# THE DIVERSITY MIRAGE: QUALITY-GATED EVALUATION FOR BIOLOGICAL SEQUENCE-DESIGN ACTIVE LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Active learning for sequence design aims to find batches of strong, diverse candidates with a limited measurement budget. Common evaluations assess overall quality but measure diversity among each method's own highest-scoring candidates. These candidate sets can differ in quality. A method can therefore look diverse even when its candidates score lower than those found by other methods. We call this the diversity mirage. We introduce quality-gated evaluation to compare methods under the same measurement budget. It uses a shared quality standard to assess both the number and diversity of high-quality candidates. We run controlled experiments on RNA design tasks and experimentally measured protein fitness data from ProteinGym. On RNA tasks, the shared standard reveals a reversal in the diversity ranking. A simple local search finds more distinct groups of high-quality sequences than a generative method designed for diversity. On ProteinGym, we use ESM features from mutated positions to select candidates. We find 42.9% more high-quality candidates on average than proxy-greedy selection, while achieving similar coverage of sequence space. With a shared quality standard, we can directly compare the diversity of high-quality candidates across methods. What matters is how many strong, distinct candidates the same budget delivers.Our code is available at `https://anonymous.4open.science/r/iclr_code-255E`.

## 1 INTRODUCTION

Active learning for biological sequence design aims to find useful candidates with a limited measurement budget. The number of possible sequences is far larger than the number we can measure. Each round uses the scores already observed to choose which sequences to measure next. A successful campaign should return a set of strong candidates that covers distinct regions of sequence space (Jain et al., 2022; Kim et al., 2024). This gives researchers several promising options to pursue.

Common evaluations assess quality and diversity separately. Optimization studies often report the best score or the average score of the top candidates (Sinai et al., 2020; Angermueller et al., 2019; Ren et al., 2022). Studies of diversity often measure how much sequences differ within each method's own highest-scoring set (Jain et al., 2022; Kim et al., 2024). These sets can have different quality cutoffs. A method can therefore appear more diverse even when its candidates score lower than those found by other methods. We call this the *diversity mirage*. It leaves the main question unanswered: *how many strong, distinct candidates does the same budget deliver?*

We introduce **quality-gated evaluation**, a shared standard for comparing what the same measurement budget delivers. It measures both the number of high-quality candidates and the distinct sequence groups they cover. Each task has one quality threshold $\tau$, set from a fixed reference pool of recorded scores. Every method is evaluated with this threshold and the same Hamming-distance rule. Candidates contribute to diversity only after meeting this shared quality standard. The evaluation uses scores already collected during a campaign, so it requires no additional measurements. Figure 1 illustrates the approach.

On three RNA tasks, we compare TR-BEAM, a simple local search, with $\delta$-CS, a generative method designed for diversity. Both methods use the same scoring model and measurement budget, with paired random seeds. The shared quality standard reverses their diversity ranking (Figure 4). TR-



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

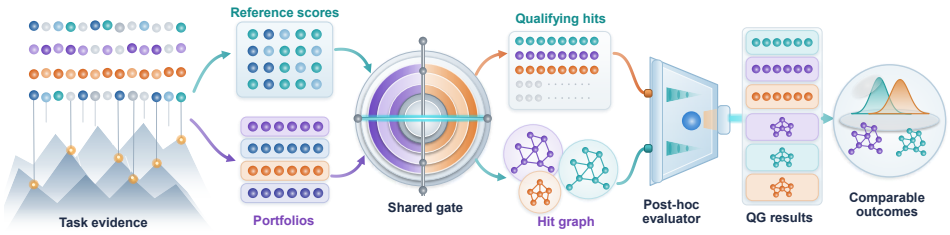


Figure 1: **Comparing sequence-design campaigns under a shared quality standard.** A fixed reference pool of recorded scores defines one threshold for each task. We apply this threshold to every method's evaluated candidates. *qg-hits* counts the candidates that meet the threshold. Among these candidates, sequences within the chosen Hamming distance are linked, and *qg-distinct* counts the resulting connected components. The evaluation uses existing scores and requires no additional measurements.

BEAM finds more distinct groups of high-quality sequences and achieves a higher mean score among its top candidates.

We also study candidate selection on experimentally measured protein fitness data from ProteinGym. We use LEF to select candidates based on ESM features from mutated positions. Across eight landscapes, this approach finds $42.9\%$ more high-quality candidates on average than proxy-greedy selection, while achieving similar coverage of sequence space. Experiments on six additional landscapes further support the gain in high-quality candidates. Together, these results show how a shared quality standard makes the output of a design campaign directly comparable across methods.

Our contributions are:

- **A shared standard for quality and diversity.** We formalize *qg-hits* and *qg-distinct* to measure how many candidates meet a common quality threshold and how many distinct sequence groups they cover.
- **A diversity ranking reversal on RNA tasks.** Under the same scoring model and measurement budget, TR-BEAM finds more distinct groups of high-quality sequences than $\delta$-CS, reversing their diversity ranking.
- **Evidence from measured protein fitness data.** Across eight ProteinGym landscapes, LEF finds more high-quality candidates at similar sequence-space coverage. Six additional landscapes support the gain in high-quality candidates.

## 2 RELATED WORK

**Model-guided sequence design and active learning.** Model-guided design uses a learned fitness model to select candidates. Active-learning variants refit that model after each measured batch. CbAS and DyNA-PPO combine learned proxies with generative or reinforcement-learning searches (Brookes et al., 2019; Angermueller et al., 2019). AdaLead follows a high-fitness local frontier (Sinai et al., 2020), PEX favors proximal exploration (Ren et al., 2022), and $\delta$-CS controls how far an off-policy generator moves from measured sequences (Kim et al., 2024). These methods improve which sequences are proposed. We study the evidence used to compare their completed portfolios. The oracle and budget must be matched, and the same declared quality standard must be applied to every portfolio.

**Diversity-aware generation and acquisition.** GFlowNets can represent several high-reward modes (Bengio et al., 2021; Malkin et al., 2022). GFlowNet-AL applies this idea to biological design and reports quality, diversity, and novelty among top-ranked candidates (Jain et al., 2022). UCB uses an upper-confidence acquisition rule. BALD selects candidates by predictive information gain (Houlsby et al., 2011). These approaches guide candidate generation or exploration, but a method-local top-$K$ statistic does not ensure a common fitness level. The $K$th candidate from two methods may clear very different quality bars. Diversity can then reward lower-quality spread.

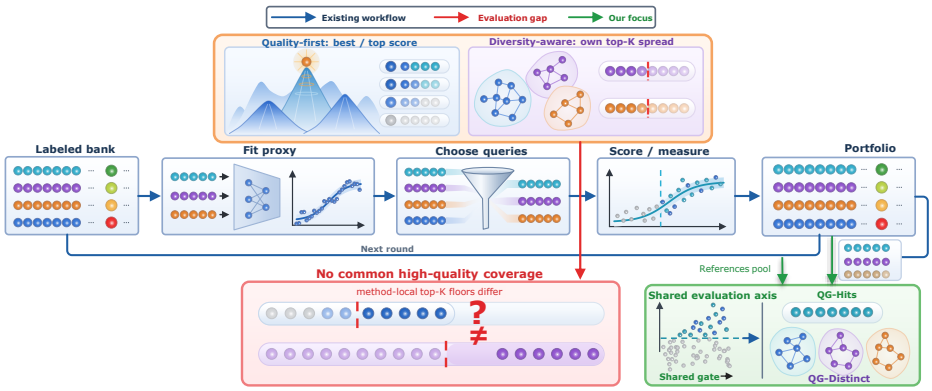


Figure 2: **A shared quality standard for sequence-design evaluation.** Method-local top sets can use different quality cutoffs. We apply one task-level threshold, then compare qualifying yield and sequence groups.

**Fitness and design benchmarks.** FLIP defines low-resource and extrapolative splits for protein fitness prediction (Dallago et al., 2021). ProteinGym provides a broader benchmark across many measured mutation-effect landscapes (Notin et al., 2023). Design-Bench standardizes datasets, oracles, and reporting rules for offline optimization (Trabucco et al., 2022). These benchmarks show the value of a common task contract. Our contribution is a common portfolio evaluator. It leaves the generator, acquisition rule, and query count unchanged. In the ProteinGym case study, mutation-position ESM features provide transferable residue information (Lin et al., 2023; Meier et al., 2021).

**Conditional quality–diversity evaluation.** Prior work establishes that quality and diversity should be read together. FLEXS reports candidates above a fixed fitness threshold and high-fitness optima found on enumerable landscapes (Sinai et al., 2020). GFlowNet-AL computes batch statistics on top-scoring candidates (Jain et al., 2022), and precision–recall views separate sample fidelity from coverage (Sajjadi et al., 2018). The Quality-Weighted Vendi Score combines item quality with similarity-based effective diversity. In active search, it measures Vendi diversity among discovered positives and uses this score to choose new queries; it also guides diverse Bayesian optimization (Nguyen & Dieng, 2024).

Our protocol addresses cross-method comparison after a campaign. A declared task-level reference pool defines one hard threshold for every method and paired run. *qg-hits* measures all portfolio members above this bar; *qg-distinct* counts connected components in their Hamming graph. This separates high-quality yield from high-quality coverage. It also exposes ranking reversals hidden by method-local top-$K$ sets. Unlike a soft weighted scalar, the hard gate keeps acceptance fixed while methods vary, and the two reported counts preserve the difference between yield and coverage.

## 3 Quality-Gated Evaluation

Quality-gated evaluation maps completed portfolios and recorded scores to two outcomes: high-quality yield and sequence-space coverage. It is an evaluator, not an acquisition rule. All methods on a task face the same quality threshold, and the search itself is unchanged. Figure 3 summarizes the approach.

### 3.1 Position in the Active-Learning Workflow

A sequence-design campaign starts from a labeled set. At each round, a method uses the available labels to fit a proxy or update its search state. Its proposal rule then generates new sequences or selects candidates from a fixed pool. The task-specific label source scores the selected batch, and

the new labels are added to the labeled set. Proxy-based methods refit on the expanded set before choosing the next batch. This loop continues until the shared query budget is exhausted.

Our two experimental settings instantiate this loop in different ways. In the RNA setting, TR-BEAM and $\delta$-CS generate new sequences and query the same seed-specific surrogate oracle within each paired run. In the ProteinGym setting, LEF and the comparison methods select candidates from the same measured DMS pool, and each query reveals an existing experimental score. These procedures differ in how they produce candidates, but both return a completed portfolio under a fixed task, initialization, label source, and query budget.

Quality-gated evaluation begins after this search process is complete. It takes the completed portfolios and their recorded scores as input, applies one declared task-specific quality threshold to every method, and reports high-quality yield and sequence-space coverage. It does not retrain a proxy, change a proposal rule, or affect any acquisition trajectory.

### 3.2 EVALUATION CONTRACT

For each comparison, we fix the task, scoring function, candidate space, initial data, query budget, and evaluated portfolio. Within a paired run, only the method-specific modeling and acquisition procedure changes. For RNA, the portfolio contains the newly acquired sequences. For ProteinGym, it contains the cumulative labeled set, including the shared initialization. A larger count therefore reflects a different use of the same measurement budget.

Quality gates are task-specific because assay scales differ. We compute each outcome within a task. For a cross-landscape ProteinGym summary, we average task-level outcomes and never average raw DMS scores across assays.

### 3.3 A SHARED QUALITY STANDARD

For task $t$ and paired run $s$, let $f_{t,s} : \mathcal{X}_t \to \mathbb{R}$ be the common scoring function and let $S_{tms}$ be the portfolio returned by method $m$. Selecting the top $\alpha$ fraction of each portfolio creates a method-local cutoff,

$$\tau^{\text{local}}_{tms} = Q_{1-\alpha}(\{f_{t,s}(x) : x \in S_{tms}\}) . \tag{1}$$

This rule targets the same portfolio fraction but not the same score level. Its diversity values can therefore compare candidates that meet different quality bars.

We instead declare a collection of reference scores for each task. Let $\mathcal{R}_t$ index its reference collections and let $A_{tj}$ be the score multiset in collection $j$. We define

$$\mathcal{P}_t = \biguplus_{j \in \mathcal{R}_t} A_{tj}, \qquad \tau_{t,q} = Q_q(\mathcal{P}_t) , \tag{2}$$

where $Q_q$ is the empirical $q$-quantile. The multiset union keeps repeated scores, so every recorded reference score contributes once. The same $\tau_{t,q}$ is then used for every method and paired run on task $t$.

The gate is computed after the campaigns from scores already recorded. It is not used for training or acquisition, requires no new oracle query, and leaves every portfolio unchanged. The reference-suite version and quantile are part of the evaluation contract because together they define the quality bar.

The pooling scope is part of the definition and changes the result. In the ProteinGym study $\mathcal{R}_t$ is the set of scored portfolios of all six compared methods across all five paired seeds on landscape $t$, so a landscape carries exactly one threshold and every seed of every method is gated at the same score. Taking the quantile within each seed instead raises the aggregate yield of LEF from $14.90$ to $15.18$ and shifts the other methods likewise. We therefore report the pooling scope alongside $q$ whenever a gate is quoted.

### 3.4 TWO COMPLEMENTARY OUTCOMES

We first count candidates that meet the shared threshold:

$$H_{tms}(q) = \{x \in S_{tms} : f_{t,s}(x) \geq \tau_{t,q}\}, \qquad \text{QGHits}_{tms}(q) = |H_{tms}(q)|. \tag{3}$$

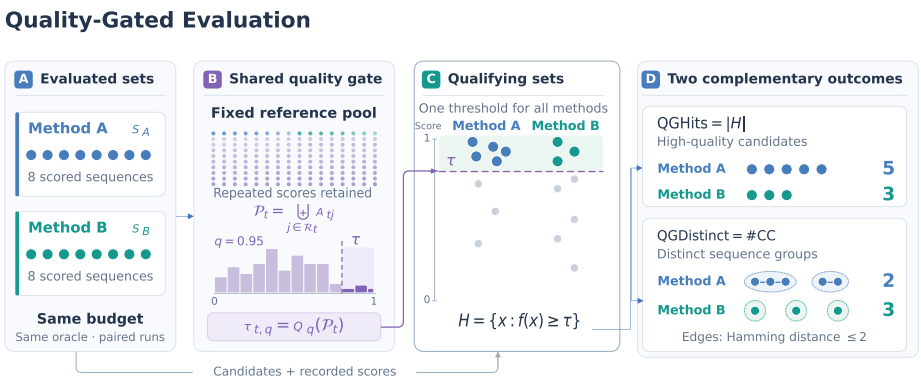


Figure 3: **Quality-gated evaluation.** A fixed pool of recorded scores defines one task-level threshold. We apply it to every portfolio produced under the same oracle and budget. *qg-hits* counts qualifying candidates; *qg-distinct* counts their Hamming-graph connected components. No additional oracle query is required.

*qg-hits* is the high-quality yield of the evaluated portfolio.

For coverage, we build a graph over $H_{tms}(q)$. Two candidates are linked when they differ at no more than $r$ positions. We define

$$\mathrm{QGDistinct}_{tms}(q, r) = \#\,\mathrm{CC}(H_{tms}(q), \{(x, x') : d_H(x, x') \leq r\}), \tag{4}$$

where $d_H$ is Hamming distance and CC denotes connected components. A chain of nearby variants forms one sequence group. Thus, dense sampling of one neighborhood increases yield without creating new groups.

The outcomes satisfy $0 \leq \mathrm{QGDistinct}_{tms} \leq \mathrm{QGHits}_{tms} \leq |S_{tms}|$. We report both instead of merging them into one score. This distinguishes greater yield, greater coverage, and a gain in both, while low-scoring spread cannot increase either count.

Our main setting is pooled $p_{95}$ ($q = 0.95$) with radius $r = 2$. Appendix C also reports $p_{90}$ and $r = 1$. The quantile sets the quality bar; the radius sets the resolution of sequence groups.

The radius has a validity condition. Any two variants that carry at most $k$ substitutions from a common wild type differ at no more than $2k$ positions, so a library of single substitutions places every qualifying candidate in one group once $r \geq 2$ and *qg-distinct* is identically one whatever the method did. The radius must therefore stay strictly below the mutational depth of the library. The landscapes we report carry between 1.85 and 13.56 mutated positions per variant, which is why $r = 2$ separates methods there; on a single-substitution assay only *qg-hits* is informative.

### 3.5 Evaluation procedure

For each task, we (1) freeze the portfolios, reference score collections, $q$, and $r$; (2) compute the single gate $\tau_{t,q}$; (3) filter every portfolio and count *qg-hits*; and (4) build each hit graph and count *qg-distinct*. We compute both values for each paired seed before aggregation.

### 3.6 Matched comparisons

The RNA methods generate sequences in the same four-token space and use the same seed-specific oracle. The ProteinGym methods select directly from the same measured variant pool. Within each paired run, we hold the initialization, oracle, budget, portfolio definition, reference pool, and metric settings fixed. Local search, uncertainty sampling, and generative policies can therefore be evaluated by the same final question: how many candidates clear one quality bar, and how many sequence groups do they cover?

# 4 EXPERIMENTAL SETUP

We test the evaluator in two settings. RNA requires each method to generate new sequences. ProteinGym requires each method to select from the same measured variant pool. Both follow the contract in Section 3.

**RNA design.** RNA1/2/3 are length-controlled, four-token tasks scored by seed-specific MLP oracles trained to approximate ViennaRNA folding scores (Lorenz et al., 2011). Within each paired seed, TR-BEAM and $\delta$-CS start from the same bank of 5,000 labeled sequences and receive 500 new queries (10 rounds of 50). We use seeds $\{0, 7, 12, 33, 44\}$.

TR-BEAM uses a one-hot MLP proxy ($512 \to 256 \to 1$, ReLU), trained for 200 epochs per round with Adam at $10^{-3}$. It starts from the top-50 labeled frontier, uses beam width 16, and performs eight local expansion steps with single-position substitutions. Random one-step proposals fill a batch when the beam returns fewer than 50 candidates. $\delta$-CS retains its published GFlowNet search pipeline, while its proposals are scored by the same seed-specific MLP oracle. Cumulative best uses the initial bank and all acquisitions. Mean-top-20, mean-top-50, *qg-hits*, and *qg-distinct* use only the 500 new sequences.

**ProteinGym.** The main pool study covers GB1, A4, YAP1, GRB2, PHOT, F7YBW8, HIS7, and GCN4 (Notin et al., 2023). Each run starts from 50 labeled variants and adds five batches of 20, giving a 150-label portfolio. Within each paired seed, all methods share the initialization, measured DMS values, and candidate pool. Pools are capped at 20,000 variants with sampling seed 123. The paired seeds are $\{0, 7, 12, 33, 44\}$.

LEF trains its proxy on ESM-2 representations at residues that differ from the wild type (Lin et al., 2023). Proxy-greedy uses the same ranking rule with one-hot features, so this comparison tests the representation. PEX denotes the project's pool-based adaptation of proximal exploration: a single-position mutation prior learned from strong labeled variants augments proxy ranking over the measured pool (Ren et al., 2022). UCB uses an upper-confidence acquisition rule, while BALD uses predictive information gain (Houlsby et al., 2011); random selection is proxy-free. All six methods choose directly from the same DMS pool, with no projection into it. Proxy-based methods refit on labels available at each round.

The evaluated portfolio contains all 150 labels. Separate representation analyses use fixed 150-label splits and held-out Spearman correlation to test proxy ranking outside the acquisition loop.

**Gates and aggregation.** For RNA, we use the archived pooled thresholds derived from the 22-configuration reference score suite. The primary ProteinGym reference-score multiset contains all scored portfolio occurrences from the six native selectors above, including their repeated shared initializations. We compute one threshold per task. The main setting is pooled $p_{95}$ and $r = 2$; $p_{90}$ and $r = 1$ provide sensitivity checks. Both outcomes are computed per paired seed, then averaged within each task. ProteinGym aggregates the eight landscape means with equal weight, so landscape size and assay scale do not control the result. Paired $t$-test markers in Table 1 use the shared seeds.

# 5 RESULTS

The experiments test four claims: a shared quality-and-coverage rule can change an evaluation ordering; *qg-hits* and *qg-distinct* separate yield from coverage; the evaluator can expose a useful representation choice; and the conclusions persist across evaluation settings.

## 5.1 A SHARED QUALITY-AND-COVERAGE RULE CHANGES THE RNA RANKING

TR-BEAM achieves higher quality on every reported summary (Table 1). Its mean-top-50 gains are $+0.088$, $+0.075$, and $+0.063$ on RNA1/2/3. Cumulative best and both portfolio summaries agree, so the gain extends beyond one exceptional sequence.

The evaluation ordering changes when we move from the method-local distinct-sequence statistic to the shared-gate radius-2 group statistic. The former gives official $\delta$-CS $50.0$ distinct top-ranked

Table 1: RNA quality under seed-matched MLP scoring and five paired runs. Cum-best spans the initial bank and acquisitions; mean-top metrics summarize the acquired set. $\Delta$ is TR-BEAM minus $\delta$-CS; stars denote paired $t$-test $p < 0.05$.

| Metric | Method | RNA1 | RNA2 | RNA3 |
|---|---|---|---|---|
| cum-best | TR-BEAM | 0.314 | 0.290 | 0.297 |
| | $\delta$-CS | 0.241 | 0.230 | 0.253 |
| | $\Delta$ | $+0.073^*$ | $+0.060^*$ | $+0.044^*$ |
| mean-top-20 | TR-BEAM | 0.305 | 0.282 | 0.290 |
| | $\delta$-CS | 0.222 | 0.212 | 0.231 |
| | $\Delta$ | $+0.083^*$ | $+0.070^*$ | $+0.059^*$ |
| mean-top-50 | TR-BEAM | 0.301 | 0.278 | 0.285 |
| | $\delta$-CS | 0.213 | 0.204 | 0.222 |
| | $\Delta$ | $+0.088^*$ | $+0.075^*$ | $+0.063^*$ |

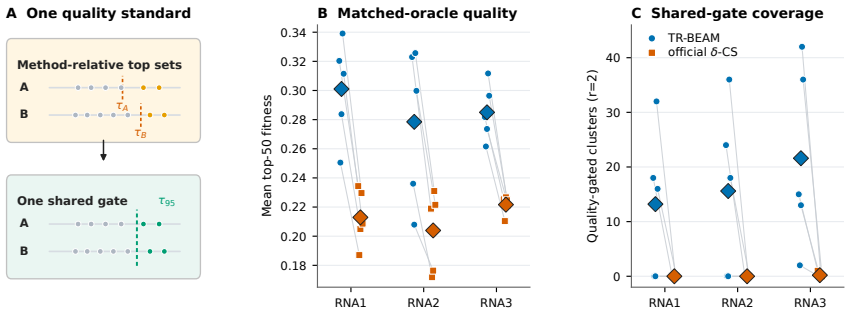


Figure 4: **A common quality-and-coverage rule changes the RNA comparison. (A)** Method-local and shared cutoffs. **(B)** Mean-top-50 fitness under matched oracles. **(C)** Radius-2 groups after the shared $p_{95}$ gate. Markers are paired seeds; diamonds are means.

Table 2: ProteinGym results under the six-method pooled-$p_{95}$ gate. Values first average five paired seeds per landscape and then average eight landscapes.

| Metric | LEF | Greedy | UCB | PEX | BALD | Random |
|---|---|---|---|---|---|---|
| *qg-hits* | **14.90** | 10.43 | 9.30 | 4.40 | 3.98 | 2.13 |
| *qg-distinct*, $r = 2$ | 4.93 | **4.98** | 4.25 | 2.88 | 2.10 | 2.10 |

sequences on each task and gives TR-BEAM 2.7 on average. Under our common rule, $\delta$-CS retains only $0.0/0.0/0.2$ radius-2 groups on RNA1/2/3, while TR-BEAM retains $13.2/15.6/21.6$. The paired means have this direction on all three tasks; paired evidence is strongest on RNA3 ($p = 0.047$; RNA1/2: $p = 0.095/0.090$). The quality-qualified radius-2 coverage of $\delta$-CS falls to near zero.

### 5.2 HIGHER YIELD AT MATCHED SEQUENCE-SPACE COVERAGE

On measured ProteinGym pools, LEF reaches 14.90 *qg-hits* and 4.93 radius-2 groups under the six-method pooled-$p_{95}$ gate (Table 2). Proxy-greedy reaches 10.43 and 4.98. Thus, LEF provides 42.9% more high-quality candidates at almost the same aggregate coverage. It also matches or exceeds UCB in hits on all eight landscapes and leads all six native selectors on five. Against proxy-greedy, its hit counts rise from 7.0 to 22.2 on A4 and from 8.8 to 18.2 on GRB2.

As a scale check, the six method means average 7.52 hits. This matches the 7.5 expected from a $p_{95}$ gate over equal 150-label portfolios, apart from ties.

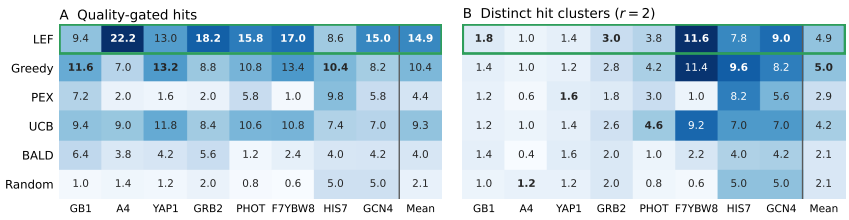


Figure 5: **High-quality yield and coverage on ProteinGym.** Cells are five-seed means under the six-method pooled $p_{95}$ gate. **Left:** hits in each 150-label portfolio. **Right:** radius-2 groups among those hits. Methods share the pool, DMS scores, budget, and initialization within each seed.

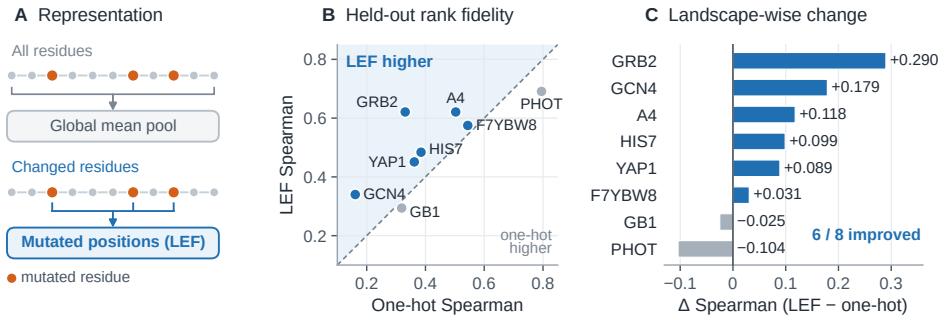


Figure 6: **Mutation-local features improve proxy ranking on six of eight landscapes. (A)** Feature construction. **(B)** Five-seed held-out Spearman correlation. **(C)** Landscape-wise gains.

### 5.3 Mutation-local representations sharpen protein ranking

On fixed 150-label splits, mutation-local ESM features improve held-out Spearman correlation over one-hot features on six of eight landscapes (Figure 6). The largest gains are GRB2 ($+0.290$), GCN4 ($+0.179$), and A4 ($+0.118$). On GB1, A4, and YAP1, mutation-position aggregation also exceeds global mean pooling: $.349/.621/.451$ versus $.221/.310/.066$ (Appendix G). Local aggregation retains signal that whole-sequence averaging can dilute.

The ranking gain is accompanied by more quality-gated hits while aggregate radius-2 coverage remains matched to proxy-greedy. This case shows how the evaluator connects a proxy choice to a campaign portfolio.

### 5.4 Consistent gains across operating points and landscapes

The ordering persists across operating points. At $p_{90}$, LEF averages $28.68$ hits, versus $19.98$ for greedy and $18.83$ for UCB; at $p_{95}$, the values are $14.90$, $10.43$, and $9.30$. With $r = 1$, LEF gives the largest coverage at both gates. An extended reference suite also preserves its lead: $11.70$ hits versus $8.33$ for greedy and $7.28$ for UCB. On six additional landscapes, LEF leads on five and averages $16.4$ hits, twice the $8.2$ of the next-best aggregate baseline. We fixed this set before its active-learning runs; Q8WTC7 was replaced before any run because its raw score scale did not match the shared proxy settings. The result therefore holds across quantiles, radii, reference sets, and assay collections. Appendix C and Appendix F give the full values.

The eight landscapes above were chosen while the method was being developed, so we fixed a selection rule in advance and applied it to assays never inspected. On that replication set LEF returns $1.89$ more qualifying candidates than proxy-greedy rather than $4.47$, and pooling gives $2.68$ over twenty-six landscapes at $p = 0.006$, where the original eight alone give $p = 0.073$. Coverage stays matched, with a *qg-distinct* difference of $-0.02$ and a $95\%$ interval of $[-0.18, +0.14]$. Appendix I gives the rule, the completion record, and the per-landscape values.

## 6 IMPLICATIONS

Quality-gated evaluation asks how much high-quality sequence space a fixed budget covers. A method-local cutoff can favor broad but weaker outputs. Our gate applies one minimum quality bar to every portfolio before coverage is counted.

*qg-hits* measures yield; *qg-distinct* measures how many sequence groups contain it. Reporting both separates depth within a strong neighborhood from breadth across neighborhoods. Cumulative best and mean-top-$K$ add the best point and the quality of the portfolio head.

RNA and ProteinGym reveal two effects. In RNA, the shared-gate group metric changes the ordering from a method-local distinct-sequence statistic. In ProteinGym, LEF and proxy-greedy cover almost the same number of high-quality groups, but LEF returns more qualifying sequences. The evaluator therefore links a representation choice to its portfolio without changing acquisition.

Future comparisons should fix the task, score source, budget, portfolio, reference suite, gate quantile, and distance rule. A frozen suite gives later methods the same bar; an expanded suite creates a named version. Reporting several $(q, r)$ pairs tests whether an ordering is stable.

When candidate identities and scores are saved, evaluation needs no new oracle calls. Hamming distance is transparent for the fixed-length spaces studied here; a validated structural or functional distance can replace it. The core contract stays fixed: one score, gate, distance, radius, and budget.

## 7 CONCLUSION

Biological sequence-design active learning should be credited for diversity only among candidates that clear a common quality bar. Quality-gated evaluation implements that principle with no extra oracle calls and exposes a clear RNA ranking reversal under seed-matched scoring. Across eight measured ProteinGym pools, mutation-local features produce substantially more high-quality hits at matched aggregate sequence-space coverage. Oracle matching within each paired run, native candidate spaces, paired seeds, and one common gate make portfolio diversity a comparable sequence-space outcome rather than a method-relative appearance. Appendix A records where the evaluation and the reported gains stop holding.

## ETHICS STATEMENT

This work is a purely computational study. All data used are publicly available and non-sensitive: standard benchmark oracles and published deep mutational scanning measurements. No human subjects, personally identifiable information, or newly generated wet-lab data are involved.

## REPRODUCIBILITY STATEMENT

Our code is available at the anonymized repository linked in the abstract, and the main text and appendices give the tasks, budgets, paired seeds, reference suites, gate construction, clustering parameters, and aggregation rules. Two offline commands regenerate every reported ProteinGym and RNA table from the released per-run records, so the values can be checked without rerunning a campaign. The gate depends on the pooling scope as well as the quantile, and the released analysis fixes both. The RNA surrogate oracles were trained inside each run and never stored, and CPU training of them is not bit-reproducible across thread counts, which moves individual scores by a few hundredths; we release oracles retrained from the same recipe with their measured disagreement against the recorded scores.

## AI USE STATEMENT

We used generative AI tools for feedback on experimental design, for interpreting and presenting results, for drafting and editing the manuscript, and for literature search and citation formatting. The authors verified all content and take responsibility for the final manuscript.

## REFERENCES


Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2019.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

David Brookes, Hahnbeom Park, and Jennifer Listgarten. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pp. 773–782. PMLR, 2019.

Jaime G Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pp. 290941–291025, 1998.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021–11, 2021.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International conference on machine learning*, pp. 9786–9801. PMLR, 2022.

Hyeonah Kim, Minsu Kim, Taeyoung Yun, Sanghyeok Choi, Emmanuel Bengio, Alex Hernández-García, and Jinkyoo Park. Improved off-policy reinforcement learning in biological sequence design. *arXiv preprint arXiv:2410.04461*, 2024.

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.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6 (1):26, 2011.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35: 5955–5967, 2022.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Quan Nguyen and Adji Bousso Dieng. Quality-weighted vendi scores and their application to diverse experimental design. *arXiv preprint arXiv:2405.02449*, 2024.

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.

Pierce J Ogden, Eric D Kelsic, Sam Sinai, and George M Church. Comprehensive aav capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science*, 366(6469):1139–1143, 2019.

Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, and Jian Peng. Proximal exploration for model-guided protein sequence design. In *International Conference on Machine Learning*, pp. 18520–18536. PMLR, 2022.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

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.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D Kelsic. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020.

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.

## A LIMITATIONS

**The coverage metric has a resolution floor.** Two variants carrying at most $k$ substitutions from a common wild type differ at no more than $2k$ positions. A single-substitution library therefore places every qualifying candidate in one group once $r \geq 2$, and *qg-distinct* equals one whatever the method did. We measured this on twelve ProteinGym assays whose libraries are purely single-substitution: all six methods score exactly $1.00$. The radius must stay strictly below the mutational depth of the library. The landscapes in Section 5.2 carry between $1.85$ and $13.56$ mutated positions per variant, which is the regime where $r = 2$ separates methods.

**Yield under a pooled gate measures a share.** When the gate is the $q$-quantile of the pooled scores of the compared methods, the mean *qg-hits* across those methods sits at $|S|(1 - q)$ by construction. Our measured means are $7.52$, $7.55$, and $7.54$ against a predicted $7.5$. A gain in *qg-hits* describes how a fixed quantity of qualifying candidates divides among the compared methods, and it moves when the comparison set changes. Table 5 shows the ordering survives an extended reference suite. Absolute values are comparable across papers only under a frozen external suite.

**Higher yield can mean depth rather than breadth.** On A4, LEF returns $22.2$ qualifying candidates that fall into $1.0$ sequence group, so its advantage there is depth inside one neighbourhood. Five of the eight main landscapes show a hits-to-groups ratio above four for LEF. Reporting the two outcomes separately is what makes this visible, and it applies to our method as much as to the baselines.

**Effect sizes shrink on landscapes fixed in advance.** The eight landscapes in the main text were chosen while the method was being developed. On the pre-registered replication set the advantage of LEF over proxy-greedy is $+1.89$ qualifying candidates rather than $+4.47$. Pooling both sets gives $+2.68$ over twenty-six landscapes at $p = 0.006$, where the original eight alone give $p = 0.073$. The pooled estimate is the one to carry forward.

**The landscape is the unit of generalization.** Each cell averages five paired seeds and each aggregate averages landscapes, so paired tests use the landscape as the unit. Pairing by seed would treat forty runs as independent and shrink the reported $p$-values. The RNA comparison spans three tasks and its $p$-values carry no multiplicity correction.

**Rescoring may cost the generative baseline.** In the RNA setting $\delta$-CS keeps its published search pipeline while its proposals are scored by our seed-matched surrogate. Matching the oracle is what makes the budget comparable, and reward scale also affects how a GFlowNet allocates probability mass, so part of the gap may follow from rescoring rather than from search.

**Distance is sequence-level and the score scale must match.** Hamming distance stands in for functional distance, which suits the fixed-length substitution spaces studied here. The shared proxy settings assume scores on a comparable scale: two assays whose score standard deviations exceed the

reported range by three orders of magnitude fall outside it, and a criterion fixed before their results were read places them out of scope. We report them unrescaled in the released artifact.

## B Evaluation Protocol Details

For every task, quality-gated evaluation is applied after all compared runs are complete:

1. Fix the campaign portfolios $S_{tms}$, paired-run scoring protocol, query budget, and reported reference score pool $\mathcal{P}_t$.
2. Compute the single threshold $\tau_{t,q}$ from Equation 2.
3. Count each portfolio member with score at least $\tau_{t,q}$.
4. Connect qualifying candidates at Hamming distance at most $r$ and count connected components.
5. Report seed-level counts together with $q$, $r$, and the fixed comparison set.

The operation is an evaluator: it changes neither candidate generation nor the number of oracle labels.

The RNA thresholds are fixed by the reported 22-configuration reference suite. The primary ProteinGym gate is computed from the six pool-native selectors. Three LEF–MMR variants (Carbonell & Goldstein, 1998) and one projected generative diagnostic form an extended reference-set analysis in Section C; they do not influence the primary gate.

Table 3: Task-level RNA gates from the fixed reference score suite.

| Task | Median | $\tau_{90}$ | $\tau_{95}$ |
|---|---|---|---|
| RNA1 | 0.1471 | 0.2694 | 0.2975 |
| RNA2 | 0.1374 | 0.2495 | 0.2874 |
| RNA3 | 0.1718 | 0.2519 | 0.2667 |

## C Threshold, Radius, and Reference-Set Sensitivity

Table 4 varies both the quality quantile and the Hamming radius. LEF retains the largest high-quality yield and radius-1 coverage at both gates. Table 5 adds the four archived auxiliary configurations to the reference set; the hit-yield ordering is unchanged.

Table 4: ProteinGym gate and radius sensitivity over eight landscapes and five paired seeds.

| Gate | Method | *qg-hits* | *qg-distinct* ($r = 1$) | *qg-distinct* ($r = 2$) |
|---|---|---|---|---|
| pooled $p_{90}$ | LEF | **28.68** | **16.45** | 7.00 |
| | Greedy | 19.98 | 13.85 | **8.08** |
| | UCB | 18.83 | 12.70 | 6.80 |
| pooled $p_{95}$ | LEF | **14.90** | **9.33** | 4.93 |
| | Greedy | 10.43 | 7.58 | **4.98** |
| | UCB | 9.30 | 6.60 | 4.25 |

Table 5: ProteinGym pooled-$p_{95}$ reference-set sensitivity. The primary gate uses the six pool-native selectors; the extended gate adds four archived auxiliary configurations.

| | Primary native gate | | Extended gate | |
|---|---|---|---|---|
| Method | *qg-hits* | *qg-distinct* ($r = 2$) | *qg-hits* | *qg-distinct* ($r = 2$) |
| LEF | **14.90** | 4.93 | **11.70** | **4.83** |
| Greedy | 10.43 | **4.98** | 8.33 | 4.40 |
| PEX | 4.40 | 2.88 | 3.90 | 2.68 |
| UCB | 9.30 | 4.25 | 7.28 | 3.98 |
| BALD | 3.98 | 2.10 | 3.35 | 1.98 |
| Random | 2.13 | 2.10 | 1.80 | 1.80 |

## D RNA Quality-Gated Results

Table 6: RNA quality-gated counts among the 500 additionally queried sequences, averaged across five paired runs. Each cell is *qg-hits*/*qg-distinct* at radius 2.

| Task | TR-BEAM, $p_{90}$ | $\delta$-CS, $p_{90}$ | TR-BEAM, $p_{95}$ | $\delta$-CS, $p_{95}$ |
|---|---|---|---|---|
| RNA1 | 268.8/26.4 | 0.4/0.4 | 172.6/13.2 | 0.0/0.0 |
| RNA2 | 262.2/32.4 | 1.2/1.2 | 156.6/15.6 | 0.0/0.0 |
| RNA3 | 225.0/39.0 | 0.8/0.8 | 134.6/21.6 | 0.2/0.2 |

## E ProteinGym Per-Landscape Results

Table 7: Pool-native ProteinGym results at the six-method pooled $p_{95}$ gate, averaged across five paired seeds. Each cell is *qg-hits*/*qg-distinct* at radius 2.

| Landscape | LEF | Greedy | PEX | UCB | BALD | Random |
|---|---|---|---|---|---|---|
| GB1 | 9.4/1.8 | 11.6/1.4 | 7.2/1.2 | 9.4/1.2 | 6.4/1.4 | 1.0/1.0 |
| A4 | 22.2/1.0 | 7.0/1.0 | 2.0/0.6 | 9.0/1.0 | 3.8/0.4 | 1.4/1.2 |
| YAP1 | 13.0/1.4 | 13.2/1.2 | 1.6/1.6 | 11.8/1.4 | 4.2/1.6 | 1.2/1.2 |
| GRB2 | 18.2/3.0 | 8.8/2.8 | 2.0/1.8 | 8.4/2.6 | 5.6/2.0 | 2.0/2.0 |
| PHOT | 15.8/3.8 | 10.8/4.2 | 5.8/3.0 | 10.6/4.6 | 1.2/1.0 | 0.8/0.8 |
| F7YBW8 | 17.0/11.6 | 13.4/11.4 | 1.0/1.0 | 10.8/9.2 | 2.4/2.2 | 0.6/0.6 |
| HIS7 | 8.6/7.8 | 10.4/9.6 | 9.8/8.2 | 7.4/7.0 | 4.0/4.0 | 5.0/5.0 |
| GCN4 | 15.0/9.0 | 8.2/8.2 | 5.8/5.6 | 7.0/7.0 | 4.2/4.2 | 5.0/5.0 |
| Mean | **14.90**/4.93 | 10.43/**4.98** | 4.40/2.88 | 9.30/4.25 | 3.98/2.10 | 2.13/2.10 |

Under the archived extended gate, LEF with MMR coefficients $0.5$, $1$, and $2$ yields aggregate *qg-hits* of $12.10$, $12.13$, and $12.10$, and radius-2 *qg-distinct* of $4.90$, $4.85$, and $5.15$. Its quality-gated counts remain similar across these reranking coefficients.

## F Additional ProteinGym Landscapes

Six further landscapes span binding, fluorescence, capsid fitness, abundance, and thermostability assays (Sarkisyan et al., 2016; Ogden et al., 2019). In this separate evaluation with 150-label portfolios and five paired seeds, LEF achieves the largest hit count on five of six and averages $16.4$ hits; proxy-greedy is next at $8.2$.

Table 8: Six additional ProteinGym landscapes, averaged across five paired seeds. Each cell is *qg-hits*/*qg-distinct* at radius 2.

| Landscape | LEF | Greedy | PEX | Random | BALD | UCB |
|---|---|---|---|---|---|---|
| GB1–Olson | 21.8/1.8 | 6.6/2.4 | 5.4/2.0 | 2.6/2.4 | 3.2/2.2 | 5.4/2.0 |
| GFP | 13.8/8.4 | 6.8/6.8 | 11.2/2.0 | 5.4/5.4 | 1.8/1.8 | 6.0/6.0 |
| CAPSD | 16.4/2.8 | 4.4/4.2 | 13.6/1.8 | 4.0/4.0 | 2.8/2.8 | 4.0/2.6 |
| RASK | 17.8/1.2 | 11.0/1.2 | 3.4/1.4 | 1.6/1.4 | 1.8/1.0 | 9.6/1.6 |
| PABP | 10.0/2.4 | 12.2/3.2 | 3.8/2.2 | 4.6/4.4 | 4.2/2.4 | 10.8/3.4 |
| HECD1 | 18.4/1.2 | 8.2/1.0 | 5.0/1.0 | 1.6/1.2 | 4.2/1.2 | 7.6/1.2 |

## G Localized ESM Feature Ablations

Table 9: Held-out Spearman correlation on fixed 150-label splits, averaged across three seeds.

| Landscape | one-hot | mean-pool | mutpos | delta | concat |
|---|---|---|---|---|---|
| GB1 | .289 | .221 | **.349** | .347 | .325 |
| A4 | .471 | .310 | .621 | .489 | **.666** |
| YAP1 | .381 | .066 | .451 | .510 | **.511** |

Table 10: Held-out Spearman correlation for one-hot and mutation-position ESM features on all eight landscapes, averaged across five seeds.

| Landscape | one-hot | LEF | Δ | Landscape | one-hot | LEF | Δ |
|---|---|---|---|---|---|---|---|
| GB1 | .319 | .294 | -.025 | PHOT | .795 | .692 | -.104 |
| A4 | .494 | .612 | +.118 | F7YBW8 | .544 | .574 | +.031 |
| YAP1 | .363 | .451 | +.089 | HIS7 | .385 | .484 | +.099 |
| GRB2 | .331 | .621 | +.290 | GCN4 | .161 | .340 | +.179 |

## H Reproducibility Checklist

The evaluation bundle records task definitions, paired seed lists, budgets, reported reference suites, gate construction, clustering code, and data provenance. Cross-landscape tables first average seeds within a landscape and then average landscape means; raw DMS scores are never averaged across assays.

## I The Pre-Registered Replication

The landscapes above were chosen while the method was being developed, so we fixed a selection rule in advance and applied it to assays never inspected: every ProteinGym substitution assay with at least $2500$ scored variants, excluding those already examined, one assay per protein, ranked by pool size, top twenty-four. Twenty completed; four exceeded the memory available to the language model on the assigned hardware, a cause independent of any result. A score-scale criterion, also fixed in advance, places two of the twenty outside the range the shared proxy settings assume. The rule, the criterion, their timestamp, and the record of which landscapes completed are released with the code.

On the replication set LEF returns $1.89$ more qualifying candidates than proxy-greedy, against $4.47$ on the original set, leading on twelve of eighteen. Pooling gives $2.68$ over twenty-six landscapes at $p = 0.006$ (Wilcoxon $p = 0.006$, leading on seventeen), where the original eight alone give $p = 0.073$. A curated set overstates the gap and a set fixed in advance still shows it, so the pooled estimate is the one to carry forward. Coverage stays matched: the *qg-distinct* difference is $-0.02$ with a $95\%$ interval of $[-0.18, +0.14]$.