PROVING THE PROXY WRONG: WITNESS RANK FOR AUDITING SCIENTIFIC DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Proxy-guided scientific design can rank far more candidates than trusted experiments can evaluate, yet standard metrics do not quantify the evidence for the single candidate ultimately selected. We ask how many trusted evaluations are needed to overturn that nomination. On a declared finite pool, the margin-aware *witness rank* R_Δ is the position, under an ordering fixed without trusted labels, of the first candidate that improves on the nomination by more than Δ. Across 230 audits on four fully measured protein and DNA landscapes, including two published offline optimizers, conditional median witness ranks are three to five despite the largest pool containing over 150,000 candidates. At sixteen trusted evaluations, proxy ordering achieves 1.68 to 8.60 times the exact random-order recall. Controlled rankings with Spearman correlation near 0.95 among the top candidates by measured performance still yield conditional median witness ranks of two to four, showing that high ranking quality alone does not determine audit cost. The same audit measurements identify replacements that cannot reduce measured performance, while independent random probes bound pool witness prevalence when no witness is found. To complement these retrospective biological audits, we demonstrate the auditing and replacement protocol in a prospective laboratory study of ionic conductivity in battery electrolytes. Together, these results make post-selection reliability measurable in experimental-budget units and guide whether to retain a nomination or replace it with a measured improvement.

1 INTRODUCTION

Scientific design pipelines increasingly fit a proxy to scarce measurements, optimize it over an admissible library, and commit one nominated candidate to trusted evaluation (Wang et al., 2023; Trabucco et al., 2022). A *trusted* score comes from an experiment or a designated reference evaluator. A poor result can originate in either the search or the proxy. An optimizer may miss a good candidate even with access to the trusted objective, or the proxy may assign its highest score to an inferior candidate. Better search can address the first problem. The second persists even when the proxy is optimized exactly. Held-out error evaluates the proxy on average, and search benchmarks evaluate whether an optimizer can maximize its objective, but neither quantifies the evidence for the realized nomination. Average prediction error can hide large errors in the candidates selected for their high proxy scores (Smith & Winkler, 2006; Gao et al., 2023).

Determining how far a nomination lies below the pool optimum requires trusted values across the pool; proving it suboptimal requires only one counterexample, which we call a *witness*. After fixing an ordering without access to trusted labels, its witness rank is the position of the first witness: the trusted-evaluation cost of falsification for that ordering. A margin-aware rank R_Δ requires the witness to improve on the nomination by more than Δ, separating consequential failures from negligible gains. To certify an improvement is to establish it through trusted evaluations.

The central empirical finding is that this audit can be cheap even when the design space is large. Across four fully measured protein and DNA landscapes, the declared selection pools contain 45,570 to 152,929 candidates, yet when a meaningful witness exists, its median rank is only three to five. Sixteen trusted evaluations certify a meaningful improvement in 78.6% to 97.1% of audits across these landscapes, and structured orderings outperform random probing at the same budget. The nominations themselves rank highly by trusted value. Even controlled rankings with Spearman correlation near

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.

0.95 among the selection pool's best candidates by trusted value retain small conditional witness ranks. Higher ranking quality can therefore reduce failure frequency while leaving the remaining failures inexpensive to expose.

The audit also guides *delivery*, the choice of the final candidate. Our *repair* rule returns the best evaluated candidate, retaining the nomination unless a measured candidate is better. This improves trusted value by exactly the certified amount under any acquisition policy. When no witness is found, an independent uniform sample gives an exact, distribution-free upper bound on pool witness prevalence.

We develop witness rank as a nomination-level audit measure through three contributions. First, we establish sharp worst-case bounds on witness rank from proxy errors and show that proxy ordering is minimax-optimal: among fixed deterministic orderings that use no trusted pool scores and evaluate the nomination first, it minimizes worst-case audit cost under the same error bound. Perfect correlation within the trusted top tail nevertheless leaves audit cost essentially unconstrained. Second, we measure witness rank across scientific design pipelines and characterize the empirical conditions for small ranks through ordering, proxy-quality, margin, and scaling studies. Third, we connect the audit evidence to standard best-observed delivery and an exact bound on pool witness prevalence, yielding a protocol for deciding whether to overturn or retain a proxy-guided recommendation.

2 Related work

Offline design and proxy exploitation. Offline model-based optimization uses a fixed dataset to propose candidates for expensive trusted evaluation. Approaches include mapping desired scores back to designs (Kumar & Levine, 2020), penalizing unsupported predictions (Trabucco et al., 2021), constraining designs through forward and backward mappings (Chen et al., 2022), and learning rankings instead of regressing scores (Tan et al., 2025). Design-Bench provides a common evaluation framework (Trabucco et al., 2022). The failure these methods contend with is documented: maximizing noisy estimates produces post-decision disappointment (Smith & Winkler, 2006), overoptimizing a learned score can reduce trusted performance (Gao et al., 2023), in-silico rankings depend on the evaluation oracle (Surana et al., 2025), and decisions need not track average prediction error (Donti et al., 2017; Elmachtoub & Grigas, 2022; Lyu et al., 2026). These studies relate prediction and optimization errors to decision quality; we instead measure the trusted evidence needed to overturn a nomination.

Distribution-free inference and best-arm identification. Conformal prediction uses held-out errors to set prediction bounds (calibration) and, under exchangeability, guarantees *coverage*, a specified minimum probability of containing a new candidate's true score (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023). Such bounds could exclude candidates unable to beat the nomination, but applying a union bound across our pools makes split-conformal bounds infinite at available calibration sizes (Section 3.5). We therefore bound the fraction of witnesses by random sampling without replacement, focusing on a fixed nomination rather than the expected losses controlled by related methods (Angelopoulos et al., 2024; 2025). Best-arm identification seeks the pool's best candidate (Even-Dar et al., 2006; Audibert et al., 2010), whereas our audit can stop at the first measured improvement (Appendix A.2). We evaluate four measured protein and DNA landscapes (Wu et al., 2016; Johnston et al., 2024; Barrera et al., 2016; Sarkisyan et al., 2016; Notin et al., 2023) and use molecular mixtures (Rajaonson et al., 2025; Zhu et al., 2024; Zohair et al., 2025; Ruza et al., 2025) for controlled mechanism studies.

3 Nomination-level auditing

3.1 Setting and the two failures

Let $\mathcal{X}$ be a finite declared pool of $N = |\mathcal{X}|$ candidates, $f : \mathcal{X} \to \mathbb{R}$ a trusted reference score, and $\hat{f} : \mathcal{X} \to \mathbb{R}$ a proxy fit to an offline dataset disjoint from $\mathcal{X}$, with higher scores better. Write $x_{\mathrm{true}} \in \arg\max_{x \in \mathcal{X}} f(x)$ and $x_{\mathrm{sel}} \in \arg\max_{x \in \mathcal{X}} \hat{f}(x)$ for the trusted optimum and the nomination returned by exhaustive proxy search over $\mathcal{X}$. A search algorithm A with a fixed evaluation budget returns $\hat{x}_A$ using proxy scores $\hat{f}$ and x_A using trusted scores f. With deterministic tie breaking,

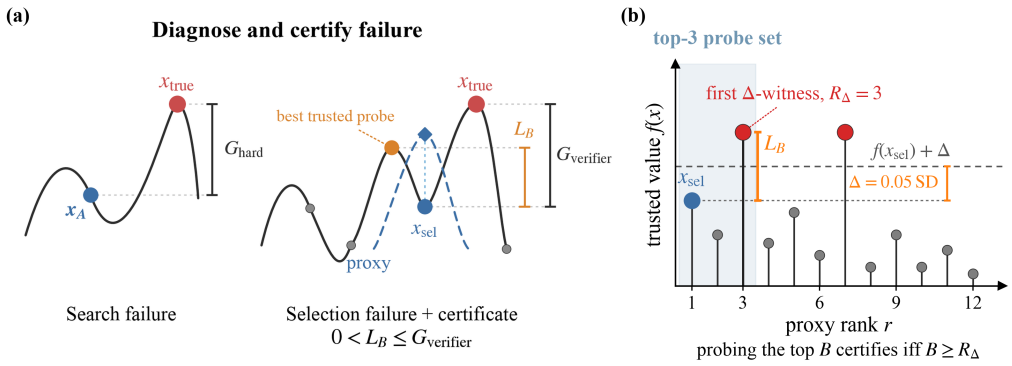

Figure 1: **Two failure modes, and how the proxy's own ranking exposes one of them.** *(a)* Search failure leaves G_{hard} when the optimizer does not reach x_{true}. Selection failure leaves G_{verifier} when the proxy (dashed) ranks x_{sel} highest. Our audit targets this selection failure: a better trusted probe gives $0 < L_B \le G_{\text{verifier}}$. *(b)* The same pool in proxy-rank order. Candidates above the dashed line at $f(x_{\text{sel}}) + \Delta$ are Δ-witnesses, the first at rank three, so the shaded top-3 probe set certifies an improvement of L_B and delivers it. In this schematic, SD denotes the standard deviation of trusted scores in the pool; Section 5.2 gives the measured distribution.

define

$$G_{\text{hard}}(A) = f(x_{\text{true}}) - f(x_A), \qquad G_{\text{search}}(A) = \hat{f}(x_{\text{sel}}) - \hat{f}(\hat{x}_A), \tag{1}$$

$$G_{\text{verifier}} = f(x_{\text{true}}) - f(x_{\text{sel}}), \qquad G_{\text{total}}(A) = f(x_{\text{true}}) - f(\hat{x}_A). \tag{2}$$

Here G_{hard} is regret under trusted-objective search, G_{search} is the remaining proxy-score gap, and G_{total} is the actual delivery's trusted regret. The nomination's *selection regret*, G_{verifier}, is the component that further proxy optimization cannot fix (Figure 1). A nomination fails when $G_{\text{verifier}} > 0$.

Proposition 1 (Audit identity and uniform bound)**.** *For any such search algorithm A, let $H(A) = f(\hat{x}_A) - f(x_{\text{sel}})$ be its output's trusted-value advantage over the exact proxy nomination. Then $G_{\text{total}}(A) = G_{\text{verifier}} - H(A)$. Let $\epsilon_\infty = \max_{x \in \mathcal{X}} |f(x) - \hat{f}(x)|$ be the largest absolute proxy error in the pool. Then $G_{\text{total}}(A) \le G_{\text{search}}(A) + 2\epsilon_\infty$ and $G_{\text{verifier}} \le 2\epsilon_\infty$, and no smaller factor holds for all such problems.*

Appendix A proves the proposition. If $H > 0$, moving from the algorithm's output to the exact proxy nomination increases trusted regret, formalizing overoptimization.

3.2 Witness rank and its budget interpretation

Computing G_{verifier} requires trusted scores over the whole pool; certifying a positive lower bound requires only a witness. Given a probe set $S_B \subseteq \mathcal{X}$ of at most B trusted evaluations containing x_{sel}, define the *audit certificate*

$$L_B = \max_{x \in S_B} f(x) - f(x_{\text{sel}}). \tag{3}$$

Because $x_{\text{sel}} \in S_B \subseteq \mathcal{X}$, $0 \le L_B \le G_{\text{verifier}}$: a positive value proves the nomination suboptimal, while zero is inconclusive. To require a practically meaningful gain, choose a margin $\Delta \ge 0$ and sort the candidates by decreasing proxy score, using the same deterministic tie breaking as for the nomination. Let $x_{(r)}$ denote the candidate in position r of this list, so $x_{(1)} = x_{\text{sel}}$. The first position with an improvement exceeding Δ is

$$R_\Delta = \min\{r : f(x_{(r)}) > f(x_{\text{sel}}) + \Delta\}, \qquad R_\Delta = \infty \text{ if none exists.} \tag{4}$$

When S_B contains the first B candidates in this ordering (Figure 1b),

$$L_B > \Delta \iff R_\Delta \le B. \tag{5}$$

Thus R_Δ is the trusted-evaluation budget needed to certify an improvement exceeding Δ. The same definition applies to any probe ordering fixed without unobserved trusted labels; Section 5.2 compares these orderings at matched cost. We use $R = R_0$ for strict improvement and $\Delta = 0.05\,\text{SD}$ for the headline analysis.

3.3 SHARP WORST-CASE BOUNDS ON AUDIT COST

Uniform proxy-error bounds constrain where a witness can appear. Let $N(t) = |\{x \in \mathcal{X} : \hat{f}(x) > \hat{f}(x_{\text{sel}}) - t\}|$ count candidates within t of the nomination's proxy score. These candidates form a prefix of the proxy ranking.

Theorem 1 (Global certification bound)**.** *If $G_{\text{verifier}} > 0$, every witness belongs to $\{x : \hat{f}(x) > \hat{f}(x_{\text{sel}}) - 2\epsilon_\infty\}$. Consequently $R \leq N(2\epsilon_\infty)$, and top-$B$ probing with $B \geq N(2\epsilon_\infty)$ certifies every selection failure.*

The $2\epsilon_\infty$ window accounts for overestimating the nomination and underestimating a witness. The bound sharpens to $R \leq N(\epsilon_- + e_{\text{sel}})$ when $G_{\text{verifier}} > 0$, where $\epsilon_- = \max_x[f(x) - \hat{f}(x)]$ is the largest trusted-minus-proxy difference in the pool, and $e_{\text{sel}} = \hat{f}(x_{\text{sel}}) - f(x_{\text{sel}})$ is the proxy-minus-trusted difference at the nomination (Appendix A.1, Proposition 3). Candidate-specific error bounds can define smaller sets to probe, but their size need not bound proxy rank (Appendix A.1, Proposition 4).

Some pools require exactly the number of evaluations given by the global bound, and for a fixed proxy and uniform error budget, proxy ordering minimizes this worst-case cost over failed nominations among fixed deterministic orderings that use no trusted pool scores and evaluate the nomination first (Appendix A.1, Proposition 5). Other orderings can attain the same bound, so their average audit costs must be compared empirically. These bounds use whole-pool error information but can greatly exceed observed ranks: in the controlled suite, the median ratio of the global bound to the observed rank is $4{,}483.8$ (Appendix E).

3.4 WHAT RANK CORRELATION DOES NOT DETERMINE

Let $k \geq 2$ be the number of top candidates, chosen by trusted score, and ρ_k their Spearman correlation between f and $\hat{f}$. This commonly used tail metric evaluates internal ordering but ignores which outside candidates the proxy places above the tail.

Proposition 2 (Top-tail correlation does not identify audit cost)**.** *For integers $k \geq 2$ and $N > k + 1$, and any $\Delta > 0$, there are a trusted score and two proxy rankings with the same suboptimal nomination and $\rho_k = 1$ for which the first Δ-witness has rank 2 in one ordering and $N - k + 1$ in the other.*

The construction, given in Appendix A.1, leaves the trusted top tail perfectly ordered while moving non-witness outsiders around it. Thus even perfect top-tail correlation can coexist with large selection regret and, when k is small relative to N, a large witness rank. Section 5.3 therefore reports nomination-level outcomes alongside correlation.

3.5 RETAINING OR REPLACING THE NOMINATION

Delivery rule. Probing the proxy's top B candidates both tests the nomination and measures candidates that could replace it. Let $x_{\text{safe}}(B)$ be the best measured candidate, with ties favoring the nomination, and G_{residual} its remaining regret relative to the pool optimum:

$$x_{\text{safe}}(B) = \arg\max_{x \in S_B} f(x), \qquad f(x_{\text{safe}}) - f(x_{\text{sel}}) = L_B, \qquad G_{\text{residual}} = G_{\text{verifier}} - L_B. \quad (6)$$

Returning x_{safe} gains exactly $L_B \geq 0$ and leaves $G_{\text{residual}} \in [0, G_{\text{verifier}}]$. This standard best-observed rule retains the nomination as the incumbent (current best candidate), preventing a loss when every new probe scores below it; Section 6 measures this gain.

When no witness is found. When the certificate is zero, the nomination is retained, but untested candidates may still be witnesses. Simultaneous upper bounds on trusted scores could exclude such candidates. For one-sided split conformal prediction with a union-bound correction over $N = |\mathcal{X}|$ candidates, a finite bound at confidence $1 - \delta$ requires $m_{\text{cal}} \geq \lceil N/\delta \rceil - 1$ calibration residuals (Appendix B.1, Corollary 1). A *residual* is the trusted score minus the proxy prediction, $f(x) - \hat{f}(x)$. At $\delta = 0.1$, TF-Bind-8 (Barrera et al., 2016; Trabucco et al., 2022) requires 593,919 calibration residuals, but only 2,048 are available. This construction returns infinite upper bounds on all four landscapes and therefore excludes no candidates (Appendix B).

We instead bound witness prevalence in the pool. Draw m candidates uniformly at random without replacement, independently of the proxy and S_B. The witness count K follows a hypergeometric distribution with unknown pool witness total W. Writing $\Pr$ for probability, exact tail inversion for an observed count k gives the one-sided upper bound at confidence $1-\alpha$ $\widehat{p}_{1-\alpha} = |\mathcal{X}|^{-1} \max\{W : \Pr(K \le k \mid W, m, |\mathcal{X}|) > \alpha\}$ with no assumption on f, on $\hat{f}$, or on their relationship. The bound quantifies witness prevalence rather than identifying the pool optimum.

In the paired study of 365 audits at $\Delta = 0$, the random-probe upper bound was below one in 99.85% of draws. Empirical coverage, the fraction of draws whose bound includes the true prevalence, met the requested confidence level for every landscape, budget, and confidence setting. At $m = 16$, the upper bound was two to fifteen times the measured prevalence (Appendix B.3). Prefix probing certified 88.00% to 95.24% of instances, and random probes bounded the remainder. In Appendix B.3, Table 4 reports the two audit outcomes across the acquisition campaigns.

4 AUDITED SYSTEMS AND PROTOCOL

A landscape maps every sequence in a library to a measured value; its declared pool is fixed before fitting any proxy. We use GB1 (Wu et al., 2016), TrpB (Johnston et al., 2024), TF-Bind-8 (Barrera et al., 2016; Trabucco et al., 2022), and avGFP (Sarkisyan et al., 2016). We exclude entries whose missing scores were filled with estimates (imputed entries), resolve duplicates and conflicts before splitting, and require disjoint training, validation, and selection sets. The resulting selection pools contain 143,781, 152,929, 59,392, and 45,570 candidates, respectively; confidence intervals resample splits. TF-Bind-10 (Trabucco et al., 2022) is excluded because its dataset files store every sequence five to eight times with conflicting labels, making the optimum depend on the order in which duplicates are processed (Appendix C). Recovering the full distribution of R_Δ requires trusted values across each pool, so these landscape audits are retrospective.

We audit four proxy families: pairwise or additive ridge regression, an ExtraTrees ensemble on token features (Geurts et al., 2006; Pedregosa et al., 2011), a pairwise ranking model, and a conservative objective surrogate (Trabucco et al., 2021), giving 200 margin-aware audits. We separately reproduce 30 runs of two published offline optimizers at their full training schedules: RaM-ListNet (Tan et al., 2025), which uses the ListNet ranking loss (Cao et al., 2007), and Conservative Objective Models (COMs) (Trabucco et al., 2021) (Appendix G). Since their original searches can propose candidates outside the declared pool, we distinguish the exact proxy optimum within the pool, the best proposal within it, and the actual output. We evaluate certification with budgets of $B \in \{8, 16, 32, 64, 128, 256\}$ trusted evaluations, always counting the nomination, and run 7,300 matched-budget multi-round acquisition campaigns (Appendices E, F, and F.1), with paired percentile bootstraps (Efron & Tibshirani, 1993) over splits, which within a landscape are repeated splits of one table rather than independent replicates. The controlled mixture family (Appendix D) lets us vary *training coverage*, the ratio of training examples to pairwise-proxy features, independently of the reference objective's additive and interaction terms, supporting the mechanism analyses in Appendices K and L. The measured landscapes provide the main empirical evidence. An independent implementation recomputes certificates and deliveries from saved predictions and trusted values, and checks split integrity, eligibility rules, and budget accounting.

Reporting conventions. We report numerical summaries to at most three decimals, rescaling small rates to preserve nonzero values and retaining exact generator parameters for reproducibility. For the landscape analyses, we set $\Delta = 0.05\,\mathrm{SD}$ of trusted values on the selection pool and use $\Delta = 0$ only as a labeled limiting case. Certification recall is the fraction of all audits that certify an improvement, including audits without witnesses in the denominator and giving each audit equal weight. Witness prevalence is the mean, across audits, of the pool fraction above $f(x_{\mathrm{sel}}) + \Delta$. Because trusted scales differ by landscape, G_{verifier}, L_B, and incumbent gains are not compared across landscapes. Quantities requiring trusted scores hidden from the algorithm are marked as oracle diagnostics and used only for retrospective analysis; Appendix C gives complete preprocessing and reporting details.

Table 1: **Sixteen trusted evaluations expose most meaningful selection failures.** Quantities use $\Delta = 0.05\,\mathrm{SD}$ and the equal-weight average over audits defined in Section 4; median ranks are conditional on a witness existing. Strict-improvement results ($\Delta = 0$) for the 200 original audits appear in Appendix E.1, Table 8.

Landscape	Audits	$\Pr(R_\Delta < \infty)$	Median R_Δ	Recall@16 / @64
TF-Bind-8	70	0.986	5	**0.786** / 0.943
TrpB	70	1.000	3	**0.971** / 1.000
GB1	70	1.000	3	**0.914** / 0.971
avGFP	20	0.950	5	**0.800** / 0.900

The 200 original audits across landscapes and models are detailed in Appendix E.1, Table 8; the remaining 30 reproduce published optimizers (Appendix G). At the headline margin, one TF-Bind-8 audit and one avGFP audit have no witness.

5 Results

5.1 Selection failure across all four landscapes

Table 1 summarizes failure rates, witness ranks, and certification recall across the four measured landscapes.

Across the four sequence landscapes and all audited proxy families, the proxy maximum is strictly worse than the 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. Only five of 230 reach the trusted top five. Selection regret varies across proxy families within each landscape, and the worst proxy family changes across landscapes, indicating a general selection effect rather than a failure of one modeling choice.

5.2 Small witness ranks

Although the selection pools contain 45,570 to 152,929 candidates, the conditional median witness rank is only three to five: a candidate near the top of the proxy ranking typically improves on the nomination by more than 0.05 SD. Sixteen trusted evaluations certify 78.6% to 97.1% of audits, and 64 evaluations yield recall between 90.0% and 100%. Most finite ranks are small, with a few much larger values (Figure 2(a)).

Ordering within the shortlist matters. 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. Shuffling only within the proxy's top 1% reduces recall@16 to 0.633 on GB1, 0.492 on TF-Bind-8, and 0.715 on TrpB, while avGFP remains at 0.807. Ordering within this shortlist therefore matters on three landscapes; on avGFP, shortlist membership matters more than internal order. Across 7,300 matched-budget campaigns, all four structured policies achieve higher certification rates than random probing in each landscape, while the best-performing policy varies across landscapes (Appendix F.1, Table 10). Split intervals describe variation within each landscape.

5.3 Small conditional ranks persist under controlled high-quality orderings

Across 525 additional combinations of model and data split, conditional median ranks remain between two and six. Three fixed, deployable model changes, including a seven-model ensemble and frozen ESM-2 embeddings (Lin et al., 2023), raise the best selection-independent top-tail Spearman correlation from 0.383 to 0.402. An oracle stress test examines higher-quality rankings.

Top-tail correlation alone cannot determine nomination reliability (Proposition 2), so we report both correlation and audit outcomes. The controlled study mixes trusted ranks with independent noise, tunes the noise level to five target correlations on validation data, and evaluates achieved quality and witness rank on the disjoint selection pool. This oracle construction isolates the effect of ranking quality. At achieved selection-pool top-tail correlation around 0.95, conditional median ranks remain two to four in every landscape (Table 2). The fraction of these audits with a Δ-witness falls to 0.468

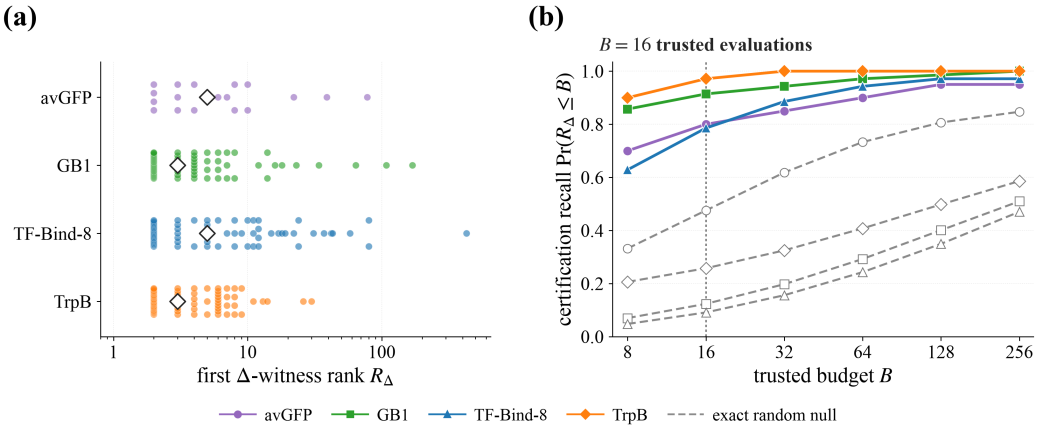

Figure 2: **Certification cost is typically small on measured landscapes.** *(a)* Finite margin-aware witness ranks from the 230 audits at $\Delta = 0.05\,\mathrm{SD}$, on a logarithmic axis; open diamonds show medians conditional on a witness existing. Audits without a Δ-witness have $R_\Delta = \infty$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

Table 2: **High-quality controlled rankings retain small conditional witness ranks.** Rows aggregate oracle rankings with achieved selection-pool top-tail Spearman correlation of at least 0.875, using $\Delta = 0.05\,\mathrm{SD}$. P90 denotes the 90th percentile of finite witness ranks; recall is shown 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

on TF-Bind-8 and 0.646 on avGFP but remains above 0.92 on GB1 and TrpB. Thus, in this controlled study, higher proxy-ranking quality can make meaningful failures rarer while many remaining failures still require few evaluations to expose.

We also evaluate the 230 headline audits at larger improvement margins. Increasing Δ from 0.05 to 0.5 SD lowers witness existence and unconditional recall most sharply on avGFP and TF-Bind-8; GB1 and TrpB retain witnesses in every audit, with median conditional ranks at most six. Larger margins therefore mostly change whether a qualifying improvement exists, not merely how deep it lies. The complete sweep over $\Delta/\mathrm{SD} \in \{0, 0.05, 0.10, 0.25, 0.50\}$ is in Appendix E.2, Table 9.

5.4 The same probes repair most of the regret

Delivering the best probed candidate removes most of the selection regret in the measured landscapes. In a separate check of 248 failure instances, an independent calculation reproduced the certified improvements and the scores of the returned candidates. At $B = 256$, the fraction of selection regret removed, $L_B/G_{\mathrm{verifier}}$, averages 0.866 on TF-Bind-8, 0.831 on avGFP, 0.803 on GB1, and 0.724 on TrpB. Paired 95% intervals for the improvement over random probing exclude zero on all four landscapes (Figure 3). The same measurements establish the nomination's failure and supply its replacement. Section 6 separately measures the gain from retaining the nomination, using the best new probe as its baseline.

Prospective electrolyte study. To illustrate auditing and repair beyond sequence design, we conducted a prospective study of battery electrolyte mixtures. We used GeoMix (Li et al., 2025) and SMI-TED-IC (Zohair et al., 2025) as proxies for ionic conductivity in five carbonate-based and

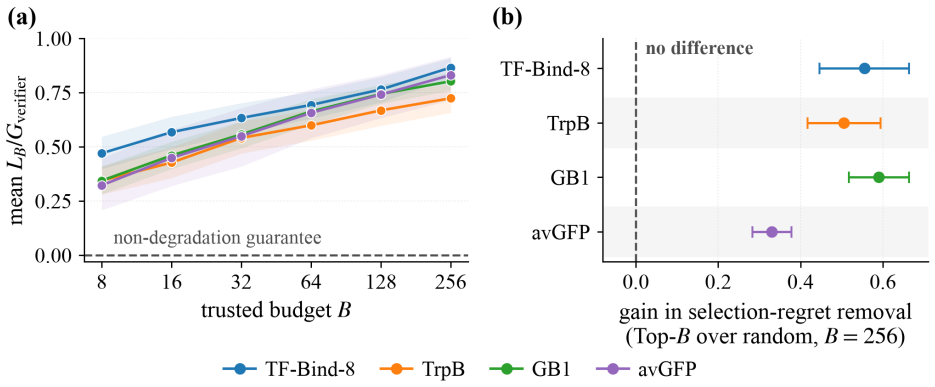

Figure 3: **The audit probes both certify and repair selection failure.** *(a)* Mean fraction of selection regret removed, L_B/G_{verifier}, against trusted budget, with 95% split-bootstrap ribbons and the non-degradation floor at zero. *(b)* Paired top-B-minus-random removal difference at $B = 256$, excluding zero on all four landscapes.

five ether-based formulations, fixing predictions before laboratory measurements. Three of the four nominations, one for each combination of model and candidate pool, are overturned: the observed ranks R_0 are two for GeoMix on carbonates and SMI-TED-IC on ethers, and five for GeoMix on ethers. Each first witness is the best measured candidate in its pool; SMI-TED-IC nominates the best carbonate directly. All candidates were measured, and ranks count prefix evaluations including the nomination. Appendix O reports the full results and experimental methods.

5.5 Published offline optimizers are not exempt

We also evaluate RaM-ListNet (Tan et al., 2025) and Conservative Objective Models (COMs) (Trabucco et al., 2021) at their full published schedules, with search restricted to the declared pool. Restricting RaM-ListNet's proposals and COMs' candidate updates makes all 30 outputs eligible without projection after optimization. RaM-ListNet has median witness ranks of 3 to 8 and top-16 certification rates of 0.6 to 1.0; COMs reaches a median rank of 27 on GB1, with top-16 rates of 0.0 to 0.2 across landscapes. For COMs, the remaining proxy-score gap reflects incomplete search, distinct from selection regret at the exact proxy optimum. Audit reports should therefore identify both the candidate being tested and the probe ordering (Appendix G).

6 Preserving performance under any acquisition policy

Incumbent-safe delivery applies to any acquisition policy or number of rounds, provided S_B includes the nomination; examples include probing in proxy order, Thompson sampling (Thompson, 1933), and batch Bayesian optimization (BO). We measure its gain over returning the best new observation alone.

Across 7,300 campaigns, 82.86% produce a positive certificate and deliver an improvement; the remaining 17.14% retain the incumbent while independent random probes bound pool witness prevalence (Appendix B.3, Table 4). For batch BO, the mean incumbent gain is 0.031 $[0.015,\ 0.048]$ on GB1 and 0.013 $[0.005,\ 0.022]$ on TrpB, while Thompson sampling shows a measurable incumbent gain only on GB1 (Appendix F.1).

Refitting on the same labels and renominating outside the probed set produces 840 deliveries below the incumbent; retaining it and restricting delivery to the measured set removes every violation (Appendix F). Across all 16 combinations of landscape and acquisition schedule, the highest mean final improvement comes from an adaptive policy: Thompson sampling on GB1 and TrpB, batch BO on TF-Bind-8, and witness-local probing on avGFP.

7 DISCUSSION

Witness rank complements predictive and optimization metrics by counting the trusted evaluations needed to overturn a realized nomination. In the measured landscapes, this cost remains small even though the declared pools contain 45,570 to 152,929 candidates. The contrast arises because falsification requires one counterexample rather than identification of the pool optimum. It cannot be inferred from top-tail correlation alone, as Proposition 2 establishes and the controlled quality ladder confirms on the evaluated pools.

Rank-shuffling shows that ordering within the proxy shortlist reduces audit cost on three landscapes, beyond the benefit of shortlist membership alone. The required improvement also matters: as Δ grows, a qualifying witness often ceases to exist on avGFP and TF-Bind-8, whereas GB1 and TrpB retain small conditional ranks through 0.5 SD. Audit cost must therefore be reported together with witness existence, and the scaling constructions in Appendix H caution that small ranks are an empirical property of these landscapes rather than a consequence of finite pool size.

Additional optimization can reduce G_{search}, while experiments on the controlled mixture family introduced in Section 4 separate selection regret due to limited data from error due to missing model structure: more training data eliminate mean selection regret for a correctly specified pairwise proxy, but regret persists when cubic interactions are omitted (Appendix K). Appendix L isolates the effects of interaction strength and the representation of decision-relevant component pairs in the training data.

A held-out mixture benchmark provides an in-domain check (Appendix N), while predictors of selection failure can help choose which nominations to audit but cannot certify failure for an individual nomination (Appendix M). Cross-landscape allocation and held-out budget prediction remain open: neither was reliable in our tests (Appendix J). Practitioners choose Δ from application requirements and B from experimental resources; observed rank distributions do not guarantee sufficient budgets on new landscapes.

A design campaign can report the nomination, label-free probe ordering, trusted budget, and certificate alongside its usual predictive and optimization metrics. Retain the nomination unless a measured probe is better; if none is, independent random probes can bound pool witness prevalence. This distinguishes measured improvements from nominations unrefuted at the available budget. Finally, the audit can be added to existing pipelines without retraining or replacing their optimizer.

8 LIMITATIONS

All exact claims are relative to a declared finite pool. Splits are repeated partitions of one measured table, not independent biological replicates; the four landscapes, not the 7,300 campaigns, determine the scope of cross-domain generalization. The observed small witness ranks are not guaranteed, and the main analysis assumes a noiseless evaluator, with 11 noisy-evaluation settings examined in Appendix I. With fixed deployable proxies reaching at most 0.402 top-tail Spearman correlation, the higher-quality study uses oracle rankings to test the mechanism rather than real-model performance. The large-scale landscape audits are retrospective; the prospective electrolyte study illustrates the protocol but does not establish generalization across chemical families or candidate-pool sizes.

9 CONCLUSION

Predictive accuracy asks how well a proxy scores candidates; witness rank asks how much trusted evidence its realized decision can withstand. On the measured landscapes, structured policies place meaningful counterexamples early in the probe ordering, a pattern that persists conditionally under controlled high-quality rankings and weakens predictably as the required improvement grows. Even when a design pool exceeds 150,000 candidates, a proxy-guided recommendation can often be falsified with only three to five trusted evaluations. Reporting witness existence, witness rank, and the required margin makes the audit cost explicit. Retaining the nomination unless a measured candidate is better then connects that evidence to the next experimental decision.

AI USE STATEMENT

Generative AI tools were used to proofread and edit the manuscript for clarity and readability, assist with mathematical proofs, and improve the visual quality of selected figures. They were not used to generate or alter the experimental measurements. The authors reviewed all AI-assisted material and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendix A states the assumptions and proves the theoretical claims, with calibration details in Appendix B. Appendices C and D document dataset provenance, cleaning, splits, candidate construction, and the training and search protocols. Appendix O reports wet-lab candidate measurements and predictions. The remaining experimental appendices provide results and robustness analyses.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

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

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.

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*, pp. 129–136, 2007.

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.

Paolo de Blasio, Jonas Elsborg, Tejs Vegge, Eibar Flores, and Arghya Bhowmik. CALiSol-23: Experimental electrolyte conductivity data for various Li-salts and solvent combinations. *Scientific Data*, 11(1):750, 2024.

Priya 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.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1993.

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

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.

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

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

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.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Anyi Li, Jiacheng Cen, Songyou Li, Mingze Li, Yang Yu, and Wenbing Huang. Geometric mixture models for electrolyte conductivity prediction. *Advances in Neural Information Processing Systems*, 38, 2025.

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.

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.

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.

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. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Ella Miray Rajaonson, Mahyar Rajabi Kochi, Luis Martin Mejia Mendoza, Seyed Mohamad Moosavi, and Benjamin Sanchez-Lengeling. CheMixHub: Datasets and benchmarks for chemical mixture property prediction. *Advances in Neural Information Processing Systems*, 38, 2025.

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

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.

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.

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.

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.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

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

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*, pp. 21658–21676. PMLR, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

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.

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

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.

Shang Zhu, Bharath Ramsundar, Emil Annevelink, Hongyi Lin, Adarsh Dave, Pin-Wen Guan, Kevin Gering, and Venkatasubramanian Viswanathan. Differentiable modeling and optimization of non-aqueous li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15(1):8649, 2024.

Murtaza Zohair, Vidushi Sharma, Eduardo A Soares, Khanh Nguyen, Maxwell Giammona, Linda Sundberg, Andy Tek, Emilio Vital Brazil, and Young-Hye La. Chemical foundation model-guided design of high ionic conductivity electrolyte formulations. *npj Computational Materials*, 11(1): 283, 2025.

APPENDIX CONTENTS

The sections below follow the order of the appendix.

Theory and proofs		
A	Formal results	Proofs of the audit and rank bounds, worst-case optimality, noisy certificates, and the comparison with best-arm identification
B	Calibration and random probes	Calibration requirements, candidate-exclusion bounds, and witness prevalence from random samples
Data and experimental setup		
C	Datasets, splits, and preprocessing	Data sources, cleaning, disjoint splits, and the exclusion of TF-Bind-10
D	Controlled mixture family	Candidate generation, generator fitting, sampling, proxies, and matched searches
Full audit and repair results		
E	Full audit results	Complete audits, proxy quality, improvement margins, rank-shuffling, and probe policies
F	Certified repair	Regret removal, label reuse, adaptive campaigns, and the benefit of retaining the nomination
G	Published offline optimizers	Full training schedules, reference runs, and whether proposed candidates belong to the declared pool
Scaling, noise, and extensions		
H	Pool size and witness rank	Scaling from one thousand to five million candidates and the distribution of witnesses
I	Noisy certificates	False-certification rates and recall under eleven noisy-evaluation settings
J	Resource allocation and audit-budget prediction	Limits of choosing between search and training and predicting audit budgets across domains
Mechanisms, diagnostics, and benchmark checks		
K	Training data and model structure	Effects of additional training data and omitted interaction terms
L	Mechanism studies in the controlled family	Interaction strength, training size, targeted data removal, optimization pressure, and robustness
M	Predictive diagnostics for selection failure	Predicting failure risk and regret without certifying an individual failure
N	Held-out mixture-benchmark audit	Evaluation on held-out benchmark candidates
Wet-lab results		
O	Prospective wet-lab electrolyte audits	Prospective protocol, materials, ionic conductivity measurements, model predictions, and witness ranks

A FORMAL RESULTS

This appendix proves the audit, rank, and replacement claims in Section 3 and derives the noisy certificate used in Appendix I.

A.1 PROOFS FOR SECTION 3

Audit identity and uniform bounds. Recall that f is the trusted score, $\hat{f}$ is the proxy score, and x_{sel} is the proxy's nomination. The largest absolute proxy error in the pool is $\epsilon_\infty = \max_{x \in \mathcal{X}} |f(x) - \hat{f}(x)|$.

Proof of Proposition 1. Using the gaps defined in Equations (1) and (2) and the definition of $H(A)$ in Proposition 1,

$$\begin{aligned} G_{\text{verifier}} - H(A) &= f(x_{\text{true}}) - f(x_{\text{sel}}) - \{f(\hat{x}_A) - f(x_{\text{sel}})\} \\ &= f(x_{\text{true}}) - f(\hat{x}_A) = G_{\text{total}}(A). \end{aligned}$$

For the bound, add and subtract the proxy scores at both candidates:

$$\begin{aligned} G_{\text{total}}(A) &= f(x_{\text{true}}) - \hat{f}(x_{\text{true}}) + \hat{f}(x_{\text{true}}) - \hat{f}(\hat{x}_A) + \hat{f}(\hat{x}_A) - f(\hat{x}_A) \\ &\le \epsilon_\infty + \hat{f}(x_{\text{sel}}) - \hat{f}(\hat{x}_A) + \epsilon_\infty = G_{\text{search}}(A) + 2\epsilon_\infty. \end{aligned}$$

Here $\hat{f}(x_{\text{true}}) \le \hat{f}(x_{\text{sel}})$ because the nomination maximizes the proxy. Setting $\hat{x}_A = x_{\text{sel}}$ makes $G_{\text{search}} = 0$ and gives $G_{\text{verifier}} \le 2\epsilon_\infty$.

To see why the factor two cannot be reduced, let $\epsilon > 0$ and $\mathcal{X} = \{x_1, x_2\}$ with

$$f(x_1) = 0, \quad f(x_2) = 2\epsilon, \qquad \hat{f}(x_1) = \hat{f}(x_2) = \epsilon.$$

Both absolute errors equal ϵ. Choose x_1 to be the candidate selected by the fixed tie-breaking rule; then $G_{\text{verifier}} = 2\epsilon_\infty$, exactly the bound. To remove the tie, replace $f(x_2)$ by $2\epsilon - \eta$ and $\hat{f}(x_2)$ by $\epsilon - \eta$, where $0 < \eta < \epsilon$. The nomination is still x_1 and $\epsilon_\infty = \epsilon$, while $G_{\text{verifier}} = 2\epsilon - \eta$ approaches $2\epsilon_\infty$ as η decreases to zero. □

Top-tail correlation and audit cost.

Proof of Proposition 2. Give k candidates distinct trusted scores above Δ, set the nomination's score to $f(x_{\text{sel}}) = 0$, and give the remaining $N - k - 1$ candidates scores below zero. In both proxy rankings, put x_{sel} first and keep the k highest trusted-score candidates in their true relative order, so $\rho_k = 1$. Placing these k witnesses immediately after the nomination gives $R_\Delta = 2$; placing all $N - k - 1$ non-witnesses before them gives $R_\Delta = N - k + 1$. □

Noisy trusted evaluations. Suppose each 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 probe set S_B, including the nomination, and each candidate's repeat count $r_x \ge 1$ before collecting these measurements. For an error probability $0 < \delta < 1$ and $m = |S_B| - 1 \ge 1$ other candidates, define

$$L_B^{\text{Gauss}} = \max\left\{0, \max_{x \in S_B \setminus \{x_{\text{sel}}\}} \left[\bar{Y}_x - \bar{Y}_{x_{\text{sel}}} - z_{1-\delta/m}\sigma\sqrt{r_x^{-1} + r_{x_{\text{sel}}}^{-1}}\right]\right\}. \tag{7}$$

Here $\bar{Y}_x$ is the mean of candidate x's measurements and z_q is the standard-normal quantile at probability q. Each expression in brackets is a lower bound on $f(x) - f(x_{\text{sel}})$ with error probability δ/m. A union bound ensures that all m bounds hold together with probability at least $1 - \delta$, even though they share the nomination's measured mean. Their maximum, or zero if all are negative, therefore satisfies $0 \le L_B^{\text{Gauss}} \le G_{\text{verifier}}$ with that probability. Appendix I evaluates this certificate and ten variants; sequential measurement requires separate control of the error probability across repeated checks.

Witness-rank bounds from proxy errors. Recall that a witness has $f(w) > f(x_{\text{sel}})$ and $R = R_0$ is the position of the first witness in decreasing proxy-score order. Error bounds restrict which candidates can be witnesses.

Let $N(t) = |\{x \in \mathcal{X} : \hat{f}(x) > \hat{f}(x_{\text{sel}}) - t\}|$ count the candidates whose proxy score falls strictly within t of the nomination's. These are the first $N(t)$ candidates in proxy order.

Proof of Theorem 1. Assume $G_{\text{verifier}} > 0$ and let w be any witness, so $f(w) > f(x_{\text{sel}})$. Then

$$\hat{f}(w) \geq f(w) - \epsilon_\infty > f(x_{\text{sel}}) - \epsilon_\infty \geq \hat{f}(x_{\text{sel}}) - 2\epsilon_\infty.$$

The first and last inequalities use $|f(x) - \hat{f}(x)| \leq \epsilon_\infty$ for every candidate. Thus every witness is among the first $N(2\epsilon_\infty)$ candidates in proxy order, giving $R \leq N(2\epsilon_\infty)$. □

The bound can be reduced by using the proxy's actual error at the nomination. Recall

$$\epsilon_- = \max_{x \in \mathcal{X}} [f(x) - \hat{f}(x)], \qquad e_{\text{sel}} = \hat{f}(x_{\text{sel}}) - f(x_{\text{sel}}).$$

Here ϵ_- is the largest trusted-minus-proxy difference in the pool, and e_{sel} is the proxy-minus-trusted difference at the nomination. Either can be negative; their sum is nonnegative. Define $\mathcal{C}_-$ to retain candidates above the resulting proxy-score threshold, together with the nomination:

$$\mathcal{C}_- = \{x : \hat{f}(x) > \hat{f}(x_{\text{sel}}) - (\epsilon_- + e_{\text{sel}})\} \cup \{x_{\text{sel}}\}. \tag{8}$$

Proposition 3 (Refined witness-rank bound)**.** *The set $\mathcal{C}_-$ in Appendix A.1 (Equation* (8)*) contains every witness, and $0 \leq \epsilon_- + e_{\text{sel}} \leq 2\epsilon_\infty$. If $G_{\text{verifier}} > 0$, then $R \leq |\mathcal{C}_-| \leq N(2\epsilon_\infty)$.*

Proof. For any witness,

$$\hat{f}(w) \geq f(w) - \epsilon_- > f(x_{\text{sel}}) - \epsilon_- = \hat{f}(x_{\text{sel}}) - (\epsilon_- + e_{\text{sel}}),$$

so $w \in \mathcal{C}_-$. Also $\epsilon_- \geq f(x_{\text{sel}}) - \hat{f}(x_{\text{sel}}) = -e_{\text{sel}}$, while $\epsilon_- \leq \epsilon_\infty$ and $e_{\text{sel}} \leq \epsilon_\infty$, giving $0 \leq \epsilon_- + e_{\text{sel}} \leq 2\epsilon_\infty$. If a witness exists, the displayed strict inequality and $\hat{f}(w) \leq \hat{f}(x_{\text{sel}})$ imply $\epsilon_- + e_{\text{sel}} > 0$. Then $\mathcal{C}_-$ consists of the first $N(\epsilon_- + e_{\text{sel}})$ candidates, including the nomination, and is contained in the first $N(2\epsilon_\infty)$ candidates. This proves both rank inequalities. □

Candidate-specific error bounds. Suppose the nomination has been measured and $u(x)$ is an error bound satisfying $f(x) \leq \hat{f}(x) + u(x)$ for every candidate. Only candidates whose upper score bounds exceed the nomination's measured score can be witnesses. Retaining these candidates and the nomination gives

$$\mathcal{C}_u = \{x : \hat{f}(x) + u(x) > f(x_{\text{sel}})\} \cup \{x_{\text{sel}}\}. \tag{9}$$

Proposition 4 (Candidate-specific error bounds)**.** *The set $\mathcal{C}_u$ in Appendix A.1 (Equation* (9)*) contains every witness. If a witness exists, evaluating this set finds one within $|\mathcal{C}_u|$ trusted evaluations, including the nomination. Taking $u(x) = \epsilon_-$ for every candidate gives $\mathcal{C}_u = \mathcal{C}_-$ and hence $R \leq |\mathcal{C}_u|$ whenever a witness exists. When u varies across candidates, this proxy-rank bound need not hold.*

Proof. For any witness, $\hat{f}(w) + u(w) \geq f(w) > f(x_{\text{sel}})$, so $w \in \mathcal{C}_u$. Taking $u(x) = \epsilon_-$ makes the threshold $\hat{f}(x) > f(x_{\text{sel}}) - \epsilon_- = \hat{f}(x_{\text{sel}}) - (\epsilon_- + e_{\text{sel}})$, exactly the definition of $\mathcal{C}_-$. The example below shows why the rank bound can fail when u varies. □

Unlike $\mathcal{C}_-$, the set $\mathcal{C}_u$ can skip higher-ranked candidates while retaining lower-ranked ones. Let $\mathcal{X} = \{x_1, x_2, x_3\}$ with

$$\hat{f} = (10,\ 0.5,\ 0.1), \qquad f = (5,\ 0,\ 6), \qquad u = (0,\ 0,\ 6).$$

These are valid upper bounds because $\hat{f} + u = (10, 0.5, 6.1) \geq f$ at every candidate. The nomination is x_1 and the only witness is x_3, giving $R = 3$ in proxy order. Yet $\mathcal{C}_u = \{x_1, x_3\}$ finds the witness in two evaluations by skipping x_2. For comparison, $\epsilon_- = 5.9$ and $e_{\text{sel}} = 5$ give $\mathcal{C}_- = \mathcal{X}$, whose size correctly bounds R. Thus $|\mathcal{C}_u|$ bounds the cost of evaluating that set, not necessarily the cost of

following the original proxy order. Constructing these bounds by split conformal prediction with a whole-pool union bound requires the calibration size derived in Appendix B.

We checked the witness-containing sets by enumerating pools of two through five candidates and generating random finite cases. Both bounds also retained every witness across 3,229 nominations from the controlled, measured, and published-optimizer studies. Among failures, the median ratio of set size to witness rank was 7,471.7 for the global bound and 2,802.4 for $\mathcal{C}_-$, which retained 35.6% fewer candidates on average. These are retrospective comparisons using complete trusted scores, which are needed to compute ϵ_-.

Why the global bound cannot be improved.

Proposition 5 (Worst-case optimality of proxy ordering)**.** *Fix $\hat{f}$ on $\mathcal{X}$ and an error limit $\epsilon > 0$, with $n = N(2\epsilon) \geq 2$. There is a trusted function f with $\epsilon_\infty = \epsilon$ whose only witness has proxy rank n, so proxy ordering requires exactly the n evaluations in the bound of Theorem 1. For every fixed deterministic ordering that uses no trusted pool scores and evaluates the nomination first, some trusted function satisfying $\epsilon_\infty \leq \epsilon$ and $G_{\text{verifier}} > 0$ places its first witness at position at least n. Hence no such ordering guarantees a smaller worst-case budget than proxy ordering.*

Proof. Let T contain the first n candidates in proxy order. For any $w \in T \setminus \{x_{\text{sel}}\}$, set $f(w) = \hat{f}(w) + \epsilon$ and $f(x) = \hat{f}(x) - \epsilon$ for every other candidate. All absolute errors are exactly ϵ, so $\epsilon_\infty = \epsilon$. Because $w \in T$, its proxy score satisfies $\hat{f}(w) > \hat{f}(x_{\text{sel}}) - 2\epsilon$, giving

$$f(w) = \hat{f}(w) + \epsilon > \hat{f}(x_{\text{sel}}) - \epsilon = f(x_{\text{sel}}).$$

Every other candidate has $f(x) = \hat{f}(x) - \epsilon \leq \hat{f}(x_{\text{sel}}) - \epsilon = f(x_{\text{sel}})$, so w is the only witness. Choosing the candidate at proxy rank n as w gives $R = n = N(2\epsilon_\infty)$.

Now fix any other permitted ordering. The last of the n members of T must appear at position at least n and cannot be the nomination, which comes first. Choosing that candidate as w in the same construction proves the lower bound for this ordering. Proxy ordering guarantees at most n evaluations whenever a witness exists and $\epsilon_\infty \leq \epsilon$, by Theorem 1. □

Retaining or replacing the nomination. Since $x_{\text{sel}} \in S_B$, the best measured candidate has score at least $f(x_{\text{sel}})$ and improves it by exactly L_B. Subtracting this improvement from G_{verifier} gives the remaining regret $G_{\text{residual}} = G_{\text{verifier}} - L_B$ in Equation (6); ties retain the nomination.

A.2 Certification versus best-arm identification

Certifying failure can require far fewer evaluations than identifying the trusted optimum. Suppose the proxy's rank-two candidate has a higher trusted score than the nomination and all other candidates score below it. Measuring rank two gives $L_B > 0$, so one evaluation beyond the nomination certifies failure. Without further assumptions, however, any unmeasured candidate could have a higher score. Identifying the exact optimum therefore requires measuring all $|\mathcal{X}| - 1$ other candidates. This distinction parallels stochastic best-arm identification, whose difficulty depends on score differences across the pool (Even-Dar et al., 2006; Audibert et al., 2010).

B Calibration and random probes

Calibration uses held-out prediction errors (residuals) to set bounds on trusted scores. Random probes are candidates sampled uniformly from the pool and evaluated with the trusted method. Appendix B.1 (Corollary 1) derives the calibration requirement used in Section 3.5, and Appendix B.2 tests two refinements. Neither usefully excludes candidates with a guarantee in our studies, so Appendix B.3 instead uses random probes to bound the fraction that beat the nomination.

The split-conformal construction uses a union-bound correction for *simultaneous coverage*: all candidate scores lie below their respective upper bounds at once, with the stated probability. When this holds, candidates whose upper bounds do not exceed the nomination's score can be excluded without losing a witness.

Table 3: **Finite bounds across the full pool require more calibration data.** For split conformal prediction with a union-bound correction at $\delta = 0.1$, we compare available and required numbers of held-out prediction errors (calibration size m). The last column gives the largest supported pool, $N_{\max} = \lfloor \delta(m+1) \rfloor$.

Landscape	Selection pool N	Available m	Required m	Largest supported pool
TF-Bind-8	59,392	2,048	593,919	204
TrpB	152,929	1,240	1,529,289	124
GB1	143,781	620	1,437,809	62
avGFP	45,570	2,048	455,699	204

B.1 Calibration data required for finite bounds

Corollary 1 (Finite-calibration barrier)**.** *Build a one-sided bound from m calibration residuals by split conformal prediction, which returns the k-th smallest residual with $k = \lceil (m+1)(1-\alpha) \rceil$ at level α, and $+\infty$ when k exceeds m. Setting $\alpha = \delta/N$ so that the bound holds at all N candidates at once with probability $1-\delta$, the returned value is an observed residual rather than $+\infty$ if and only if $m + 1 \geq N/\delta$, equivalently $m \geq \lceil N/\delta \rceil - 1$. The requirement grows linearly in the number of candidates and inversely in δ, and no choice of score function or residual distribution relaxes it.*

The condition follows from one-sided split conformal prediction (Vovk et al., 2005; Lei et al., 2018) under a union bound over the N candidates. Solving the same condition for candidate-family size gives $N_{\max} = \lfloor \delta(m+1) \rfloor$, the largest family admitting a finite quantile. In Appendix B.1, Table 3 instantiates both thresholds at $\delta = 0.1$ on the four measured landscapes.

All 55 measured split rows return $+\infty$ before any residual is examined. This matches the eleven constructions tested on 320 leakage-free instances containing 261 failures. The distribution-free construction covered every candidate and contained every witness, yet returned the entire pool in all 320 instances. The finite Gaussian-based alternatives fell short of the requested 90% simultaneous coverage, bounding all candidate scores in only 18.44% to 36.56% of instances. Median set-to-rank ratios on the same instances were 20,540 for the two-sided set and 14,179 for the one-sided one. The obstruction is therefore arithmetic rather than a property of the fitted residual model.

B.2 Restricting the pool and using residual smoothness

Restricting the family to a proxy-selected tail. Applying the bound only within the proxy's top-K tail shrinks the family the union bound must cover. Over $K \in \{8, 16, \ldots, 1024\}$ on 200 measured model-split instances and 3,200 rows, the two available versions fail for different reasons. Choosing the tail from the data breaks the interchangeability that split conformal prediction relies on, so reusing all validation residuals needs the extra assumption that the tail is uninformative about them; under it 51.3% of quantiles became finite but empirical simultaneous coverage among finite bounds was only 74.3%, so the assumption is not supported by these results. Restricting calibration to validation candidates above the realized proxy threshold preserves interchangeability but leaves a median relevant calibration count of one, and no tested quantile was finite at any K. Because a within-tail bound returning $+\infty$ retains all K candidates, the trusted cost of top-K probing is unchanged, and we therefore report top-K recall as the rank quantity $\Pr(R_\Delta \leq K)$ and keep it separate from the coverage of such bounds.

Bounding changes in proxy error. The residual $r(x) = f(x) - \hat{f}(x)$ is the trusted score minus the prediction. Suppose a known constant L bounds how much residuals can increase with candidate distance d, so that $r(x) \leq r(z) + L\, d(x, z)$. Then $u_L(x) = \min_z [r(z) + L\, d(x, z)]$, with the minimum over measured reference candidates z, bounds the residual at x and gives the score bound $f(x) \leq \hat{f}(x) + u_L(x)$ (Appendix A.1, Proposition 4). We tested this on eleven representative instances, using the number of differing sequence positions or the total absolute difference in component amounts, measured in mixture grid units, as distance.

The oracle constant $L^\star$ is the smallest valid L, computed using the complete trusted table. It gave 100% coverage and retained every witness, but median retained pool fractions of 0.995 to 1.000 show that even this choice excluded very few candidates. Estimating L from the largest residual

change per unit distance among calibration candidates bounded all pool scores in zero of eleven instances; doubling these estimates raised this to seven, while generally retaining essentially the full pool. Calibration data alone cannot guarantee such a bound for unseen candidates without a justified assumption on residual variation; within this construction, $L = \infty$ retains the entire pool.

B.3 Bounding witness prevalence with random probes

When probing finds no improvement, random sampling can bound the fraction of the pool that still beats the nomination. As in Section 3.5, we draw m candidates uniformly from $\mathcal{X}$ without replacement, independently of the proxy and the probe set S_B. The number of sampled witnesses follows a hypergeometric distribution; its observed value gives an exact one-sided upper confidence bound on the pool's witness fraction. This requires no calibration of prediction errors and does not identify which untested candidates are witnesses.

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 rows. Empirical coverage met the requested confidence level in every combination of landscape, budget, and confidence level. In Appendix B.3, Table 5 reports the mean bounds at 95% confidence against measured prevalence at $\Delta = 0$.

Table 4: **The two audit outcomes cover every campaign.** Shares are over 7,300 matched-budget campaigns. A positive certificate proves failure and supplies a replacement; a zero certificate retains the nomination while an independent random sample bounds pool witness prevalence.

Outcome	Share	Guarantee
$L_B > 0$	82.86%	Failure certified; measured improvement delivered; remaining regret $G_{\text{verifier}} - L_B$
$L_B = 0$	17.14%	Incumbent retained; random probes bound pool witness prevalence

Table 5: **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 $f(x_{\text{sel}})$. At $m = 16$, the bounds are two to fifteen times the measured prevalence.

Landscape	Prevalence	Bound, m=16	Bound, m=64	Coverage (%) m=16	Coverage (%) m=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

At $B = 16$, proxy-prefix probing certified 88.00%, 94.29%, 92.38%, and 95.24% of avGFP, GB1, TF-Bind-8, and TrpB instances, respectively, while random sampling supplied a finite upper bound for every uncertified case. Thus every audit produced either a measured improvement or a bound on witness prevalence. The two procedures are budgeted separately: B trusted evaluations for the proxy prefix and m for the random sample. We keep the random sample separate from prefix probing to preserve the independence required by the hypergeometric calculation.

C Datasets, splits, and preprocessing

Complete or near-complete measurements support evaluation of witness-rank distributions, exact selection regret, and certification recall within each declared pool. In Appendix C, Table 6 summarizes the pools and cleaning rules.

GB1. Each candidate specifies the amino acid at four variable sites in the B1 domain of protein G, and measured binding fitness is the trusted score (Wu et al., 2016). We use the measured supplement

Table 6: **Declared pools after cleaning, with all train, validation, and selection intersections empty.** Exactness of every reported quantity is relative to the measured pool.

Landscape	Declared candidates	Train/val/selection	Cleaning
TF-Bind-8	65,536	4,096/2,048/59,392	256 identical-label duplicate rows deduplicated
TrpB	159,129 measured	4,960/1,240/152,929	non-imputed measured non-stop variants only
GB1	149,361 measured	4,960/620/143,781	directly measured variants only, imputed table excluded
avGFP	51,714 measured	4,096/2,048/45,570	unique finite measured sequences

and exclude the separately imputed table. A monotone $\log(1+y)$ transform preserves ranking and leaves 149,361 unique measured sequences, or 93.35% of the theoretical 20^4 possibilities. Removing training and validation sequences within each split yields a fixed selection pool of 143,781 candidates.

TrpB. The TrpB active-site landscape provides a combinatorially complete four-site measurement (Johnston et al., 2024). We retain non-imputed measured non-stop variants only, giving 159,129 candidates and a selection pool of 152,929.

TF-Bind-8. The eight-mer transcription-factor binding table (Barrera et al., 2016) is used in the Design-Bench packaging (Trabucco et al., 2022). Removing 256 duplicate rows with identical labels leaves 65,536 distinct sequences; after splitting, the selection pool contains 59,392 candidates.

avGFP. The ProteinGym (Notin et al., 2023) table for the Sarkisyan avGFP landscape (Sarkisyan et al., 2016) has 51,714 unique measured 238-residue variants. Preprocessing found zero duplicate mutation strings, zero duplicate reconstructed sequences, zero conflicting labels, zero missing or nonfinite scores, and zero reference-residue mismatches.

Excluded: TF-Bind-10. We hold TF-Bind-10 out of the headline evidence. Its 84 canonical shards contain 8,321,066 rows for 1,048,576 sequences, and every sequence occurs five to eight times with conflicting labels; canonical Design-Bench iteration overwrites prior values in shard and row order. Reproducing that iteration matched the packaged mapping exactly, but every alternative ordering changed all candidate values and the identity of the optimum. Any exact statement about a pool defined this way is a statement about an arbitrary iteration order, so we retain TF-Bind-10 only as benchmark-relative sensitivity evidence.

D CONTROLLED MIXTURE FAMILY

This family lets us vary training-data quantity and interaction strength separately while knowing every candidate's trusted score. We specify the mixtures, score generator, sampling, and matched searches below.

Candidate mixtures. A candidate $x \in \mathbb{R}^n_{\geq 0}$ contains at most k of n components. Its fractions sum to one and are multiples of $1/D$, so D sets the grid resolution and the support size counts components with nonzero fractions. The reference score combines individual-component and pairwise effects,

$$q(x) = c + \sum_{i=1}^{n} a_i x_i + \sum_{1 \leq i < j \leq n} b_{ij} x_i x_j, \tag{10}$$

where c is a constant offset, a_i a component effect, and b_{ij} a pairwise interaction coefficient. The model has $p = n + \binom{n}{2}$ features. For a mixture containing s components, there are $\binom{n}{s}$ choices of components and $\binom{D-1}{s-1}$ ways to assign their grid units, giving

$$|\mathcal{X}(n,k,D)| = \sum_{s=1}^{\min(k,n,D)} \binom{n}{s}\binom{D-1}{s-1}.$$

Our setting $(n, k, D) = (12, 4, 8)$ gives 22,419 candidates and feature vector $\phi(x) = (x_1, \ldots, x_n, x_1x_2, \ldots, x_{n-1}x_n) \in \mathbb{R}^{78}$; the model excludes self-interaction terms x_i^2.

Fitting the score generator. Five enthalpy and five viscosity mixture tasks determine how often coefficients are nonzero, their signs, and the spread of their magnitudes. Additive coefficients have log absolute magnitudes with standard deviation 1.0593 and are positive with probability 0.613. Pair coefficients are nonzero with probability 0.2335, have log absolute magnitudes with standard deviation 1.1876, and are positive with probability 0.5232. These fitted parameters retain their original precision for reproducibility. Because both task families informed the generator, the mixture benchmark in Appendix N is an in-domain check rather than independent validation.

Varying interactions at a fixed score scale. Let $A(x)$ denote the normalized additive score from individual components and $I(x)$ the normalized score from pairwise interactions. We subtract each score's mean and remove the pairwise score's linear dependence on A before rescaling. Both components then have unit variance and zero covariance over the complete grid. We vary their contributions using

$$f_\alpha(x) = \sqrt{1-\alpha}\, A(x) + \sqrt{\alpha}\, I(x), \qquad \alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}, \tag{11}$$

where the square-root weights make α the fraction of score variance due to interactions: $\mathrm{Var}(f_\alpha) = (1-\alpha) + \alpha = 1$. The fitted coefficient statistics above apply before this normalization. For the omitted-interaction study in Appendix K (Equation (12)), $J(x)$ denotes the normalized three-component interaction score. We construct it from $\sum_{i<j<k} c_{ijk}x_ix_jx_k$ by subtracting its mean and linear dependence on A and I, then rescaling to unit variance. The final combined score is also scaled to unit variance, so the coefficient γ changes the strength of omitted interactions without changing the score scale.

Training and proxies. Training samples include mixtures with one to four components, with sampling weights that prevent four-component mixtures from dominating. Training sets are nested, so each larger set contains the smaller one, while training, validation, and test partitions remain disjoint. Training coverage is M/p, the number of training candidates M per pairwise-proxy feature.

Validation root-mean-squared error selects regression hyperparameters. Pairwise ridge uses all 78 additive and pairwise features. The composition model is a multilayer perceptron (MLP), a neural network whose inputs are the 12 component fractions; we also compare an ExtraTrees ensemble of regression trees. Evaluating each fitted proxy over the complete pool determines its exact nomination.

Comparing proxy and trusted searches. To separate search difficulty from proxy error, we run the same search algorithm and budget using either proxy or trusted scores. Both searches propose moving one grid unit of mixture from one component to another, subject to the mixture constraints above. Local search keeps improvements and periodically restarts. Simulated annealing also accepts ties and sometimes worse moves: a score decrease $d > 0$ is accepted with probability $\exp(-d/T)$, where the temperature $T > 0$ decreases linearly. The trusted-score run provides a retrospective comparison using the known pool scores. Appendix D (Table 7) gives the experimental settings.

Table 7: **Settings for the controlled mixture experiments.** The primary studies vary interaction strength, training size, and search budget; the misspecification study adds omitted three-component interactions. Training sizes are given as candidates per model feature, M/p.

Item	Value
Candidate set	$n = 12$ components, at most $k = 4$ present; fractions in multiples of $1/D$ with $D = 8$; 22,419 candidates
Primary landscapes	50 landscapes; $\alpha \in \{0, 0.10, 0.25, 0.50, 0.75, 1\}$
Models	pairwise ridge and composition MLP in the primary studies; pairwise ridge, MLP, and ExtraTrees in the misspecification comparison
Primary training	nested $M/p \in \{0.5, 1, 2\}$; probabilities $(0.15, 0.50, 0.25, 0.10)$ for mixtures with one, two, three, and four components
Misspecification training	nested $M/p \in \{0.5, 1, 2, 4, 8, 16\}$
Validation/test	19/39 disjoint candidates per task
Search	restarted local search and simulated annealing
Search budgets	16, 32, 64, 128, 256, and 512 proxy evaluations
Uncertainty	5,000 landscape bootstrap resamples

E Full audit results

We expand the certification results from Section 5.2 across measured and controlled landscapes, then compare probe policies and stopping rules.

E.1 Witness ranks and recall

Across 1,364 controlled failures, witness rank had median 3, interquartile range 2 to 9, 95th percentile 1,031, and maximum 16,344, matching the concentration at small ranks with a few much larger values in Figure 2(a). The global bound remained conservative, with a median bound-to-rank ratio of 4,483.8. For the measured landscapes, Appendix E.1 (Table 8) reports strict-improvement results ($\Delta = 0$) for the 200 original audits by landscape and model; Table 1 summarizes the 230 headline audits at $\Delta = 0.05$ SD.

The same pattern appears in the controlled-suite budget curves in Appendix E.1 (Figure 4), which show recall and the fraction of selection regret removed, L_B/G_{verifier}, under proxy-prefix, local-neighborhood, and random probing. Proxy-prefix probing certifies most failures by sixteen evaluations, with little further gain in recall above 64.

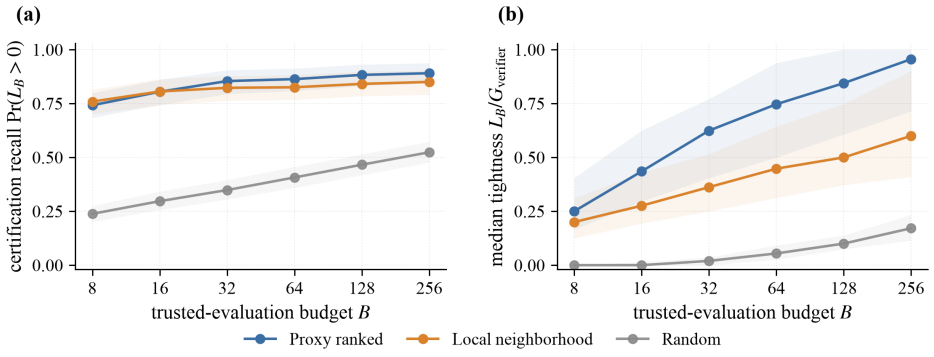

Figure 4: **In the controlled suite, proxy-prefix probing certifies most failures by sixteen evaluations.** *(a)* Fraction of instances with $G_{\text{verifier}} > 0$ certified by $L_B > 0$ against total trusted budget, the nomination included in every budget. *(b)* Median fraction of selection regret captured by the certificate, L_B/G_{verifier}, among instances with $G_{\text{verifier}} > 0$. The certificate is valid for every policy regardless of its recall or the fraction of regret it captures.

Exact random-order baseline. For Figure 2(b), budget B includes the nomination and $B - 1$ uniform probes without replacement from the remaining candidates. We compute the exact probability

Table 8: **Selection failure in all original combinations of landscape and model across 200 audits.** Ranks and recall use strict improvement ($\Delta = 0$, $R = R_0$). Brackets give 95% split-bootstrap intervals; the two avGFP rows are point estimates. G_{verifier} is in native trusted units and is not comparable across landscapes.

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

of finding a Δ-witness for each audit at $\Delta = 0.05\,\mathrm{SD}$, then average with equal weight per audit, including audits without witnesses. The reported recall ratios divide structured recall by this mean baseline within each landscape.

How ordering contributes: rank-shuffle ablation. We permute the proxy ordering at three levels over 16,000 rows to distinguish the effects of shortlist membership and ordering within the shortlist. A global shuffle drops recall@16 to 0.467, 0.108, 0.072, and 0.208 on avGFP, GB1, TF-Bind-8, and TrpB. Shuffling separately within each tenth of the proxy ranking recovers part of the recall, at 0.754, 0.356, 0.255, and 0.433. Shuffling only within the top 1% is the discriminating condition: recall falls to 0.633 on GB1, 0.492 on TF-Bind-8, and 0.715 on TrpB, but stays at 0.807 on avGFP. The fine ordering of the highest-ranked candidates therefore matters on three of four landscapes. On avGFP, reordering the shortlist has little effect, consistent with its higher witness prevalence. 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.

E.2 Proxy quality and margin sensitivity

Fixed deployable proxies. The expanded proxy study contains 525 rows spanning five to seven model families per landscape. Without selection-pool labels, we tested score-weighted ridge regression, a simple ensemble, and ridge regression on frozen ESM-2 embeddings (Lin et al., 2023) for GB1 and avGFP. The weighted ridge gives more weight to the top-scoring 10% of examples; the ensemble averages model predictions after standardizing them to zero mean and unit variance. Replaying the saved predictions reproduces the original and expanded top-tail Spearman ceilings to numerical precision; their rounded values are 0.383 and 0.402. Conditional median witness ranks remain between two and six. An ensemble with nonnegative weights fitted on the highest-scoring validation candidates reaches 0.743 on its GB1 tuning labels but only 0.319 on the disjoint selection pool. Over a fortyfold increase in training-set size within a fixed candidate pool, failure probability remains 1.0 in six of eight combinations of landscape and model and falls only to 0.867 in two TrpB model groups, with conditional median rank two to three.

Oracle quality ladder. To study quality beyond what the fixed models achieve, normalized complete-table trusted ranks are mixed 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 disjoint selection pool. Higher validation

correlation generally corresponds to higher selection-pool correlation, though the values differ. 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 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. Table 2 reports audit outcomes binned by achieved selection-pool quality, not by the requested validation target.

Margin sensitivity on the headline audits. In Appendix E.2, Table 9 recomputes the 230 headline audits at five margins. Witness existence and unconditional recall decrease with the required improvement, while conditional ranks remain small in many experimental conditions. The avGFP median of seven at $\Delta = 0.5$ is conditional on two audits with a qualifying witness.

Table 9: **Margin sensitivity of the 230 headline audits.** Recall@16 is unconditional; medians are conditional on a qualifying witness. P90 denotes the 90th percentile of finite witness ranks.

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.3 Probe policies and stopping rules

In Appendix E.3, Figure 5 compares five single-round policies on 60 matched measured failures: proxy-prefix probing follows descending proxy score; local-neighborhood probing favors nearby sequences; disagreement probing favors candidates on which the proxy ensemble disagrees; diversity probing spreads evaluations across feature space; and random probing samples uniformly without replacement. Proxy-prefix probing has the highest recall at every tested budget. Its recall is 0.900 against random probing's 0.197 at $B = 8$, and 1.000 against 0.675 at $B = 256$. Mean fraction of regret removed by certified delivery, L_B/G_{verifier}, rises from 0.307 to 0.680 over the same range. Local-neighborhood probing is competitive at small budgets and then plateaus, consistent with witnesses lying near the nomination in sequence space but not only there. This comparison is internal to the fixed-shortlist family; adaptive multi-round policies are compared at matched budget in Section 5.2 and Appendix F.1.

Adaptive stopping. Because R_Δ is unknown in advance, we also evaluated rules that stop probing when continuation appears unprofitable. Across 1,800 policy rows no rule produced a false certification, since adaptivity does not alter the lower-bound property of Equation (3), and none improved on the fixed shortlist: the best reached 0.944 recall with 5.08 mean trusted evaluations against 4.76 for fixed top-16 at the same recall. We therefore use fixed top-B probing with B read from the empirical rank distribution.

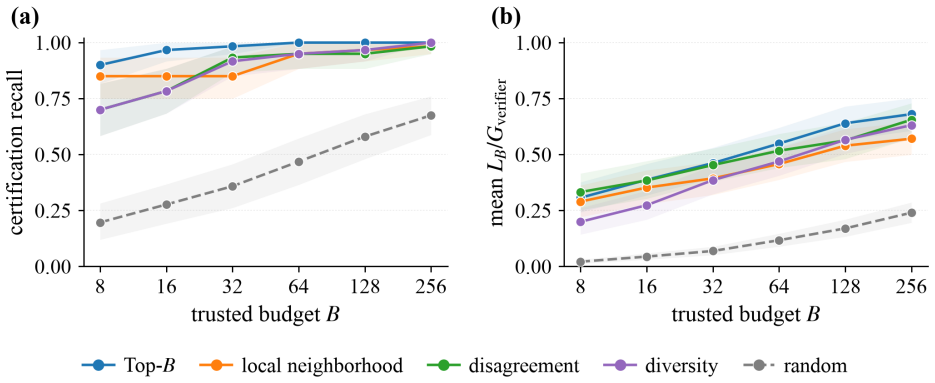

Figure 5: **Proxy-prefix probing has the highest recall among the five single-round policies at every tested budget.** *(a)* Certification recall against trusted budget over 60 matched measured failures. *(b)* Mean fraction of selection regret removed by certified delivery, L_B/G_{verifier}. Ribbons are 95% bootstrap intervals. Adaptive multi-round policies are compared separately in Appendix F.1, Table 10.

F CERTIFIED REPAIR

We expand the delivery result from Section 5.4 with an independent recomputation and comparisons with refitting and other uses of the same audit labels. Figure 3(a) shows the fraction of regret removed, L_B/G_{verifier}, across budgets; the remaining fraction is $1 - L_B/G_{\text{verifier}}$.

Regret removal and independent recomputation. An independent calculation reproduced the certified improvements and the scores of returned candidates across 248 failure instances and 34,224 metric rows. The paired comparisons against random probing are positive with 95% intervals excluding zero on all four landscapes, showing that the gain comes from the probe policy rather than from trusted-label expenditure.

Replacement compared with 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 changed mean selection regret by -1.087 and -0.441 on the controlled suite at $B = 16$ and $B = 64$, but by $+0.016$ and $+0.035$ on TF-Bind-8 and $+0.005$ and $+0.078$ on TrpB, so refitting increased selection regret in every measured combination of landscape and budget. A targeted follow-up added disagreement neighborhoods, validation-residual ranking, and a hybrid of certification and pairwise regression over 60 controlled, 15 TF-Bind-8, 15 TrpB, and 10 avGFP cases, producing 2,000 outcomes across five equal-label policies. The hybrid improved mean normalized regret on avGFP and TrpB but worsened it on TF-Bind-8, beat plain top-B significantly in only one of six measured experimental conditions, and was significantly worse than random on the controlled suite. Returning the best measured candidate with the nomination retained guarantees non-degradation. Section 6 compares the improvement obtained with adaptive acquisition policies.

Reusing audit labels. Adjusting proxy scores without changing their order leaves the shortlist unchanged. A tested rule for extending the shortlist after finding a witness also selected the same candidates as continuing in proxy order, because the first witness was already near the top. Both therefore reproduced the original policy. A reranker based on prediction errors did change the ordering and improved delivery in some GB1 and TrpB experimental conditions, but not consistently across landscapes. Label reuse can therefore alter acquisition only by changing the ordering, and does not guarantee that the next nomination improves on the current one.

F.1 MATCHED-BUDGET CAMPAIGNS AND THE DELIVERY WRAPPER

This subsection reports the 7,300 matched-budget multi-round campaigns behind Section 6, comprising 36,500 policy-round rows. The five policies are proxy-prefix probing, Thompson sampling,

batch Bayesian optimization (BO), witness-local probing, and random probing. 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.

Improvement by policy. Mean final trusted-value improvement is highest for an adaptive policy in every combination of landscape and acquisition schedule. Thompson sampling leads on GB1 and TrpB, reaching 0.617 against prefix probing's 0.428 on GB1 after 16 new probes and significantly exceeding batch BO on both landscapes; batch BO leads on TF-Bind-8 and is significantly better than prefix probing there and on TrpB, while being significantly worse on avGFP; witness-local probing has the highest means on avGFP, where its paired interval against prefix excludes zero for one schedule, at $+0.018\,[0.001,\ 0.035]$. Values are in native trusted units and are not comparable across landscapes; means for each combination of landscape and acquisition schedule are in the released campaign table.

Certification by policy. Recomputing L_B over each campaign's evaluated set, which holds the nomination and every distinct candidate probed in any round, gives the rates in Appendix F.1, Table 10: prefix probing certifies 0.905, 0.945, 0.905, and 0.991 of campaigns on avGFP, GB1, TF-Bind-8, and TrpB against 0.630, 0.219, 0.231, and 0.279 for random probing, with the three adaptive policies between 0.860 and 0.993. Within each landscape, paired split-cluster intervals exclude zero, with advantages over random probing from 0.240 to 0.750. Pool SD enters only the post-hoc threshold $0.05\,\mathrm{SD}$ and is never visible to a policy. Rates among the structured policies are close: prefix probing and Thompson sampling show no statistically detectable difference on avGFP, GB1, and TrpB, Thompson certifies more often on TF-Bind-8, and a strict zero-margin reanalysis gives the same ordering.

Table 10: **Structured acquisition certifies more often than random probing, but no structured policy dominates.** Rates are over 7,300 matched-budget campaigns at $\Delta = 0.05\,\mathrm{SD}$; bold marks the largest 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

The gain from retaining the nomination. Reusing the same campaigns without rerunning acquisition, we compare standard best-observed delivery, the trusted best among probed candidates excluding the nomination, against incumbent-safe delivery, the trusted best among probes and nomination together. The only difference is whether the nomination remains an option for the final output. Mean incumbent gain in native trusted units is $0.031\,[0.015,\ 0.048]$ for batch BO 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 experimental conditions having intervals that include zero. Standard delivery 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. There were $1{,}251$ zero-certificate campaigns (17.137%). Retaining the nomination improved the final trusted score in $1{,}243$ of these campaigns; in the other eight, the best TF-Bind-8 probe tied the nomination exactly. This is the distinction described in Section 3.5. No wrapped delivery fell below either the standard best-observed value or the nomination.

G Published offline optimizers

We document the two reproductions from Section 5.5, including exact schedules, native reference runs, and checks of whether returned candidates belong to the declared pool.

RaM-ListNet. We used the official RaM-ListNet implementation (Tan et al., 2025), with the ListNet objective (Cao et al., 2007) and Adam search. A TF-Bind-8 reference run completed 100 epochs over the full 10,000-list corpus with the official search and achieved normalized score 0.991, reproducing the published behavior. The finite-landscape audit contains fifteen runs, five splits each of TF-Bind-8, TrpB, and GB1. Every run used 10,000 fixed lists of length 1,000, an $8{,}000/2{,}000$ fixed-list partition, batch size 128, width $[2048, 2048]$, 100 epochs, exactly 6,200 optimizer and 1,500 validation steps, and 128-start 100-step search. The fixed corpus is presented in full, at 800,000 training-list presentations per run, rather than sampled online.

Median selection regret and witness rank were 0.027 and 3 on TF-Bind-8, 0.345 and 2 on TrpB, and 0.407 and 4 on GB1; top-8 recall was 0.8, 1.0, and 0.8 and top-32 recall 1.0 throughout, with median regret removal at $B = 256$ of 1.0, 0.803, and 1.0. Independent checks confirmed the schedule, data separation, nomination, certificate, and replacement calculations over all 270 reported rows.

Conservative Objective Models (COMs). We ran the official COMs implementation (Trabucco et al., 2021) on native TF-Bind-8 for the full 50 epochs with the official particle search, producing normalized score 0.936. Large adversarial-overestimation diagnostics prompted a check of the original and updated implementations using identical batches and initial weights. Their post-update weights agreed to 1.30×10^{-7} on average, with no difference in the objective, optimizer, or step size. The published TF-Bind-8 hyperparameters were used throughout and transferred unchanged to TrpB and GB1: width 2,048, 50 epochs, batch 128, 50 train and evaluation particle steps, particle learning rate 2, overestimation limit 2, and 128 starts.

Over five splits each, mean selection regret, median witness rank, and top-8 recall were 0.013, 7, and 0.6 on TF-Bind-8; 0.426, 3, and 1.0 on TrpB; and 0.678, 2, and 0.6 on GB1, with mean regret removal at $B = 256$ of 1.000, 0.771, and 0.795. In Appendix G, Figure 6 places both pipelines side by side on a common logarithmic regret axis.

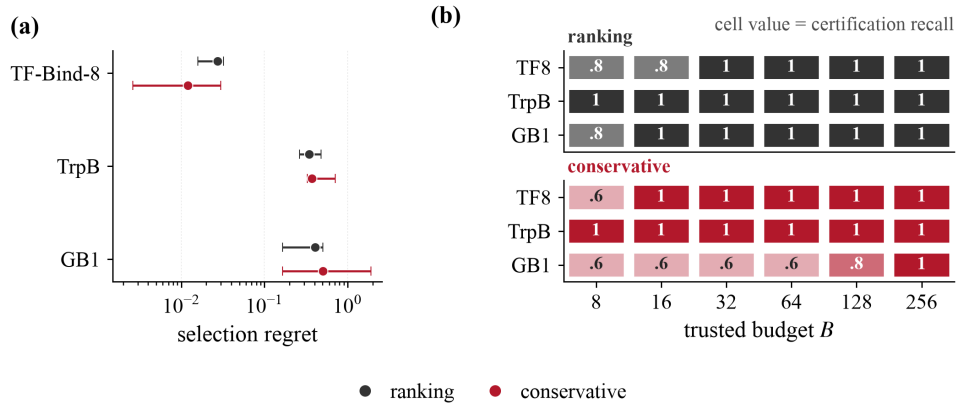

Figure 6: **Both published optimizers show the same pattern at their full published schedules.** *(a)* Selection regret of the exact learned-score optimum, five splits each, median with range on a logarithmic axis because the values span more than an order of magnitude. *(b)* Certification recall against trusted budget as a heatmap, one row per landscape within each pipeline, recall printed in every heatmap entry; RaM-ListNet reaches recall one by $B = 32$ on all three landscapes.

Unconstrained proposal audit. 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 highest-scoring decoded proposal inside the pool, and the method's actual output may lie outside the pool. RaM-ListNet's actual output was eligible in five of five TF-Bind-8 runs and zero of five on both TrpB and GB1, with every ineligible one inside the training split; COMs' output was eligible in nine of fifteen, namely zero of five on TF-Bind-8 and four and five of five on TrpB and GB1. Ineligible deliveries have undefined finite-pool metrics. We did not project or replace these outputs, or estimate their missing finite-pool scores.

Constrained-pool rerun. We then reran all 30 optimizer instances, restricting RaM-ListNet's proposed candidates and COMs' candidate updates to the declared pool. All final deliveries were eligible, no output was projected or substituted after optimization, and all configurations matched the reported schedules. For RaM-ListNet, median witness rank was 8 on GB1, 3 on TF-Bind-8, and 4 on TrpB; top-16 certification rates were 0.6, 1.0, and 1.0, and the exact proxy optimum within the declared pool was delivered in 1.0, 0.8, and 1.0 of runs. For COMs, the corresponding median ranks were 27, 3, and 20, top-16 rates were 0.0, 0.2, and 0.2, and the exact proxy optimum within the declared pool was never delivered. Witnesses remain detectable when proposals are restricted to the declared pool. COMs' larger search gap accompanies witnesses appearing later in the probe ordering.

H POOL SIZE AND WITNESS RANK

We test whether witnesses remain easy to find as the candidate pool grows, comparing both their frequency and their positions in the proxy ranking. The measured study contains 840 audits on progressively larger subsets of the TF-Bind-8, TrpB, and GB1 selection pools, each containing the preceding subset. At full pool size, top-8 recall was 0.817, 0.883, and 0.900 on TF-Bind-8, TrpB, and GB1 and top-64 recall was 1.0 on all three, with median witness rank between two and six depending on model and landscape. Witness rank was only weakly related to the fraction of candidates that were witnesses. To test whether the proxy ordering finds witnesses sooner than chance, we use the exact recall of uniform random probing in Section 5.2 and the equal-budget comparison in Appendix F.1 (Table 10). A prespecified analysis of high-rank cases classified 20 as involving rare witnesses alone, two as involving witnesses ranked too low by the proxy alone, none as both, and 30 as neither, so neither diagnostic accounts for all high ranks.

A controlled study then evaluated 848 instances at pool sizes of one thousand, ten thousand, one hundred thousand, one million, and five million candidates, alongside the 840 measured rows. We constructed witness placements at a fixed rank, at $\sqrt{N}$, at $0.01N$, and at $N/2$, where N is the pool size. In Appendix H, Table 11 gives the five-million-candidate behavior and Figure 7 shows the full size sweep.

Table 11: **Only fixed-rank witness placement remains auditable within 64 probes at five million candidates.** Measured pools behave like the first row; the other three place witnesses at $\sqrt{N}$, $0.01N$, and $N/2$ and are not certifiable by a small fixed budget.

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 pool	2,500,000	0	0	0

The ratio $|\mathcal{C}_-|/R$ compares the number of candidates retained by the error bound with the actual witness rank (Appendix A.1, Proposition 3). Its medians are 6,751, 14.29, 2.14, and 1.02 in table order; values near one indicate close agreement between the bound and observed cost, found here for witnesses furthest down the ranking.

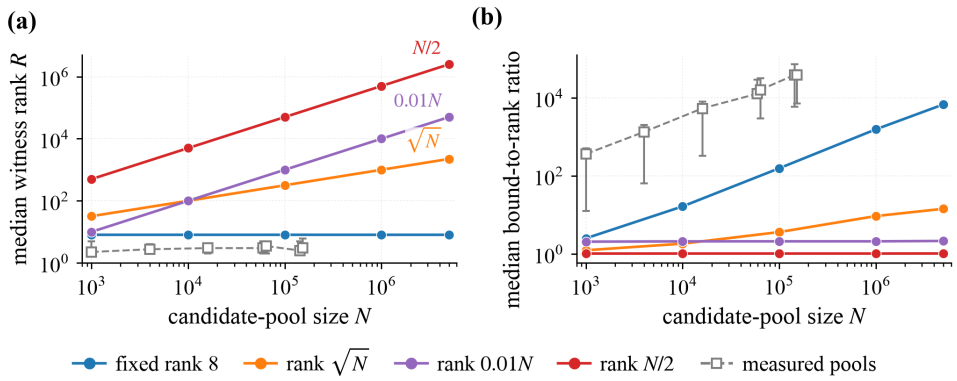

Figure 7: **Measured pools track fixed-rank witness placement.** *(a)* Median R against pool size for four constructed witness mechanisms and for measured nested pools, both axes logarithmic, with $\sqrt{N}$, $0.01N$, and $N/2$ marked. *(b)* Median ratio $|\mathcal{C}_-|/R$ of the number of candidates retained by the error bound to actual witness rank (Appendix A.1, Proposition 3). At the largest pool size, the bound is closest to the observed cost for witnesses placed furthest down the ranking.

I NOISY CERTIFICATES

To assess whether measurement noise leads to false claims of improvement, we test the bound in Appendix A.1 (Equation (7)) and ten variants across 396,000 trials: 180 controlled top-8 audits, 11 noise-and-certificate settings, and 200 trials per setting. Methods with a fixed sample size use 128 measurements per candidate. The sequential method, which distributes the allowed error probability across repeated checks, uses 216 on average.

Appendix I (Table 12) focuses on six settings testing variance estimation, large noise outliers, and systematic bias. Settings satisfying the stated assumptions had low observed false-certification rates. Uncorrected systematic bias raises this rate to 0.214, whereas accounting for a known bias bound restores validity with recall 0.520, close to 0.521 for known Gaussian noise. For Student-t_3 noise, a distribution with three degrees of freedom and more frequent large deviations than Gaussian noise, Chebyshev's inequality (Wasserman, 2004, Chapter 4) gives a valid but conservative bound; assuming Gaussian noise instead gives no guarantee despite a low observed error rate. The other five settings, covering bounded and random-sign errors, sequential checks, unequal variances, and correlated errors, also had low observed false-certification rates when their assumptions held.

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, at most three measurements per candidate suffice for 60.0% of avGFP instances, 88.6% of GB1, 47.1% of TF-Bind-8, and 94.3% of TrpB. This calculation assigns an infinite required repeat count to nonpositive observed certificates. TF-Bind-8's lower share reflects its smaller improvements relative to the pool's score spread.

Table 12: **Variance estimation, heavy tails, and bias affect noisy certification.** Selected settings from the eleven-setting study. 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

J RESOURCE ALLOCATION AND AUDIT-BUDGET PREDICTION

We tested whether witness rank could also guide resource allocation or predict audit budgets across domains. First, a matched-cost study over 37,800 rows asked whether a budget-32 certificate should determine whether the next block extends search or collects training data. Certificate guidance has the lowest mean area under the regret-versus-cost curve on the controlled suite and TF-Bind-8, while random allocation is lower on TrpB and GB1; the comparisons do not support a transferable allocation rule. Second, predicting log witness rank within known landscape families gives Spearman correlation 0.262, while predicting whether a budget B suffices ($R \leq B$) gives an area under the receiver operating characteristic (ROC) curve, or AUROC, of 0.855 to 0.891. The ROC curve plots true-positive against false-positive rates as the prediction threshold varies. When an entire landscape is withheld from training, Spearman correlation for predicted log witness rank ranges from -0.463 to 0.269 and budget-prediction AUROC falls to 0.556 to 0.686. These results support measuring realized audit costs within the evaluated pools but do not establish a transferable budget predictor or allocation policy.

K TRAINING DATA AND MODEL STRUCTURE

To distinguish error due to limited data from error due to model misspecification, we use the controlled mixture family of Appendix D. We vary training size through M/p, the number of training candidates M per pairwise-model feature, with $p = 78$. Separately, we add interactions between three components that the pairwise proxy cannot represent, using

$$f_{\alpha,\gamma}(x) = A(x) + \alpha I(x) + \gamma J(x), \quad \alpha = 1, \quad \gamma \in \{0, 0.25, 0.5, 1, 2\}, \tag{12}$$

where A, I, and J are the additive, pairwise, and three-component scores defined in Appendix D, each with zero mean, unit variance, and zero covariance with the others. The coefficient γ controls the strength of the omitted three-component term J. Across 30 matched landscapes, we vary $M/p \in \{0.5, 1, 2, 4, 8, 16\}$ using nested training sets, so each larger set includes the smaller one.

Selection regret vanishes with additional data when the proxy class is correctly specified. At $\gamma = 0$, mean standardized G_{verifier} for pairwise ridge reaches zero at $M/p = 2$ and remains zero as training size increases, so failures in that regime come from limited data rather than missing model terms. Omitted three-component interactions produce persistent error even with more data: at $M/p = 16$, mean standardized G_{verifier} remains 0.435 for $\gamma = 1$ and 1.281 for $\gamma = 2$, with intervals $[0.209, 0.711]$ and $[0.923, 1.663]$. On a prespecified subset of ten matched landscapes, pairwise ridge, the neural network using component fractions, and ExtraTrees all retain nonzero regret when the generator includes three-component interactions, showing that this persistence is not limited to the pairwise model.

Appendix K (Figure 8) shows why more training data need not correct a nomination when relevant interactions are missing from the model.

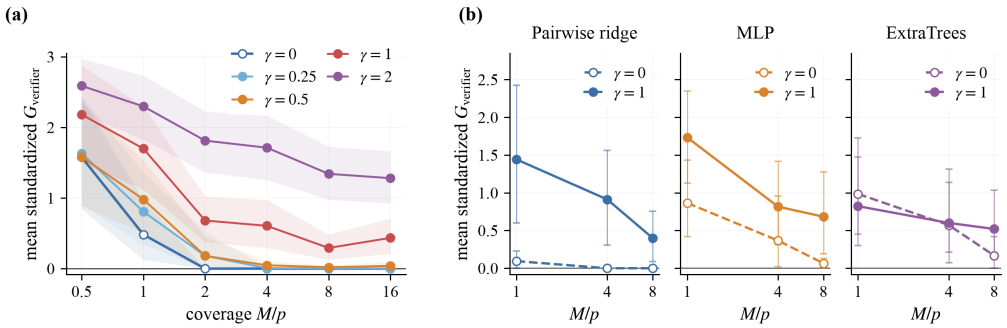

Figure 8: **More training data remove estimation error but not missing model terms.** *(a)* Mean standardized exact G_{verifier} for pairwise ridge across training examples per feature, M/p, and omitted three-component interaction strength γ; ribbons are 95% landscape-bootstrap intervals over 30 matched landscapes. *(b)* Pairwise ridge, an MLP, and ExtraTrees on 10 prespecified matched landscapes, showing that regret persists with three-component interactions across all three models.

A five-round sequential loop with a changing candidate pool shows the same separation dynamically: correctly specified mean G_{verifier} falls from 1.585 to 0.075 as training data increase, while the misspecified value remains 1.078. Evaluated candidates leave the selection pool between rounds, so the comparison is no longer over a fixed set, but additional data reduce selection failure far more strongly under correct specification. Optimizer dependence behaves as the audit identity predicts: at $\alpha = 1$ and $M/p = 0.5$, four search algorithms reduce normalized G_{search} to 0.033 to 0.042 by budget 512 while exact G_{verifier} stays at 0.173.

L Mechanism studies in the controlled family

These studies ask why proxy-selected candidates fail: whether the cause is interaction strength, missing training examples, or more extensive optimization of the proxy. Building on Appendix K, we reuse the same 50 controlled landscapes while independently varying α, the fraction of score variance due to pairwise interactions, and M/p, the number of training candidates per pairwise-model feature (training coverage). Thus we can test how additional data change the effect of interactions on selection failure.

Interaction strength and training size. To test whether interactions mainly hinder search or mislead the proxy, we compare each landscape with its additive version ($\alpha = 0$), holding model class and search budget 512 fixed (Appendix L, Figure 9). Interactions increase regret only modestly when search uses trusted scores, but sharply when selection uses the proxy. Increasing training size from $M/p = 0.5$ to $M/p = 2$ reduces this added selection regret by more than an order of magnitude at each tested nonzero α. The effect on trusted search is unchanged because that search does not use training data. Here, ΔG_{hard} and $\Delta G_{\text{verifier}}$ denote changes from the matched additive landscape.

Missing training examples and longer proxy searches. Appendix L (Figure 10) separates two ways that proxy-guided selection can fail. In panel (a), we remove training mixtures containing both components of the trusted optimum's active pair. We replace them with unused mixtures, preserving training size and the frequencies of other components and pairs. Selection regret increases relative to removing the same number of examples at random, at every training size tested. The difference disappears when examples containing the pair are restored, and prespecified control removals do not reproduce it. This links the selection error to missing examples of that pair, rather than to fewer training data overall.

Panel (b) tests whether a larger proxy-search budget can worsen the selected candidate. It measures how often increasing the budget improves the proxy score but reduces the trusted score, and how often the final search output is better under trusted evaluation than the exact proxy optimum ($H > 0$). These two comparisons capture the overoptimization described in Proposition 1. Proxy gains accompanied

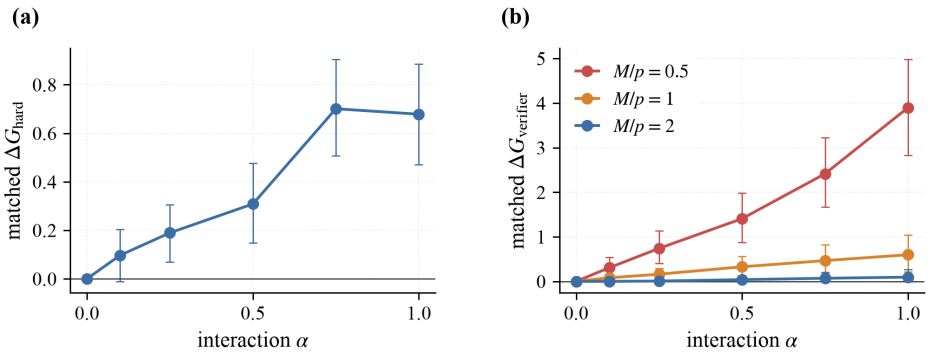

Figure 9: **Additional training data reduce interaction-induced selection regret; trusted-search regret is unchanged.** Interaction strength α is the fraction of score variance due to pairwise interactions. Both panels show changes from the same landscape without interactions ($\alpha = 0$): *(a)* regret when search uses trusted scores and *(b)* selection regret when the proxy is pairwise ridge, at three training sizes M/p (examples per model feature). Error bars are 95% percentile intervals over 50 landscapes, reused across conditions.

by trusted losses become more common with stronger interactions, with the largest effect for the neural network that takes component fractions as inputs (composition MLP).

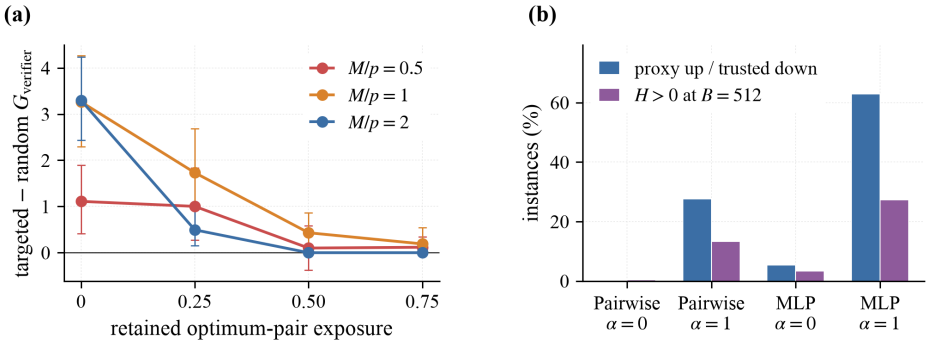

Figure 10: **Representation of decision-relevant component pairs in the training data drives selection failure.** *(a)* Removing the training examples covering the trusted optimum's active pair raises exact selection regret relative to random removal of the same size; paired 95% landscape intervals, $n = 50$. *(b)* Fraction of instances where increasing the search budget improves the proxy score but lowers the trusted score, and fraction where the search output at budget 512 has a higher trusted score than the exact proxy optimum ($H > 0$).

Local search and alternative interaction coefficients. Stronger interactions make nearby mixtures less similar in score and create more places where local search can stop. Raising α from 0 to 1 lowers their mean score correlation from 0.945 to 0.829 and raises the mean number of local maxima from 1.00 to 3.46. At a local maximum, moving one grid unit between two components cannot improve the score. The fraction of starting candidates reaching the most common local maximum falls from 1.000 to 0.546, so search outcomes depend more on the starting point.

To check whether selection failure depends on the generator's coefficient choices, we test six ways of assigning pairwise interaction coefficients. Interactions increase selection regret in all six: $\Delta G_{\text{verifier}}$ ranges from $0.128\,(0.091, 0.175)$ with all negative coefficients to $2.491\,(1.674, 3.386)$ with all positive coefficients. Their effect on trusted search is less consistent: negative coefficients make it easier, while Gaussian coefficients for all component pairs give an interval for ΔG_{hard} that includes zero. Changes to candidate-set size, model capacity, measurement noise, training sampling, and ridge regularization alter the amount of selection regret but do not remove it in these tests.

M PREDICTIVE DIAGNOSTICS FOR SELECTION FAILURE

We test whether prediction-error and training-data summaries can identify nominations worth auditing before collecting trusted probes. Over 900 controlled instances, with failure defined as $G_{\text{verifier}} \geq 0.1$ and complete landscapes held out during cross validation, the combined classifier reaches AUROC 0.910 and area under the precision-recall curve (AUPRC) 0.700. Held-out root-mean-squared error (RMSE) performs almost as well, while candidate density, representation of component pairs in training data, disagreement between models, and local uncertainty are substantially weaker. These measures predict failure risk; a trusted improvement is still needed to certify failure for an individual nomination. In Appendix M, Figure 11 reports discrimination and probability calibration, the agreement between predicted failure probabilities and observed failure frequencies, and Table 13 summarizes the corresponding continuous-regret analysis.

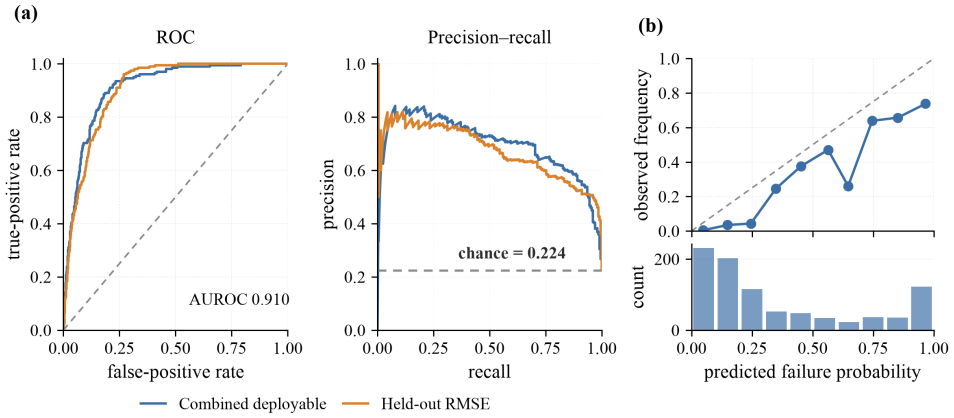

Figure 11: **Prediction errors help identify nominations at risk.** *(a)* ROC and precision-recall curves from predictions made with each landscape excluded from model fitting, for the combined classifier and held-out RMSE, with the combined AUROC of 0.910 annotated and the chance level of 0.224 marked. *(b)* Observed failure frequency by predicted probability, instances per bin below, diagonal marking perfect agreement between predicted probability and observed frequency.

For continuous regret prediction on these controlled landscapes, each model holds out one landscape and includes interaction strength, training examples per feature (M/p), and proxy family as context.

Table 13: **Prediction of selection regret from context and error metrics.** Pooled R^2 from predictions made with each landscape excluded from model fitting. Top-tail metrics use complete-table trusted labels. Intervals resample landscapes while holding these predictions fixed.

Predictors beyond context	Pooled out-of-fold R^2 (95% interval)
None	0.283 (0.223, 0.324)
Global root-mean-square error	0.441 (0.351, 0.498)
Training representation of component pairs	0.330 (0.258, 0.374)
Top-tail root-mean-square error	0.797 (0.745, 0.834)
All metric blocks	0.814 (0.768, 0.847)

Global RMSE improves on context alone, while the oracle top tail and combined metrics perform best; the oracle tail, however, uses trusted labels from the complete table and loses most of its advantage when restricted to offline labels.

N HELD-OUT MIXTURE-BENCHMARK AUDIT

We audit 875 molecular-dynamics enthalpy and 1,813 NIST viscosity formulations from CheMixHub (Rajaonson et al., 2025) at 298.15 K and 101 kPa, restricting local search and annealing to the test split. This is an in-domain check because the same benchmark's coefficient statistics were used to

fit the controlled generator (Appendix D). In Appendix N, Figure 12 shows that range-normalized G_{verifier} and G_{total} remain nonzero in every held-out experimental condition, and top-B certification recall on this benchmark reaches 0.829 at $B = 8$ and 1.000 by $B = 128$. An unrestricted whole-table comparison places 65.8% of approximate selections in training and a further 16.7% in validation; because it lets candidates used for fitting or model selection remain eligible, we include it only as a sensitivity analysis of the candidate-pool definition.

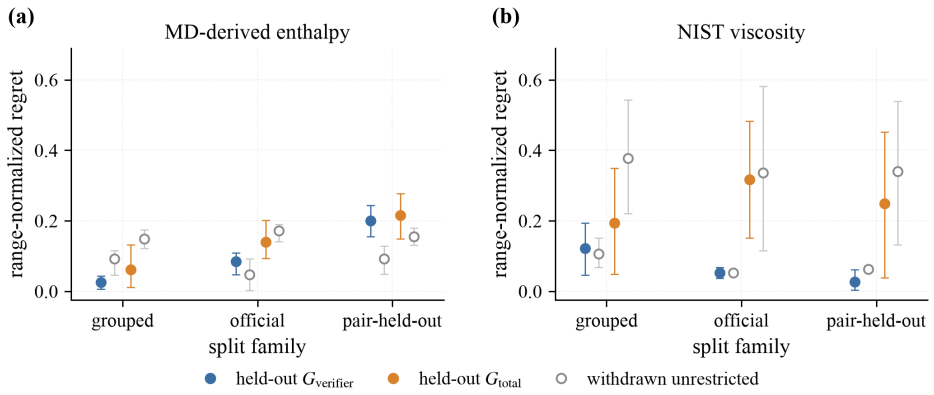

Figure 12: **Selection regret is nonzero in every held-out experimental condition of the mixture benchmark.** Range-normalized G_{verifier} and G_{total} for three split families in *(a)* molecular dynamics enthalpy and *(b)* NIST viscosity, both panels on the same scale. Filled marks are the held-out test-only protocol; open gray marks are the unrestricted whole-table sensitivity analysis.

O PROSPECTIVE WET-LAB ELECTROLYTE AUDITS

O.1 CANDIDATE POOLS, PREDICTIONS, AND AUDIT ACCOUNTING

We evaluated two finite pools of 1 M lithium bis(trifluoromethanesulfonyl)imide (LiTFSI) electrolytes: five carbonate-based formulations and five ether-based formulations. The carbonate-based pool contains propylene carbonate (PC), diethyl carbonate (DEC), and their 3:1, 1:1, and 1:3 mixtures by volume. The ether-based pool contains 1,2-dimethoxybenzene (12DMBn), tetrahydrofuran (THF), 2-methyltetrahydrofuran (2-MeTHF), 1,2-diethoxyethane (DEE), and diethylene glycol diethyl ether (DEGDEE). All reported optima and ranks are defined within these pools.

We used pretrained GeoMix (Li et al., 2025) and SMI-TED-IC (Zohair et al., 2025) weights. The published SMI-TED-IC model was fine-tuned on literature data including CALiSol-23 (de Blasio et al., 2024). The candidate pools and predictions were fixed before obtaining the in-house conductivity measurements. Within each family, each model nominates its highest-predicted candidate; descending predicted conductivity defines its probe ordering.

Both models share the same 10 formulation measurements, giving four audits, one for each combination of model and candidate pool. All candidates were measured. The observed R_0 is the first-witness position in each prediction ordering, counting the nomination as rank one.

O.2 MATERIALS AND CONDUCTIVITY MEASUREMENTS

Materials. LiTFSI was purchased from Sigma-Aldrich. Lithium bis(fluorosulfonyl)imide (LiFSI, 99.9%), used for conductivity calibration, was purchased from Arkema. All solvents were purchased from Sigma-Aldrich and used as received: 1,2-dimethoxyethane (DME, 99.5%), PC and DEC (both 99.5%, anhydrous), THF, 2-MeTHF, 12DMBn, DEE, and DEGDEE.

Conductivity measurements. Electrolyte formulations were prepared inside an argon-filled glovebox. Ionic conductivity was measured using a two-electrode Swagelok cell, a chamber holding the electrolyte. The polished stainless steel electrodes had an area of 1 cm^2 and were separated by 1 mm, without a separator. The cell constant, the factor converting measured conductance to conductivity,

was determined using 1 M LiFSI in DME as the 15 mS/cm conductivity reference. Electrolyte resistance was measured using a Biologic VSP potentiostat, and conductivity was calculated as $\sigma = k/R$, where k is the calibrated cell constant and R is the electrolyte resistance at 500 kHz. All measurements were conducted at room temperature (23 °C), with one measurement per formulation. Ranks and gains were computed from these measured values.

O.3 Complete observations and witness ranks

In Appendix O, Table 14 summarizes the four audits across models and candidate pools, and Table 15 reports every measured conductivity, proxy prediction, and rank in the proxy ordering.

Table 14: **Wet-lab witness ranks and measured improvements across four audits of model and candidate-pool combinations.** IC denotes ionic conductivity in mS/cm; N is the pool size. Both IC columns contain measured values. Witness rank R_0 locates the first improvement over the nomination in proxy order; gain is the best measured IC minus the nomination's measured IC.

Pool (N)	Model	Nomination measured IC	Best measured IC	Witness rank R_0	Gain
Carbonate-based (5)	GeoMix	5.40	5.91	2	0.51
Carbonate-based (5)	SMI-TED-IC	5.91	5.91	∞	0.00
Ether-based (5)	GeoMix	4.72	10.60	5	5.88
Ether-based (5)	SMI-TED-IC	1.93	10.60	2	8.67

Table 15: **Complete electrolyte measurements, predictions, and proxy ranks.** Every formulation contains 1 M LiTFSI; PC:DEC ratios are by volume. IC denotes ionic conductivity. Proxy rank is each formulation's position in descending predicted IC within its pool, with rank one denoting the nomination. Audit-level witness ranks appear in Appendix O (Table 14).

		GeoMix		SMI-TED-IC	
Solvent formulation	Measured IC (mS/cm)	Predicted IC (mS/cm)	Proxy rank	Predicted IC (mS/cm)	Proxy rank
Carbonate-based formulations ($N = 5$)					
PC:DEC 1:1	5.91	6.27	2	6.26	1
PC:DEC 3:1	5.62	5.67	3	6.04	2
PC	5.37	4.22	4	5.90	3
PC:DEC 1:3	5.40	9.03	1	5.86	4
DEC	4.04	4.02	5	5.44	5
Ether-based formulations ($N = 5$)					
12DMBn	1.93	9.96	3	12.87	1
THF	10.60	8.12	5	10.69	2
2-MeTHF	3.04	8.93	4	8.98	3
DEE	4.37	10.79	2	7.98	4
DEGDEE	4.72	10.81	1	5.63	5

For GeoMix on carbonates, the nomination has measured conductivity 5.40 mS/cm, and the second-ranked candidate reaches 5.91 mS/cm. Hence $R_0 = 2$ and $L_2 = G_{\text{verifier}} = 0.51$ mS/cm for the observed finite pool. SMI-TED-IC selects the 5.91 mS/cm candidate directly; no measured carbonate exceeds it, giving $R_0 = \infty$ and $L_B = 0$ for every prefix of the fully measured five-candidate pool.

GeoMix's first four ether candidates yield no improvement; THF is the first witness at $R_0 = 5$, giving $L_5 = G_{\text{verifier}} = 5.88$ mS/cm. SMI-TED-IC ranks THF second, giving $R_0 = 2$ and $L_2 = G_{\text{verifier}} = 8.67$ mS/cm. GeoMix therefore requires the complete ether pool, while SMI-TED-IC reaches the same best measured ether with one additional evaluation after the nomination. In all three overturned cases, the first witness recovers all observed within-pool selection regret. Gains relative to the nominations are 9.4%, 124.6%, and 449.2% for GeoMix on carbonates, GeoMix on ethers, and SMI-TED-IC on ethers, respectively.

A post-hoc improvement margin of 0.50 mS/cm preserves all three observed failures and their witness ranks.