The nomination remains eligible, so $f(x_{\text{safe}}) - f(x_{\text{sel}}) = L_B \geq 0$ and the remaining regret is $G_{\text{verifier}} - L_B$. If no probe improves on the nomination, retain it. The returned candidate therefore performs at least as well as the nomination, but may still fall short of the application's requirements.

With noisy evaluations, a candidate is a certified Δ-witness only if a simultaneous lower confidence bound on $f(x) - f(x_{\text{sel}})$ exceeds Δ; a positive bound suffices for replacement. For a probe set and repeat counts fixed before measurement, independent zero-mean Gaussian noise with known variance permits a certificate L_B^{Gauss} satisfying $\Pr(L_B^{\text{Gauss}} \leq G_{\text{verifier}}) \geq 1 - \delta$ by a union bound, where δ is the allowed error probability, distinct from the improvement margin Δ. Under these assumptions, replacing the nomination only when a candidate's bound is positive guarantees a returned true score at least as high as the nomination's with probability at least $1 - \delta$. Appendix H gives the formula and reports six combinations of measurement error and confidence bounds.

3.5 The zero-certificate case

If candidates remain unmeasured, a zero certificate does not establish optimality: the structured shortlist may simply have missed existing witnesses. We therefore optionally evaluate an independent uniform random sample from the pool. If that sample finds few or no witnesses, standard finite-population sampling gives a one-sided upper confidence bound on witness prevalence, the fraction of pool candidates that are witnesses. This statement does not depend on the proxy or on any assumed relationship between f and $\hat{f}$. The random evaluations are additional to the structured audit budget and bound witness prevalence rather than proving the nomination optimal. Appendix I gives the exact calculation and its validation.

4 Audited systems and protocol

We use four fully measured landscapes: GB1 [18], TrpB [19], TF-Bind-8 [20, 4], and avGFP [21, 22]. After removing entries with estimated rather than measured scores, resolving duplicates and conflicts, and making disjoint training, validation, and selection splits, the declared selection pools contain 143,781, 152,929, 59,392, and 45,570 candidates, respectively. The study is retrospective because estimating the full witness-rank distribution requires measured values across each pool.

Four proxy families span common modeling choices: additive or pairwise ridge regression, Extra-Trees ensembles of randomized trees on token features [23, 24], a pairwise ranking model, and a conservative objective surrogate [9]. These give 200 margin-aware audits. We additionally reproduce 30 runs from two offline optimizers using their official components: a ranking pipeline [11, 25] and Conservative Objective Models [9]. All candidates used in the finite-pool audit must belong to the declared pool. If an optimizer proposes a candidate outside that pool, we exclude that proposal from the finite-pool evaluation rather than mapping it to a different valid candidate. Appendix B reports dataset details, and Appendix C reports the full optimizer schedules.

We evaluate prefix budgets $B \in \{8, 16, 32, 64, 128, 256\}$, always counting the nomination. The primary margin is $\Delta = 0.05$ standard deviations of the selection pool, a retrospective normalization that no policy observes. Certification recall, written Recall@B, is the fraction of all audits that find a qualifying witness within B trusted evaluations; audits without a qualifying witness count as uncertified. Intervals use paired percentile bootstrap resampling of data splits within a landscape. They show how results vary when the same measured dataset is divided into training, validation, and selection sets in different ways. They do not measure variation across repeated biological experiments or establish whether the findings apply to other scientific domains.

We also retrospectively simulate 7,300 matched-budget, multi-round campaigns using proxy prefixes, Thompson sampling, batch Bayesian optimization (BO), witness-local probing, and uniform random probing. Every campaign retains the original nomination in its evaluated set. Policies never observe the trusted selection-pool scale or labels outside their acquired set. Appendix F defines the information available to each policy, gives the schedules, and reports the acquisition and budget checks.

To test whether the result is merely an artifact of moderate-quality proxies, we construct a controlled quality ladder. Normalized trusted ranks are mixed with independent noise; mixing strengths are calibrated on validation data for five correlation targets and evaluated on the disjoint selection pool. Because these rankings use measured scores that would be unknown before an experiment, this oracle study tests the effect of ranking quality rather than a prediction method available in a prospective campaign. We also shuffle the full ranking, shuffle within each tenth of the ranking, or shuffle only the top one percent to test which parts of the ordering help locate witnesses. Appendix E gives the complete quality-control construction.

5 Results

We measure nomination-level audit cost and test whether pool size or proxy quality explains the advantage over random probing. We then ask whether the same measurements can improve the final recommendation while retaining the nomination when no better candidate is found.

5.1 Large candidate pools can have small audit costs

As shown in Table 1, selection failure appears in every audited landscape and proxy family, although the nominations themselves lie at high trusted percentiles. Across candidate pools of 45,570–152,929 designs, median margin-aware witness rank is only three to five. Unconditional certification recall is 78.6–97.1% at 16 trusted evaluations and 90.0–100% at 64. One TF-Bind-8 audit and one avGFP audit have no witness at the headline margin.

In every audited landscape and proxy family, the proxy maximum is strictly worse than the trusted pool optimum. Nevertheless, the nominations are strong: their median trusted percentile is 0.926 on avGFP and at least 0.996 on the other landscapes, and only five of 230 reach the trusted top five. Selection regret varies across proxy families within each landscape, while the worst proxy family changes across landscapes. The phenomenon is therefore not confined to one unusually poor model or one landscape; Table 4 gives results for every combination of landscape and model.

Table 1: **A small trusted budget certifies most audits at the stated margin.** All quantities use $\Delta = 0.05$ standard deviations; recall is unconditional, and median ranks are conditional on a witness existing. The 200 proxy-family audits and 30 published-optimizer reproductions sum to 230.

Landscape	Pool size	Audits	$\Pr(R_\Delta < \infty)$	Median R_Δ	Recall@16 / @64
TF-Bind-8	59,392	70	0.986	5	0.786 / 0.943
TrpB	152,929	70	1.000	3	0.971 / 1.000
GB1	143,781	70	1.000	3	0.914 / 0.971
avGFP	45,570	20	0.950	5	0.800 / 0.900

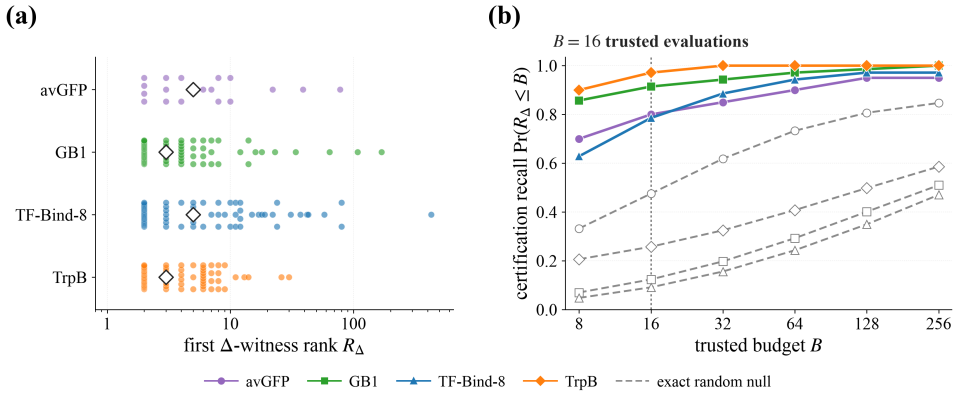


Figure 2: **Certification cost is typically small on measured landscapes.** *(a)* Finite margin-aware witness ranks from the 230 audits at $\Delta = 0.05$ SD, on a logarithmic axis; open diamonds show medians conditional on a witness existing. Audits without a Δ-witness have $R_\Delta = \infty$ and are not plotted. *(b)* Certification recall $\Pr(R_\Delta \leq B)$ as a function of trusted budget, with the exact finite-population random-order baseline shown dashed. At $B = 16$, the reported recall ratio is the value of each solid curve divided by its corresponding dashed baseline.

The contrast is not a rescaled form of pool size. Figure 2(a) shows a distribution concentrated at small ranks with a heavy upper tail. The exact random-order null in Figure 2(b) accounts for witness prevalence and finite pool size. At $B = 16$, certification recall under proxy ordering is $1.68\times$ the exact random-order baseline on avGFP, $8.60\times$ on TF-Bind-8, $7.39\times$ on GB1, and $3.77\times$ on TrpB. Structured orderings therefore do more than benefit from the mere existence of many witnesses.

5.2 Why the order of the highest-ranked candidates matters

To understand why proxy ordering finds witnesses more often than random ordering, we shuffle the rankings to test whether the advantage comes from placing promising candidates near the top or from their precise order within that group. A global shuffle reduces recall@16 to 0.467, 0.108, 0.072, and 0.208 on avGFP, GB1, TF-Bind-8, and TrpB, confirming that the learned ordering contains substantial information about where witnesses lie. Shuffling only the candidates within the proxy's top one percent keeps the same candidates in this group but changes their relative order. Recall then falls to 0.633 on GB1, 0.492 on TF-Bind-8, and 0.715 on TrpB, showing that the precise order of the highest-ranked candidates reduces audit cost on these landscapes. On avGFP, recall remains 0.807, indicating that selecting the right group of candidates matters more than their precise order. Thus the small witness ranks are not explained only by a positively correlated proxy concentrating good candidates near the top; on three landscapes, their ordering within that region is itself important.

5.3 Small conditional ranks persist under controlled high-quality orderings

Across 525 additional combinations of model and data split, conditional median witness ranks remain between two and six. Three model changes that do not require measurements from the selection pool, including a seven-model ensemble and fixed sequence representations from the ESM-2 protein language model [26], raise the best top-tail Spearman correlation only from 0.383 to 0.402. These

Table 2: **High-quality rankings retain small witness ranks.** Rows aggregate rankings with achieved selection-pool top-tail Spearman correlation at least 0.875. P90 denotes the 90th percentile of finite witness ranks. Recall is unconditional / conditional on a witness existing.

Landscape	Mean ρ_{tail}	$\Pr(R_\Delta < \infty)$	Median R_Δ	P90 R_Δ	Recall@16
avGFP	0.950	0.646	4	16.8	0.580 / 0.897
GB1	0.958	0.927	2	13.0	0.845 / 0.912
TF-Bind-8	0.948	0.468	4	15.0	0.435 / 0.930
TrpB	0.951	0.966	2	8.0	0.926 / 0.959

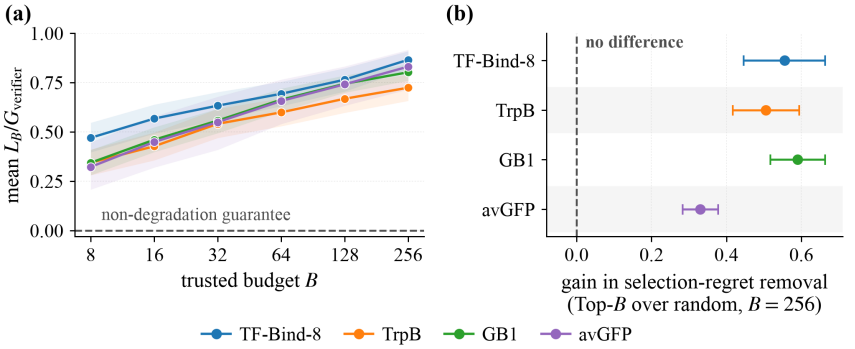


Figure 3: **The audit probes are also the repair.** *(a)* Mean fraction of selection regret removed, L_B/G_{verifier}, against trusted budget; ribbons show 95% split-bootstrap intervals. Because the nomination remains available, this fraction cannot fall below zero; the horizontal zero line marks that guarantee. Curve heights are empirical. *(b)* Paired prefix-minus-random removal difference at $B = 256$, excluding zero on all four landscapes.

models remain only moderately accurate at the top of the ranking, so we next test whether small witness ranks persist under substantially higher ranking quality. The controlled ladder tests that mechanism beyond the observed model ceiling. At achieved selection-pool top-tail correlation near 0.95, conditional median witness ranks remain two to four in all four landscapes (Table 2). Higher quality changes how often a meaningful failure exists: incidence falls to 0.468 on TF-Bind-8 and 0.646 on avGFP, while remaining above 0.92 on GB1 and TrpB. Conditional on a failure still existing, however, it does not become uniformly expensive to reveal.

The ladder does not show that a particular trainable model attains this quality. It tests only whether small conditional ranks must disappear as ranking quality improves. Proposition 1 also shows why the Spearman axis is not sufficient; the audit outcomes remain the quantities of interest.

Increasing Δ from 0.05 to 0.5 standard deviations lowers witness existence and unconditional recall most sharply on avGFP and TF-Bind-8. GB1 and TrpB retain a qualifying witness in every audit, with conditional median ranks at most six. The complete sweep is in Appendix D. These findings must be reported separately: a meaningful counterexample often exists, and, conditional on existence, it is usually near the head.

5.4 Using the audit measurements to improve the recommendation

A measured witness can also replace the nomination. For each prefix, we return the trusted best candidate measured so far while carrying the nomination as an incumbent. The improvement is exactly $L_B \geq 0$, so the delivery cannot be worse than the original recommendation.

An independent calculation reproduced all certificate values L_B and the trusted scores of the returned candidates across 248 failure instances. At $B = 256$, prefix probing removes a mean fraction 0.866 of selection regret on TF-Bind-8, 0.831 on avGFP, 0.803 on GB1, and 0.724 on TrpB. Paired improvements over random probing exclude zero on all four landscapes (Figure 3). The reported fractions measure the reduction in selection regret achieved in practice; the guarantee is simply that the final candidate performs at least as well as the nomination.

Retaining the nomination when choosing the best evaluated candidate provides this safeguard for both fixed proxy-ordered shortlists and policies that select new candidates using earlier measurements. Across 7,300 matched-budget campaigns, every structured policy certifies more often than uniform random probing in every landscape (Table 8 in Appendix F). The policy with the highest certification rate differs by landscape: prefix probing performs best on GB1, batch BO on TF-Bind-8, witness-local probing on avGFP, and Thompson sampling on TrpB. Overall, 82.86% of campaigns find a candidate with a strictly higher trusted score than the nomination; the remaining campaigns retain the nomination. Keeping the nomination as an option improves the final score in 1,243 campaigns compared with choosing only among newly evaluated candidates, eliminating every outcome below the nomination's score. Full policy and refitting comparisons appear in Appendix F.

Published optimizers exhibit the same pattern. At their full training schedules, auditing each learned model's exact in-pool proxy optimum gives median witness ranks 3, 2, and 4 for the ranking pipeline on TF-Bind-8, TrpB, and GB1, with recall one by $B = 32$ on all three. The conservative surrogate gives median ranks 7, 3, and 2 and top-8 recall 0.6, 1.0, and 0.6. In-pool reruns retain the pattern while exposing larger delivery ranks when search does not reach the proxy optimum; Appendix C reports the training schedules and checks whether proposed candidates belong to the declared pool.

Imperfect and absent certificates. Across the six noisy-evaluator settings, those satisfying the stated assumptions had low observed false-certification rates, whereas uncorrected systematic bias raised this rate to 0.214 (Appendix H). Separately, across 876,000 noiseless random draws, the exact residual bound was below one in 99.85% of cases. Coverage, measured as the fraction of draws whose upper bound was at least the true fraction of witnesses, met or exceeded the requested confidence level for every tested combination of landscape, budget, and confidence level (Appendix I).

6 Discussion

Witness rank asks a concrete experimental question: how many trusted measurements would overturn the recommendation about to be used? On four measured landscapes the answer is often a handful, because falsification needs one counterexample rather than identification of the optimum. The result remains empirical and policy-dependent: head shuffling increases audit cost, better rankings make failures rarer but not uniformly deeper, larger margins can eliminate witnesses, and finite pool size alone gives no small-budget guarantee (Appendix G). A campaign should therefore report its frozen nomination, policy, physical-unit margin, total trusted budget, and observed rank. With exact evaluations, a measured witness can replace the nomination with a higher-scoring candidate. With noisy measurements, replacement requires a positive simultaneous lower confidence bound on the improvement under the stated assumptions. If no candidate meets the replacement criterion, we retain the nomination without claiming that it is optimal.

7 Limitations

All audits are retrospective and assume a fixed finite pool, excluding searches that change the pool during the audit. Repeated splits neither provide independent biological replication nor establish generalization beyond the four measured datasets; deployable proxies have moderate tail correlation, and the high-quality ladder is oracle-only. The noise study uses simulated measurements, and its main formula assumes independent repeats and provides no guarantee under unknown, unbounded systematic bias.

8 Conclusion

Witness rank measures the trusted cost of falsifying a nomination under a policy that cannot read unobserved outcomes. On the measured landscapes, meaningful counterexamples often appear within three to five evaluations even when the pool exceeds 150,000 candidates; this is empirical, not a consequence of finiteness. Reporting witness existence, rank, margin, and total budget, then replacing the nomination only with a certified measured candidate, turns post-selection reliability into a protocol for fixed or adaptive scientific acquisition.

References

[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026.

[2] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026.

[3] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[4] Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-Bench: Benchmarks for data-driven offline model-based optimization. In *International Conference on Machine Learning*, pages 21658–21676. PMLR, 2022.

[5] Jurgis Ruža, Michael A Stolberg, Sawyer Cawthern, Jeremiah A Johnson, Yang Shao-Horn, and Rafael Gómez-Bombarelli. Autonomous discovery of polymer electrolyte formulations with warm-start batch Bayesian optimization. *Chemistry of Materials*, 37(19):7771–7777, 2025.

[6] James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[7] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pages 10835–10866. PMLR, 2023.

[8] Aviral Kumar and Sergey Levine. Model inversion networks for model-based optimization. *Advances in neural information processing systems*, 33:5126–5137, 2020.

[9] Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In *International Conference on Machine Learning*, pages 10358–10368. PMLR, 2021.

[10] Can Chen, Yingxue Zhang, Jie Fu, Xue Steve Liu, and Mark Coates. Bidirectional learning for offline infinite-width model-based optimization. *Advances in Neural Information Processing Systems*, 35:29454–29467, 2022.

[11] Rong-Xi Tan, Ke Xue, Shen-Huan Lyu, Haopu Shang, Yao Wang, Yaoyuan Wang, Sheng Fu, and Chao Qian. Offline model-based optimization by learning to rank. In *International Conference on Learning Representations*, 2025.

[12] Priya L. Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

[13] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management science*, 68(1):9–26, 2022.

[14] Shikha Surana, Nathan Grinsztajn, Timothy Atkinson, Paul Duckworth, and Thomas D Barrett. Overconfident oracles: Limitations of in silico sequence design benchmarking. *arXiv preprint arXiv:2502.17246*, 2025.

[15] Shen-Huan Lyu, Rong-Xi Tan, Ke Xue, Yi-Xiao He, Yu Huang, Qingfu Zhang, and Chao Qian. On the learnability of offline model-based optimization: A ranking perspective. *arXiv preprint arXiv:2603.04000*, 2026.

[16] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[17] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory*, pages 41–53, 2010.

[18] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016.

[19] Kadina E Johnston, Patrick J Almhjell, Ella J Watkins-Dulaney, Grace Liu, Nicholas J Porter, Jason Yang, and Frances H Arnold. A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences*, 121(32):e2400439121, 2024.

[20] Luis A Barrera, Anastasia Vedenko, Jesse V Kurland, Julia M Rogers, Stephen S Gisselbrecht, Elizabeth J Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, 351(6280):1450–1454, 2016.

[21] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533 (7603):397–401, 2016.

[22] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36:64331–64379, 2023.

[23] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[27] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

Appendix contents

The appendix collects the formal checks, experimental details, and robustness results that support the claims in the main text. Each section is cross-referenced where its evidence is used, and the table below summarizes its role in the paper.

Section	Contents	Role in the paper
A	Proofs and formal remarks	Certificate and delivery identities, and the correlation-separation construction
B	Datasets and protocol	Provenance, cleaning, split disjointness, checks of which candidates belong to the pool, and landscape–model results for all four landscapes
C	Published optimizers	Full training schedules, reference runs, and in-pool delivery accounting for two reproduced methods
D	Margin sensitivity	Witness existence, conditional ranks, and recall over the complete margin sweep
E	Model quality and ordering	Model comparisons, artificial rankings, and rank-shuffling tests
F	Repair and acquisition	Independent recomputation, incumbent-preserving delivery, and the matched-budget policy comparison
G	Candidate-pool scaling	Constructed regimes showing that small witness rank is not implied by a finite candidate pool
H	Noisy evaluators	Six combinations of measurement error and confidence bounds
I	Fraction of better candidates	Upper bounds on the fraction of better candidates from an independent random sample

Frequently used notation

Symbol	Meaning
$\mathcal{X},\ N = \lvert\mathcal{X}\rvert$	Fixed candidate pool and its size.
$f(x),\ \hat{f}(x)$	Trusted and proxy scores of candidate x; higher scores are preferred.
$x_{\mathrm{true}},\ x_{\mathrm{sel}}$	Best candidate under the trusted score and under the proxy score, respectively; x_{sel} is the nomination.
G_{verifier}	Selection regret of the nomination, $f(x_{\mathrm{true}}) - f(x_{\mathrm{sel}})$.
$B,\ S_B$	Evaluation budget and evaluated candidates, including the nomination, with $\lvert S_B\rvert \le B$.
x_{safe}	Best evaluated candidate under exact trusted scores.
L_B	Largest measured improvement over the nomination under exact evaluation; a lower bound on G_{verifier}.
Δ	Required improvement margin; a witness satisfies $f(x) > f(x_{\mathrm{sel}}) + \Delta$.
$R_\Delta,\ R = R_0$	First qualifying witness rank, counting the nomination; R uses zero margin. The rank is infinite if no witness exists.
ρ_{tail}	Spearman correlation between proxy and trusted rankings among the highest-scoring trusted candidates.
SD, σ	Standard deviation of trusted pool scores and of measurement noise, respectively.
$\delta,\ L_B^{\mathrm{Gauss}}$	Allowed error probability and Gaussian-noise certificate, with $\Pr(L_B^{\mathrm{Gauss}} \le G_{\mathrm{verifier}}) \ge 1-\delta$ under the stated assumptions.

A Proofs and formal remarks

The two gaps. In Equation (1), A is a search algorithm run with a fixed budget on the trusted score f, and $x_{A,f}$ is the candidate it returns from the pool $\mathcal{X}$. The search gap $G_{\mathrm{hard}}(A)$ compares its score with that of the best trusted candidate, x_{true}. The selection gap G_{verifier} instead compares x_{true} with the proxy nomination x_{sel}, the candidate with the highest learned score $\hat{f}$. Both gaps are nonnegative because neither candidate can score above the best candidate in the pool. The nomination audit concerns only G_{verifier}, because it tests the proxy's choice; $G_{\mathrm{hard}}(A)$ describes the separate task of searching the trusted score directly. These gaps are not assumed to add up to an overall error.

Certificate validity and witness-budget equivalence. Let S_B contain the at most B candidates evaluated by the audit, including the nomination x_{sel}, and let $\mathcal{X}$ be the full candidate pool. Thus $x_{\mathrm{sel}} \in S_B \subseteq \mathcal{X}$, which gives

$$0 \le L_B = \max_{x \in S_B} f(x) - f(x_{\mathrm{sel}}) \le \max_{x \in \mathcal{X}} f(x) - f(x_{\mathrm{sel}}) = G_{\mathrm{verifier}}$$

For a fixed top-B prefix, $L_B > \Delta$ exactly when at least one of the first B candidates exceeds $f(x_{\mathrm{sel}}) + \Delta$, which by Equation (3) is equivalent to $R_\Delta \le B$. For an adaptive policy, the same set inclusion proves validity after every round because it does not depend on how S_B was selected. Even if a policy decides when to stop after seeing the exact scores, a positive certificate still corresponds to a measured candidate that truly improves on the nomination. A value $L_B = 0$ means that none of the evaluated candidates improves on the nomination; if candidates remain unmeasured, this does not rule out a witness among them.

Exact finite-population random-order null. For an audit with pool size $N = |\mathcal{X}|$, budget $1 \le B \le N$, and W qualifying witnesses satisfying $f(x) > f(x_{\mathrm{sel}}) + \Delta$, a random ordering that keeps the nomination first and samples the remaining candidates uniformly without replacement finds at least one witness by budget B with probability

$$p_{\mathrm{null}}(B \mid N, W) = 1 - \frac{\binom{N-1-W}{B-1}}{\binom{N-1}{B-1}} \tag{4}$$

The dashed curves in Figure 2 average Equation (4) over the audits in each landscape. For these exact audits, unconditional recall (Recall@B or top-B recall) is the fraction of all audits finding a qualifying witness within B candidate evaluations, including the nomination; audits without a qualifying witness count as uncertified. The reported recall ratios at $B = 16$ are the observed unconditional recall divided by this averaged random-order recall. Thus the null preserves the pool size and witness prevalence of every audit and randomizes only the ordering after the nomination.

Incumbent-preserving delivery and regret removal. Let $x_{\mathrm{safe}} = \arg\max_{x \in S_B} f(x)$ be the best evaluated candidate, and recall that x_{true} is the best candidate in the full pool. Because the evaluated set S_B includes the nomination x_{sel}, Equation (2) gives

$$f(x_{\mathrm{safe}}) - f(x_{\mathrm{sel}}) = L_B \ge 0$$

Thus the delivered candidate cannot be worse than the nomination. Its remaining selection regret is

$$f(x_{\mathrm{true}}) - f(x_{\mathrm{safe}}) = f(x_{\mathrm{true}}) - f(x_{\mathrm{sel}}) - \{f(x_{\mathrm{safe}}) - f(x_{\mathrm{sel}})\} = G_{\mathrm{verifier}} - L_B$$

which lies in $[0, G_{\mathrm{verifier}}]$. Whenever $G_{\mathrm{verifier}} > 0$, the fraction of the measurable selection regret removed is therefore exactly $L_B / G_{\mathrm{verifier}}$, the quantity plotted in Figure 3.

Noisy certificate validity. Fix the evaluated candidates S_B and their repeat counts before collecting measurements, and assume independent Gaussian measurement errors with mean zero and known variance, as in Appendix H. Let δ be the allowed probability of an incorrect comparison across all $m = |S_B| - 1 \ge 1$ candidates other than the nomination. For each candidate x, its lower confidence bound on the true improvement $f(x) - f(x_{\mathrm{sel}})$ exceeds that improvement with probability at most δ/m. Adding these m error probabilities by the union bound shows that the probability of any such error is at most $m(\delta/m) = \delta$.

The comparisons share $\bar{Y}_{x_{\mathrm{sel}}}$, the mean of the nomination's repeated measurements, so they are not independent. The union bound does not require independence between comparisons: with

probability at least $1-\delta$, all their lower bounds are simultaneously no greater than the corresponding true improvements. The certificate L_B^{Gauss} is the largest of these lower bounds, or zero if all are negative. On this event, it cannot exceed G_{verifier}, because no candidate's improvement exceeds the nomination's shortfall from the pool optimum and $G_{\mathrm{verifier}} \geq 0$. This proves $\Pr(L_B^{\mathrm{Gauss}} \leq G_{\mathrm{verifier}}) \geq 1-\delta$.

On the same event, any candidate whose own lower bound is positive truly improves on the nomination. Returning such a candidate, or retaining the nomination otherwise, therefore preserves or improves the trusted score with probability at least $1-\delta$. If measurements are collected sequentially and the bounds are checked repeatedly to decide when to stop, the error calculation must also account for those repeated checks; the fixed-sample argument alone does not provide that guarantee.

Proof of Proposition 1. For a pool of N candidates, let k be the number of highest-scoring trusted candidates considered in the proposition, and write $s = x_{\mathrm{sel}}$ for the nomination. Choose these k candidates, $t_1, \ldots, t_k$, in strictly decreasing trusted-score order, with $f(t_i) > f(s) + \Delta$ for every i. The remaining $N-k-1$ candidates have scores at most $f(s)+\Delta$, so none is a qualifying witness.

Both proxy rankings place s first and preserve the relative order $t_1, \ldots, t_k$. Their nomination is therefore identical, and their Spearman correlation within this trusted top-k group is one. The first ranking places t_1 immediately after s, so its first witness has rank two. The second places all $N-k-1$ remaining non-witnesses between s and the top-k group. There are then $1+(N-k-1) = N-k$ candidates before the first witness, giving rank $N-k+1$. Thus getting the order right within the highest-scoring group does not determine how early that group appears in the full proxy ranking.

B Datasets and protocol details

Each measured dataset is divided into separate training, validation, and selection sets. The selection set is the declared pool, meaning the fixed set of candidates to which each audit applies.

Table 3: **Measured datasets and separate training, validation, and selection sets.** The counts reflect the data preparation described in the last column, and no candidate appears in more than one set within a split. Pool optima and witness ranks are defined relative to the measured selection set.

Landscape	Measured	Train/val/selection	Data checks
TF-Bind-8	65,536	4,096/2,048/59,392	Removed 256 duplicate rows whose sequences and scores matched an existing entry.
TrpB	159,129	4,960/1,240/152,929	Measured scores only; variants with stop codons (signals that end protein synthesis) are excluded.
GB1	149,361	4,960/620/143,781	Only directly measured scores are used; the separate dataset of estimated scores is excluded.
avGFP	51,714	4,096/2,048/45,570	Retained sequences are unique and have measured scores that are neither missing nor infinite.

For each split, training and validation candidates are sampled without replacement from the measured dataset after the checks in Table 3, and the remaining candidates form the declared selection pool. Split identifiers fix the sampling, model initialization, tie breaking, and every stochastic policy. No unobserved selection-pool score, or statistic computed from those scores, is used to fit a deployable proxy, select a hyperparameter, normalize a policy input, or choose an action. Pool standard deviation enters only the retrospective reporting margin. TF-Bind-10 is excluded because its canonical shards repeat every sequence under multiple conflicting labels, so iteration order changes the apparent optimum; TF-Bind-8 has no such ambiguity after cleaning.

GB1 uses $\log(1+y)$, where y is measured binding fitness; this transformation preserves the ordering of candidates [18]. TrpB uses measured enzyme fitness [19]. TF-Bind-8 uses the transcription-factor binding dataset from Design-Bench after duplicate records are removed [20, 4], and avGFP uses ProteinGym's measured fluorescence values for the Sarkisyan landscape [21, 22]. The selection pool is fixed after splitting, so the trusted optimum, selection regret, witness set, and random-order baseline are defined relative to that finite pool.

The primary proxy suite uses additive one-hot ridge for avGFP, additive-plus-pairwise one-hot ridge for the four-site and eight-mer landscapes, and ExtraTrees on integer token features. Ridge penalties

and tree settings are chosen on validation data. Ranking and conservative rows use the published objectives and schedules detailed in Appendix C. Ties in proxy score are resolved using a fixed ordering of candidate identifiers, without consulting trusted scores.

The 200 proxy-family audits consist of repeated splits across the models in Table 4: 15 splits for each of four model families on TF-Bind-8, TrpB, and GB1, and 10 splits for each of two families on avGFP. The additional 30 audits reproduce a ranking pipeline and Conservative Objective Models at their published schedules. These methods search over continuous numerical representations and then convert, or decode, the results into candidate sequences. We distinguish the pool candidate with the highest proxy score, the generated sequence with the highest proxy score among those belonging to the pool, and the sequence actually returned by the method. Finite-pool evaluation metrics and witness ranks are computed only for candidates in the declared pool.

Every audit counts the nomination as its first trusted evaluation, and a candidate is a qualifying witness only if its trusted score satisfies the strict improvement condition $f(x) > f(x_{\mathrm{sel}}) + \Delta$. We form confidence intervals by resampling data splits within each landscape and taking percentile bounds, keeping results from the same split together when comparing methods. Because all resampled splits reuse the same measured datasets, these intervals do not establish whether the findings generalize to other scientific domains.

Table 4: **Selection failure in every combination of landscape and model across 200 proxy-family audits.** Brackets show 95% confidence intervals from bootstrap resampling of data splits, while the two avGFP rows are reported without intervals. Selection regret G_{verifier} is expressed on each landscape's score scale, so its numerical values should not be compared across landscapes. The rank and recall columns use zero margin ($\Delta = 0$, $R = R_0$), meaning that any strictly higher trusted score qualifies as a witness; Top-8 recall is the fraction of audits finding such a witness within eight evaluations, including the nomination.

Landscape	Proxy family	Splits	Mean G_{verifier}	Median / max R	Top-8 recall
TF-Bind-8	pairwise ridge	15	0.037 [.032, .043]	3 / 11	0.933
TF-Bind-8	ExtraTrees	15	0.102 [.071, .137]	2 / 17	0.933
TF-Bind-8	ranking	15	0.049 [.035, .066]	3 / 43	0.733
TF-Bind-8	conservative	15	0.076 [.047, .111]	5 / 48	0.667
TrpB	pairwise ridge	15	0.258 [.172, .352]	3 / 30	0.933
TrpB	ExtraTrees	15	0.510 [.433, .596]	6 / 13	0.933
TrpB	ranking	15	0.702 [.615, .787]	3 / 8	1.000
TrpB	conservative	15	0.924 [.823, .982]	3 / 26	0.667
GB1	pairwise ridge	15	0.406 [.318, .494]	2 / 23	0.867
GB1	ExtraTrees	15	0.869 [.690, 1.084]	5 / 14	0.867
GB1	ranking	15	1.352 [1.203, 1.513]	3 / 5	1.000
GB1	conservative	15	1.053 [.756, 1.389]	2 / 34	0.867
avGFP	additive ridge	10	0.456	3 / 2,067	0.800
avGFP	ExtraTrees	10	0.352	4 / 29	0.800

C Published optimizers at full schedule

We audit two published optimization methods, distinguishing the best candidate under each learned model from the candidate its search actually returns.

Ranking pipeline. We use RaM-ListNet [11], which trains a ranking model with the ListNet loss [25]. This loss compares probability distributions formed from measured and predicted scores across each list of candidates. After training, Adam updates candidate representations to increase their scores under the fixed model. A TF-Bind-8 reference run completes 100 epochs over the full 10,000-list corpus and achieves a normalized score consistent with the published results. The finite-landscape audit contains five splits each of TF-Bind-8, TrpB, and GB1, using 10,000 fixed lists of length 1,000, an $8{,}000/2{,}000$ partition, batch size 128, width $[2048, 2048]$, 100 epochs, and 128-start, 100-step search. Median witness ranks are 3, 2, and 4; top-32 recall is one on all three landscapes.

Conservative surrogate. The Conservative Objective Models surrogate [9] is trained for 50 epochs. Its search follows the original method, updating candidate representations by gradient ascent to increase the learned score. Across five splits per landscape, median witness ranks are 7, 3, and 2 on TF-Bind-8, TrpB, and GB1, with top-8 recall 0.6, 1.0, and 0.6. Figure 4 summarizes the full-schedule results for both pipelines.

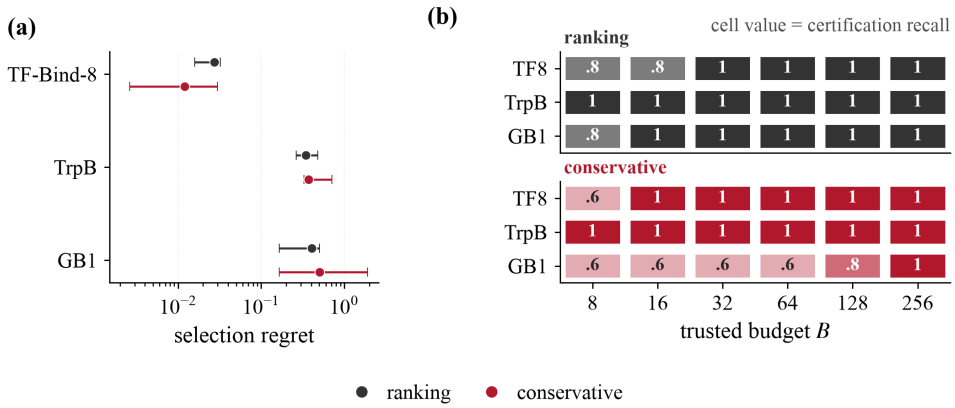


Figure 4: **Both published optimizers exhibit nomination-level failures at their full schedules.** *(a)* Selection regret of the exact in-pool learned-score optimum, five splits each, median with range. *(b)* Certification recall against trusted budget. The ranking pipeline reaches recall one by $B = 32$ on all three landscapes.

Checking whether proposals belong to the declared pool. The original searches can propose candidates outside the declared pool, so we keep three quantities separate: the *exact proxy optimum within the declared pool* maximizes the learned score over the declared pool and is the object to which our theory applies, the *best eligible proposal* is the generated sequence with the highest proxy score among those that belong to the pool, and the method's actual output may lie outside the pool. Ineligible deliveries have undefined finite-pool metrics. We did not project or replace these outputs, or estimate their missing finite-pool scores.

We then reran all 30 optimizer instances, restricting the ranking pipeline's proposed candidates and the conservative pipeline's candidate updates to the declared pool. All returned candidates belonged to the pool, without projection or substitution after optimization. Here, Recall@16 is the fraction of runs that find a qualifying witness within 16 trusted evaluations, including the method's returned candidate. For the ranking pipeline, median witness ranks relative to that returned candidate were 8 on GB1, 3 on TF-Bind-8, and 4 on TrpB, with Recall@16 values of 0.6, 1.0, and 1.0. For the conservative pipeline, the corresponding median ranks were 27, 3, and 20, with Recall@16 values of 0.0, 0.2, and 0.2. Thus, restricting both searches to the pool still leaves better candidates that the audit can find. The conservative pipeline never returns the candidate with the highest proxy score, and its witnesses generally require more evaluations to find.

D Margin sensitivity

Table 5 reports the margin sweep summarized in Section 5.3. The ratio Δ/SD expresses the required improvement in units of the standard deviation of trusted scores across the selection pool. The column $\Pr(R_\Delta < \infty)$ gives the fraction of audits with a qualifying witness anywhere in the pool. Reporting witness existence separately from conditional rank is essential here: as the required improvement grows, meaningful witnesses may disappear, and the conditional rank distribution then describes only the audits in which such a witness still exists.

At large margins, conditional medians can be based on few surviving audits. For example, the avGFP median at $\Delta/\mathrm{SD} = 0.5$ uses only two audits with a qualifying witness. We therefore report witness existence together with conditional rank.

Table 5: **Margin sensitivity across all 230 audits.** Recall@16 is the fraction of all audits finding a qualifying witness within 16 evaluations, including the nomination. The median and P90 (90th-percentile) witness ranks are calculated only for audits in which a qualifying witness exists.

Landscape	Δ/SD	$\Pr(R_\Delta < \infty)$	Median R_Δ	P90 R_Δ	Recall@16
avGFP	0.00	1.000	3.5	30.0	0.800
	0.05	0.950	5.0	25.4	0.800
	0.10	0.950	8.0	56.2	0.650
	0.25	0.800	17.5	249.5	0.400
	0.50	0.100	7.0	11.0	0.100
GB1	0.00	1.000	3.0	14.0	0.929
	0.05	1.000	3.0	14.2	0.914
	0.10	1.000	3.0	18.5	0.886
	0.25	1.000	4.0	34.1	0.843
	0.50	1.000	5.5	67.1	0.800
TF-Bind-8	0.00	1.000	3.0	12.1	0.914
	0.05	0.986	5.0	32.2	0.786
	0.10	0.900	5.0	397.0	0.500
	0.25	0.443	4.0	255.0	0.271
	0.50	0.243	6.0	239.4	0.171
TrpB	0.00	1.000	3.0	8.1	0.971
	0.05	1.000	3.0	8.1	0.971
	0.10	1.000	3.5	8.1	0.957
	0.25	1.000	4.0	12.2	0.914
	0.50	1.000	6.0	24.5	0.843

E Model quality and candidate ordering

To distinguish the effects of model accuracy and candidate ordering on witness rank, we compare changes to the proxy with controlled rankings and shuffling tests.

The proxy-quality study evaluates 525 combinations of model and data split, covering five to seven model families per landscape. Three model changes were evaluated without using selection-pool labels: a top-decile-weighted ridge model, an equal-weight ensemble after standardizing predictions, and frozen ESM-2 embeddings with ridge heads on GB1 and avGFP. Replaying the saved predictions confirms top-tail Spearman ceilings of 0.383 for the baseline proxy set and 0.402 for the extended set. An ensemble whose weights were fitted using the highest-scoring GB1 validation candidates reaches a top-tail Spearman correlation of 0.743 on those candidates but only 0.319 on the separate selection pool. Its high validation correlation therefore does not establish better ranking accuracy on the candidates used for the audit.

The oracle quality ladder is a set of artificial rankings constructed by combining rescaled trusted ranks of all candidates in the measured dataset with independent Gaussian noise. For each of 30 fixed seeds per landscape, we tune the noise level on validation data to target correlations of 0.2, 0.4, 0.6, 0.8, and 0.95, then evaluate the resulting rankings on the separate selection pool. Because these rankings use known trusted scores, they test how ranking accuracy affects witness discovery; they cannot be used to predict scores for candidates that have not yet been measured. Here ρ_{tail} denotes Spearman correlation between proxy and trusted rankings within the candidates with the highest trusted scores. At the 0.95 target, mean validation / selection top-tail correlations are $0.950/0.948$ on avGFP, $0.965/0.902$ on GB1, $0.950/0.953$ on TF-Bind-8, and $0.949/0.946$ on TrpB. Selection-pool global Spearman correlation is at least 0.996 and normalized discounted cumulative gain for the top 1% (NDCG@1%) is at least 0.989 in these experimental conditions. Higher validation correlation generally corresponds to higher selection-pool correlation, but the values differ. Table 2 therefore groups runs by the correlation measured on the selection pool, rather than the target used during tuning. Table 6 summarizes the size, construction, and reported outcome of the model comparisons, artificial rankings, and shuffling tests.

The rank-shuffle ablation permutes the same learned scores globally, within score deciles, or within the top one percent while preserving the remaining ordering. Table 7 reports the complete landscape-

Table 6: **Studies of model quality and candidate ordering.** The proxy models are fitted and tuned without trusted selection-pool scores, which are used to evaluate them. The artificial rankings use those scores to test how ranking accuracy affects witness discovery.

Component	Size	Construction	Reported check
Trained proxy models	525 combinations	Five to seven model families per landscape	Conditional median R_Δ of 2–6; best $\rho_{\mathrm{tail}} = 0.402$
Oracle quality ladder	600 rankings	Four landscapes, five targets, 30 fixed seeds	High-quality outcomes in Table 2
Rank-shuffle ablation	16,000 rows	Global, score-decile, and top-1% permutations	Recall@16 in Table 7

level recall summary. The three granularities distinguish broad head membership from fine ordering near the nomination. This ablation averages results by model group rather than giving every audit equal weight as in Section 4, so recall for the unshuffled ordering differs from Table 1 by at most 0.048. All shuffle comparisons use the same aggregation.

Table 7: **Recall@16 under the rank-shuffle controls.** Every comparison holds the candidate pool, witness set, and nomination fixed; only the ordering available to the audit changes.

Landscape	Global shuffle	Within score deciles	Within top 1%
avGFP	0.467	0.754	0.807
GB1	0.108	0.356	0.633
TF-Bind-8	0.072	0.255	0.492
TrpB	0.208	0.433	0.715

F Repair and acquisition-policy details

All policy comparisons begin with the nomination already evaluated, acquire distinct in-pool candidates, and hide the trusted pool scale and every unacquired trusted label. A fixed-prefix policy evaluates candidates in descending proxy-score order, while uniform random probing selects from the remaining unevaluated candidates with equal probability. The local policy prioritizes unmeasured sequence neighbors of the best trusted probe, breaking distance ties by proxy score; before its first improvement it begins from the nomination. Diversity shortlists favor candidates that differ from those already evaluated in their input features, while disagreement shortlists favor candidates whose predicted scores vary most across the fixed collection of proxy models. In the multi-round comparison, Thompson sampling selects candidates from a posterior whose prior and update rule are fixed before the campaign, while batch Bayesian optimization selects each batch using its acquisition score; both methods update only from labels acquired in earlier rounds. The exact feature maps, posterior settings, acquisition optimizers, and random seeds are fixed before the policy receives any trusted scores from the selection pool.

For the fixed single-round policies, Figure 5 shows certification recall in panel (a) and the mean fraction of selection regret removed in panel (b).

Independent recomputation. The repair experiment uses a separate collection of 248 recorded cases from TF-Bind-8, TrpB, GB1, and avGFP in which the nominated candidate was suboptimal; this is not a recount of the 230 audits in Table 1. Comparing prefix and random probing at multiple budgets produces 34,224 saved results for the audit metrics, including the certificate and the trusted score of the returned candidate. Independently recalculating these quantities reproduced every recorded value.

At budget $B = 256$, we compare prefix and random probing on the same failure instance i. For each policy, $L_B/G_{\mathrm{verifier}}$ is the fraction of the nomination's original selection regret removed by the evaluated candidates, so the difference between the policies is

$$d_i = \left(\frac{L_B}{G_{\mathrm{verifier}}}\right)_{\mathrm{prefix},i} - \left(\frac{L_B}{G_{\mathrm{verifier}}}\right)_{\mathrm{random},i}$$

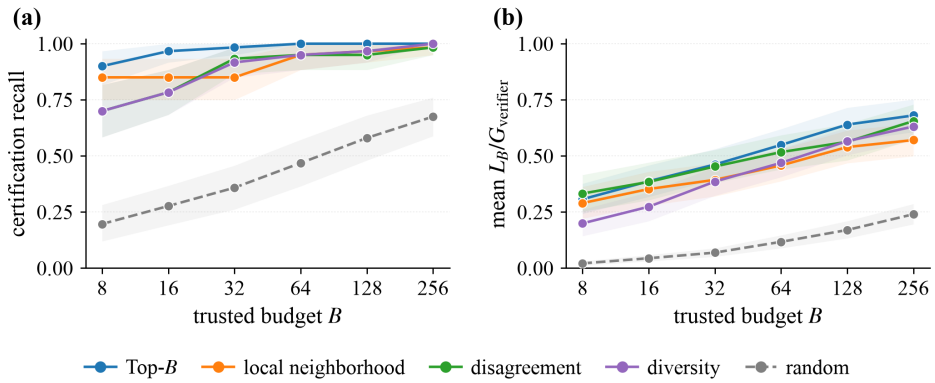


Figure 5: **Within the fixed single-round family, the proxy prefix is the strongest baseline.** *(a)* Certification recall and *(b)* mean fraction of selection regret removed, L_B/G_{verifier}, over 60 matched measured failures. A run is certified when an evaluated candidate strictly improves on the nomination ($L_B > 0$, corresponding to zero margin). The shaded areas around the curves show 95% confidence intervals obtained by bootstrap resampling of the data. The adaptive multi-round policies are compared separately in Table 8.

A positive d_i means that prefix probing removes more of the original regret than random probing on that instance. The 95% percentile intervals for the mean difference resample splits within each landscape while keeping the two policy results paired. All four intervals lie above zero, as shown in Figure 3(b). Because the candidate pool, trusted budget, and included nomination are matched, this comparison isolates the contribution of ordering.

Incumbent-preserving delivery versus refitting. The same labels can instead be used to refit the proxy and renominate, but this carries no non-degradation guarantee. Across 1,008 outcomes with probed candidates excluded from the next nomination, refitting increased mean selection regret in every measured combination of landscape and budget. Returning the best measured candidate with the nomination retained guarantees non-degradation. Across the multi-round campaigns, refitting produced 840 deliveries below the incumbent, whereas retaining it and restricting delivery to the measured set removed every violation. Of the 1,251 campaigns with a zero certificate, retaining the nomination improved the final trusted score in 1,243 cases; in the other eight, the best TF-Bind-8 probe tied the nomination exactly.

Reusing the same campaigns without rerunning acquisition, we compare the best measured candidate among new probes alone with the best among probes and nomination together. The only difference is whether the nomination remains an option for the final output. The mean gain from retaining it is $0.031\,[0.015,\ 0.048]$ for batch Bayesian optimization on GB1, $0.013\,[0.005,\ 0.022]$ on TrpB, $0.006\,[0.001,\ 0.017]$ on avGFP, and $0.005\,[0.001,\ 0.011]$ for Thompson sampling on GB1; the remaining confidence intervals include zero. Returning the best new probe alone falls below the nomination in 14.1% of batch-BO campaigns on GB1 and 11.9% on TrpB, compared with 4.1% and 1.0% for prefix probing. No incumbent-preserving delivery falls below either the best new probe or the nomination.

Matched-budget campaigns. The 7,300 matched-budget multi-round campaigns comprise 36,500 policy-round rows. Each schedule has T rounds of b new probes, with $(T, b) \in \{(4, 4), (4, 8), (8, 8), (4, 16)\}$. Including the nomination, every policy receives the same $1 + Tb$ trusted evaluations. Checks confirmed that delivered scores never fell below the nomination or decreased between rounds, no candidate was probed twice, and trusted-label access stayed within budget. A separate implementation reproduced every delivery directly from the trusted tables.

For each campaign, we compute L_B using the nomination and all candidates evaluated across the rounds. Table 8 reports the fraction of campaigns that find a witness exceeding the required margin, $L_B > \Delta$, with $\Delta = 0.05\,\text{SD}$. Every structured policy improves this fraction over random probing within each landscape, with differences ranging from 0.240 to 0.750. The confidence intervals keep policy comparisons paired within each data split and treat each split as a group; all exclude zero. They

describe variation across splits of the same measured dataset and do not establish how the advantage would vary across other scientific domains.

An adaptive policy achieves the highest mean final improvement for every combination of landscape and schedule, but which policy leads depends on the landscape. Keeping the nomination available as a final choice provides a performance safeguard regardless of how new candidates are selected. Witness rank measures how many evaluations are needed to find an improvement; it does not prescribe one search policy for every landscape.

Table 8: **All four policies find witnesses more often than random probing, with the best policy varying by landscape.** Each entry is the fraction of campaigns for that landscape and policy that find a witness exceeding the margin $\Delta = 0.05\,\mathrm{SD}$. The study contains 7,300 matched-budget campaigns in total, and bold marks the highest certification rate in each landscape.

Landscape	Prefix	Thompson	Batch BO	Witness-local	Random
avGFP	0.905	0.875	0.870	**0.925**	0.630
GB1	**0.945**	0.933	0.860	0.941	0.219
TF-Bind-8	0.905	0.979	**0.981**	0.921	0.231
TrpB	0.991	**0.993**	0.881	0.991	0.279

G Candidate-pool scaling and witness geometry

We examine how witness rank changes as the candidate pool grows, both on measured landscapes and in constructed examples. Here $N = |\mathcal{X}|$ is the number of candidates in the pool and $R = R_0$ is the first rank with a strictly higher trusted score than the nomination. A nested-pool study contains 840 audits over growing prefixes of TF-Bind-8, TrpB, and GB1. At full pool size, top-8 recall is 0.817, 0.883, and 0.900, respectively, and top-64 recall is one on all three, with median witness rank between two and six depending on model and landscape. Rank and witness prevalence are only weakly related in these rows.

A controlled study then evaluates 848 instances at pool sizes of one thousand, ten thousand, one hundred thousand, one million, and five million. At each size, the first witness is placed at rank 8, $\lceil\sqrt{N}\rceil$, $0.01N$, or $N/2$, representing constant, sublinear, one-percent, and half-pool audit costs. Table 9 gives the five-million-candidate endpoint and Figure 6 shows the full sweep. In this diagnostic, R denotes the strict, zero-margin witness rank. Witness rank remains small only under the fixed-rank construction; the other placements grow strongly with pool size.

Table 9: **At five million candidates, only the fixed-rank example reveals a witness within 64 evaluations.** The measured pools also have small median witness ranks as their size grows. The other three constructed examples place witnesses at $\sqrt{N}$, $0.01N$, and $N/2$ and are not certifiable by a small fixed budget. The Top-8, Top-16, and Top-64 columns give certification recall at those budgets.

Witness mechanism	Median R	Top-8	Top-16	Top-64
Fixed rank 8	**8**	**1**	**1**	**1**
Square root of pool	2,237	0	0	0
One percent of pool	50,000	0	0	0
Half of pool	2,500,000	0	0	0

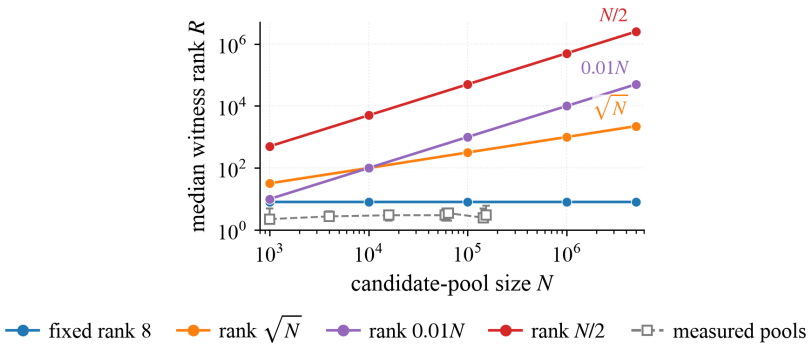


Figure 6: **Small witness rank is an empirical property, not a finite-pool consequence.** Median strict witness rank R against pool size for four constructed placement rules and the measured nested pools. Gray square markers show the median across measured audits, and their vertical bars show the range. The measured landscapes occupy the favorable small-rank regime but do not establish it universally.

H Noisy trusted evaluations

With noisy measurements, certification must account for uncertainty in each candidate's score.

Suppose the jth measurement Y_{xj} of candidate x has independent Gaussian noise, $Y_{xj} \sim \mathcal{N}(f(x), \sigma^2)$, with known standard deviation σ. Fix the evaluated candidate set S_B, including the nomination x_{sel}, and each candidate's repeat count $r_x \geq 1$ before collecting these measurements. The budget B counts candidates, including the nomination, whereas r_x counts repeated measurements of candidate x. For an error probability $0 < \delta < 1$ and $m = |S_B| - 1 \geq 1$ other candidates, define

$$L_B^{\mathrm{Gauss}} = \max\left\{0, \max_{x \in S_B \setminus \{x_{\mathrm{sel}}\}} \left[\bar{Y}_x - \bar{Y}_{x_{\mathrm{sel}}} - z_{1-\delta/m}\sigma\sqrt{r_x^{-1} + r_{x_{\mathrm{sel}}}^{-1}}\right]\right\}. \tag{5}$$

Here $\bar{Y}_x$ is the mean of the r_x repeated measurements of candidate x, and $\bar{Y}_{x_{\mathrm{sel}}}$ is the mean of the nomination's $r_{x_{\mathrm{sel}}}$ measurements. The quantity $z_{1-\delta/m}$ is the standard-normal quantile at probability $1 - \delta/m$. A candidate certifies the margin-aware audit when its lower bound exceeds Δ; comparison with zero is sufficient for non-degrading replacement. Appendix A gives the union-bound argument, while sequential measurement requires separate control of the error probability across repeated checks.

To assess whether measurement noise leads to false claims of improvement, we report six combinations of measurement error and confidence bounds, including the bound in Equation (5). Each setting is evaluated on 180 controlled top-8 audits, with 200 trials per audit and 128 measurements per candidate.

Table 10 reports the six settings testing variance estimation, large noise outliers, and systematic bias. When the measurement errors satisfy the assumptions used to construct the confidence bounds, false claims of improvement are rare in these simulations. Uncorrected systematic bias instead gives a reported false-certification rate of about 21.4%. Correcting for a known bound on the bias gives valid confidence bounds and recall 0.520, close to 0.521 for Gaussian noise with known variance. For Student-t_3 noise, a distribution with three degrees of freedom and more frequent large deviations than Gaussian noise, Chebyshev's inequality [27, Chapter 4] gives a valid but conservative bound; assuming Gaussian noise instead gives no guarantee despite a low observed error rate.

Using the observed score differences as fixed inputs, a separate calculation on the measured landscapes estimates how many repeated measurements are needed for certification. At $B = 16$ and $\sigma/\mathrm{SD} = 0.05$, where σ is measurement-error standard deviation and SD is the spread of trusted scores across the pool, the calculation requires at most three measurements per candidate for 60.0% of avGFP instances, 88.6% of GB1, 47.1% of TF-Bind-8, and 94.3% of TrpB. When the observed certificate is nonpositive, the calculation records infinity to indicate that it has no finite estimate of the repeats needed for certification. TF-Bind-8's lower share reflects its smaller improvements relative to the pool's score spread.

Table 10: **Variance estimation, heavy tails, and bias affect noisy certification.** The table reports six combinations of measurement error and confidence bounds. Status indicates whether the method's assumptions hold; a low observed error rate alone does not establish validity. False-certification rates are per 1,000 trials.

Noise and method	Status	False certification per 1,000	Recall
Gaussian noise			
Gaussian, known variance	valid	0.132	0.521
Gaussian, estimated variance (pooled t)	valid	0.526	0.511
Heavy-tailed noise (Student-t_3)			
Chebyshev bound	valid	0.044	0.238
Gaussian bound	no guarantee	1.842	0.522
Systematic bias			
Uncorrected bias	**invalid**	**213.947**	0.835
Correction using a known bias bound	valid	0.482	0.520

I Bounding the fraction of better candidates

When the structured shortlist finds no witness, an independent random sample can still bound how common witnesses are in the full pool.

The random sampling procedure draws q candidates uniformly without replacement from $\mathcal{X}$, independently of the proxy, structured shortlist, and their outcomes. If k witnesses are observed and W is the unknown total number in a pool of size N, the random witness count K has the hypergeometric distribution $K \sim \mathrm{Hypergeom}(N, W, q)$ for sampling without replacement. For error level $0 < \alpha < 1$, exact lower-tail inversion gives

$$\widehat{p}_{1-\alpha} = \frac{1}{N} \max\{W : \Pr(K \leq k \mid N, W, q) > \alpha\}. \tag{6}$$

The quantity $\widehat{p}_{1-\alpha}$ is a one-sided upper confidence bound on W/N, the fraction of candidates in the full pool that are witnesses, at confidence level $1 - \alpha$. It needs no calibration set, exchangeability model, or assumption linking the trusted score f to the proxy score $\hat{f}$. It does require unambiguous witness labels; with noisy evaluation, the confidence-adjusted comparisons of Appendix H must be used instead, and the exact hypergeometric guarantee does not automatically transfer.

The paired study uses 365 audits at $\Delta = 0$: the 200 original audits and 165 additional proxy fits, excluding the 30 optimizer audits. This gives 50 audits on avGFP and 105 on each other landscape; the headline results instead use 230 audits at $\Delta = 0.05\,\mathrm{SD}$. Over 876,000 draws spanning four landscapes, two budgets, and several confidence levels, the exact inversion returned an upper bound below one in 99.85% of draws. Empirical coverage is the fraction of draws whose upper bound is at least the true fraction of witnesses. It met or exceeded the requested confidence level for every tested combination of landscape, budget, and confidence level. Table 11 reports the mean bounds at 95% confidence against measured prevalence at $\Delta = 0$. At $q = 16$, the bounds are two to fifteen times the measured prevalence. The shortlist and random sample have separate evaluation budgets: an audit requesting both spends B labels on the shortlist and q on the independent random draw.

Table 11: **Random probes bound pool witness prevalence.** Mean upper bounds at 95% confidence against measured witness prevalence for the 365-audit study at $\Delta = 0$. This cohort and margin differ from the headline analysis in Figure 2. Prevalence is the mean over audits of the pool fraction exceeding $f(x_{\mathrm{sel}})$; coverage is reported as a percentage.

Landscape	Prevalence	Bound, q=16	Bound, q=64	Coverage (%), 16	Coverage (%), 64
avGFP	0.141	0.338	0.218	99.16	97.45
GB1	0.013	0.189	0.065	99.95	99.88
TF-Bind-8	0.021	0.200	0.075	99.95	99.61
TrpB	0.076	0.250	0.130	99.62	99.46

In this 365-audit study, evaluating the first 16 candidates in proxy order, including the nomination, finds a candidate with a higher measured score in 88.00%, 94.29%, 92.38%, and 95.24% of avGFP, GB1, TF-Bind-8, and TrpB audits, respectively. For every audit where the shortlist finds no improvement, the independent random sample still provides an upper confidence bound on the fraction of the full pool that beats the nomination. Finding a witness establishes a specific improvement over the nomination, while the random-sample bound limits the fraction of the pool that could offer such improvements.