FROM PLAUSIBILITY TO RELIABILITY: FAILURE-AWARE CLOSED-LOOP LEARNING FOR PROTEIN DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein design is increasingly able to produce large sets of computationally plausible candidates, while physical measurements remain the scarce resource that establishes reliability. We study folding stability as a controlled reliability endpoint and formulate measurement selection as a sequential learning problem under sparse physical feedback. We test whether an acquisition objective that combines model disagreement with predicted low stability can improve discovery of low-stability candidates under a fixed budget. PRIB Failure-Aware implements this objective as a scale-free rank aggregation of bootstrap-ensemble disagreement and failure relevance. In a frozen retrospective replay over 30,000 MGnify Stability sequences, each acquisition round reveals a batch of previously hidden experiment-derived ΔG targets and refits a three-member Ridge ensemble on frozen ESM-2 representations. Across five algorithmic seeds and cumulative budgets from 512 to 8,192 labels, PRIB achieves higher low-stability recall than Random and generic uncertainty sampling at every checkpoint. At 8,192 labels, recall is $53.95 \pm 0.77\%$, compared with $27.62 \pm 0.81\%$ for Random and $27.55 \pm 1.01\%$ for Uncertainty; at 4,096 labels, PRIB already exceeds both baselines at 8,192 labels. On this fixed pool, the direction holds in all 25 seed-by-budget comparisons against each baseline. A target-provenance audit fixes the released ΔG convention before production replay, and predictive test evaluation is reserved for the final 8,192-label checkpoint. These results show that the acquisition objective can materially change which low-stability risks are resolved within a finite measurement budget and motivate experimental allocation as a first-class component of protein reliability workflows.

1 INTRODUCTION

Deep-learning methods now support de novo protein design, structure prediction, and low-data protein engineering at scales that were recently impractical (Watson et al., 2023; Lin et al., 2023; Biswas et al., 2021; Jiang et al., 2025). At the same time, high-throughput biophysical assays can measure folding stability across hundreds of thousands to millions of small protein domains (Tsuboyama et al., 2023; Cho et al., 2026a). Together, these advances expand the number of plausible candidates that can be scored computationally and the amount of physical evidence that can be collected experimentally. They also sharpen a practical decision problem: experimental capacity remains finite, so a reliability workflow must decide which unresolved candidates are most valuable to measure next.

Active learning treats labels as a resource and chooses measurements according to their expected information or decision value (Settles, 2009; Cohn et al., 1996). For reliability screening, measurement utility also depends on where a candidate lies relative to an operational risk region. We therefore test a simple acquisition objective that jointly ranks ensemble disagreement and predicted low stability. The rule gives equal percentile-rank weight to the two terms, avoiding any assumption that their raw numerical scales are directly comparable.

We evaluate this question in a retrospective closed-loop replay with an offline oracle. At each round the system predicts over the unrevealed pool, selects the next batch, reveals the corresponding experiment-derived stability targets, records the decision state and observation, and refits before the next round. This state-transition structure follows the experiment-planning logic used in active

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.

learning and self-driving laboratories (Häse et al., 2019; Burger et al., 2020). The MGnify Stability pool provides a controlled testbed for the post-generation measurement-allocation stage that arises when many candidates can be scored but only a subset can be assayed. We ask: *under a fixed experimental budget, can failure-aware sequential acquisition recover experimentally low-stability proteins more efficiently than passive sampling or generic uncertainty sampling?*

Our contributions are threefold. First, we formulate protein reliability screening as sequential evidence allocation and implement an auditable closed-loop replay with explicit separation between pre-reveal model state and newly observed physical evidence. Second, we introduce a rank-based failure-aware acquisition rule that combines ensemble disagreement with predicted proximity to the low-stability region, directly matching the reliability objective. Third, on a frozen 30,000-sequence pool with five algorithmic seeds and a fixed budget grid, we show that this objective changes where experiments are spent and substantially increases low-stability discovery relative to Random and generic uncertainty sampling. The complete registry, target-provenance audit, validation diagnostics, and final-checkpoint test diagnostics preserve the evidence trail for each conclusion.

2 Related Work

Protein stability and representations. Protein language models provide transferable sequence representations that can support downstream structure and property modeling (Lin et al., 2023). In parallel, cDNA-display proteolysis has enabled high-throughput thermodynamic folding-stability measurement (Tsuboyama et al., 2023). The MGnify Stability study extends this experimental scale to a diverse collection of short protein domains and releases the machine-learning index used in our replay (Cho et al., 2026a;b). Our controlled study freezes the ESM-2 representation and uses a lightweight Ridge ensemble so that the experimental variable of interest is the acquisition policy.

Active learning and closed-loop experimentation. Machine-learning-guided protein engineering uses sparse experiments to learn sequence–function relationships and prioritize promising variants (Yang et al., 2019; Biswas et al., 2021). Recent prospective campaigns iterate between model updates and physical screening, including active-learning-assisted directed evolution and PLM-guided evolution (Yang et al., 2025; Jiang et al., 2025). Pool-based active learning provides a complementary language for experimental allocation, with uncertainty sampling as a widely used query strategy (Settles, 2009; Cohn et al., 1996). Batch-aware methods such as BatchBALD model dependencies among selected points and can reduce redundant acquisitions (Kirsch et al., 2019). Protein-engineering benchmarks also show that the value of uncertainty-based acquisition depends on the landscape, uncertainty estimator, and evaluation objective (Greenman et al., 2025). Our study focuses this decision problem on reliability triage: the reward is discovery of an experimentally defined low-stability region under a fixed measurement budget, and the closed loop is evaluated with a label-hiding oracle and a complete evidence registry.

3 Problem Formulation and Method

3.1 Budgeted evidence allocation

Let $\mathcal{P} = \{x_i\}_{i=1}^{N}$ be a fixed candidate pool with hidden experimental targets y_i. At round t, the revealed set is $\mathcal{L}_t$ and the available candidates are $\mathcal{U}_t = \mathcal{P} \setminus \mathcal{L}_t$. A policy chooses a batch $B_t \subset \mathcal{U}_t$, the offline oracle reveals $\{y_i : i \in B_t\}$, and the state is appended to a registry before the model is refit. The event of interest is low stability,

$$e_i = \mathbb{1}[y_i \leq 0 \text{ kcal/mol}], \tag{1}$$

For this study, reliability screening is operationalized through this endpoint: "failure-aware" refers to acquisition targeted toward the predefined low-stability event. Under the dataset convention, larger ΔG denotes greater folding stability and $\Delta G = 0$ is the operational free-energy boundary between folded and unfolded states; the threshold was fixed before production replay (Tsuboyama et al., 2023). The primary metric at cumulative budget b is the fraction of all low-stability pool members revealed by that budget,

$$\text{Recall}_{\text{fail}}(b) = \frac{\sum_{i \in \mathcal{L}_b} e_i}{\sum_{i \in \mathcal{P}} e_i}. \tag{2}$$

This metric quantifies how much of the predefined low-stability region has been resolved by cumulative budget b.

3.2 Frozen representation and uncertainty model

Each sequence is represented by the final hidden state of the frozen `facebook/esm2_t36_3B_UR50D` checkpoint, mean-pooled over non-special amino-acid tokens. The representation has dimension 2,560 and uses the unadapted base checkpoint throughout the replay. We fit a three-member bootstrap Ridge regression ensemble to the currently revealed targets. For an unlabeled candidate x at round t, its ensemble mean and memberwise standard deviation are

$$\mu_t(x) = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{m,t}(x), \qquad U_t(x) = \sqrt{\frac{1}{M}\sum_{m=1}^{M} \left(\hat{y}_{m,t}(x) - \mu_t(x)\right)^2}, \tag{3}$$

with $M = 3$. Thus U_t is the population standard deviation across the three ensemble members and serves as an ordinal dispersion score. The lightweight learner keeps representation and regression capacity fixed across policies, concentrating the comparison on how measurements are allocated.

3.3 Failure-aware acquisition

We compare three policies fixed in the experiment configuration before production replay. **Random** samples uniformly from the unrevealed pool. **Uncertainty** selects candidates with the largest $U_t(x)$. **PRIB Failure-Aware** uses a failure relevance score $F_t(x) = -\mu_t(x)$, so lower predicted stability receives a larger score. Let $R_t(s; x) \in [0, 1]$ denote the ascending average-percentile rank of $s(x)$ within $\mathcal{U}_t$: the largest score receives rank one, and ties receive their average rank. PRIB independently rank-normalizes uncertainty and failure relevance, then defines

$$A_{\text{PRIB},t}(x) = R_t(U_t; x) + R_t(F_t; x), \qquad B_t = \mathrm{TopK}_{x\in\mathcal{U}_t}\left(A_{\text{PRIB},t}(x), k_t\right), \tag{4}$$

where $k_t = |B_t|$ is the pre-specified batch size and TopK returns the k_t largest scores. The weights are fixed to $1 : 1$, with no diversity term and no post-hoc weight sweep. Because only within-pool ranks enter the score, the ordering is unchanged by independently strictly increasing reparameterizations of uncertainty or failure relevance; this avoids treating the two quantities as calibrated on a common physical scale. The score therefore favors candidates that are simultaneously high-disagreement and low-predicted-stability; the experiment evaluates this combined acquisition rule as a whole.

3.4 Closed-loop state and evidence registry

Figure 1 depicts the replay. The registry records seed, cumulative budget, candidate identity, prediction, uncertainty, acquisition score, selected rank, label-reveal transition, and observed outcome. Pre-reveal quantities are therefore preserved separately from the subsequently observed target, and every acquisition can be traced through the full predict–select–observe–update cycle. After configuration and launch, the replay proceeds across rounds without human candidate selection; the observation interface is an offline oracle over previously measured stability values.

4 Experimental Setup

Target and provenance. The primary oracle is the official `mgnify_training_index.csv` `dG` value clipped to $[-1, 5]$ kcal/mol, matching the released benchmark convention (Cho et al., 2026a;b). The low-stability event is $\Delta G \leq 0$ kcal/mol and its membership is invariant to this clipping transform. A provenance audit resolves the target convention before production replay: the clipped official index exactly reproduces all 3,283 values in the released test benchmark (MAE, RMSE, and maximum absolute error each 0), while the recalibrated full table contains non-unique final sequence-level targets for 14,149 indexed entities. We retain the full recalibrated table as a provenance and sensitivity record.

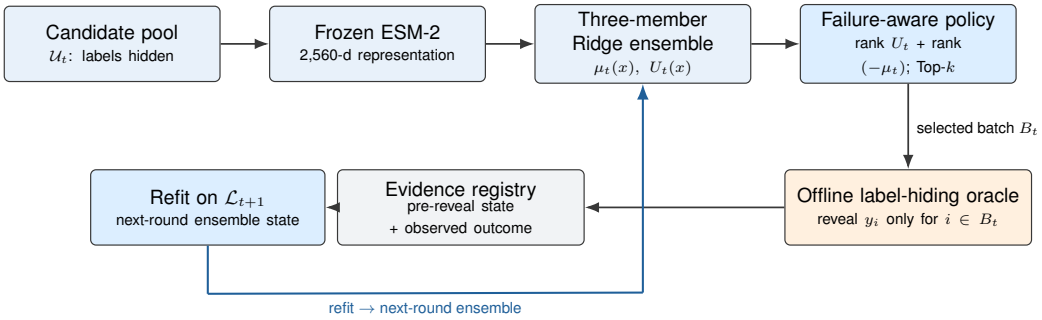


Figure 1: Auditable retrospective closed-loop replay. A frozen ESM-2 representation feeds a three-member Ridge ensemble that produces μ_t and U_t; the acquisition policy selects a batch while its targets remain hidden. The offline oracle then reveals the selected experiment-derived ΔG values, the transaction is written to the evidence registry, and the learner is refit for the next round.

Replay protocol. We sample one fixed pool of 30,000 official training-split sequences using seed 20260922; 3,865 pool members (12.883%) meet the low-stability event. Every policy starts from the same 256 revealed targets within an algorithmic seed. We evaluate cumulative budgets of 512, 1,024, 2,048, 4,096, and 8,192 labels over five algorithmic seeds (0–4). Within the production replay, predictive test evaluation occurs only at the final 8,192-label checkpoint. The separate provenance audit uses the released test-benchmark targets solely to resolve the published target convention before replay; test predictive performance plays no role in policy, weight, or checkpoint selection.

Metrics and comparisons. The primary outcome is cumulative low-stability recall. We also report cumulative failure enrichment (observed low-stability proportion divided by pool prevalence), validation Spearman correlation, and final-checkpoint test Spearman correlation as diagnostics. Means and sample standard deviations summarize variation over the five algorithmic seeds. These seeds vary initialization and replay randomness on the same fixed pool, so the reported dispersion summarizes algorithmic variability on that fixed evaluation setting.

5 Results

5.1 RQ1: Does failure-aware selection improve low-stability discovery over passive allocation?

On the frozen budget grid, PRIB Failure-Aware exceeds Random at every checkpoint (Figure 2). At 8,192 labels, it reaches $53.95 \pm 0.77\%$ low-stability recall, versus $27.62 \pm 0.81\%$ for Random. In counts, this is 2,085.0 of 3,865 low-stability pool members on average for PRIB, compared with 1,067.4 for Random. The result is about $1.95\times$ the Random recall at that checkpoint.

5.2 RQ2: Does failure-aware allocation improve over generic uncertainty sampling?

On the fixed 30,000-sequence pool, PRIB Failure-Aware exceeds the tested generic Uncertainty policy at every checkpoint, reaching $53.95 \pm 0.77\%$ versus $27.55 \pm 1.01\%$ recall at 8,192 labels ($1.96\times$). Across the 25 paired seed-by-budget comparisons, PRIB has higher recall in all 25 comparisons against Random and all 25 against Uncertainty. These comparisons establish fixed-pool algorithmic consistency; transfer across independently constructed pools remains a separate generalization test.

5.3 RQ3: How does the allocation behavior differ?

The selection diagnostic in Figure 3 visualizes the pre-reveal prediction–uncertainty plane for the two active policies. Generic Uncertainty concentrates on high-disagreement candidates across the predicted-stability range, whereas PRIB shifts acquisition toward candidates that combine appreciable disagreement with low predicted stability, matching its fixed rank-sum objective. At the final budget,

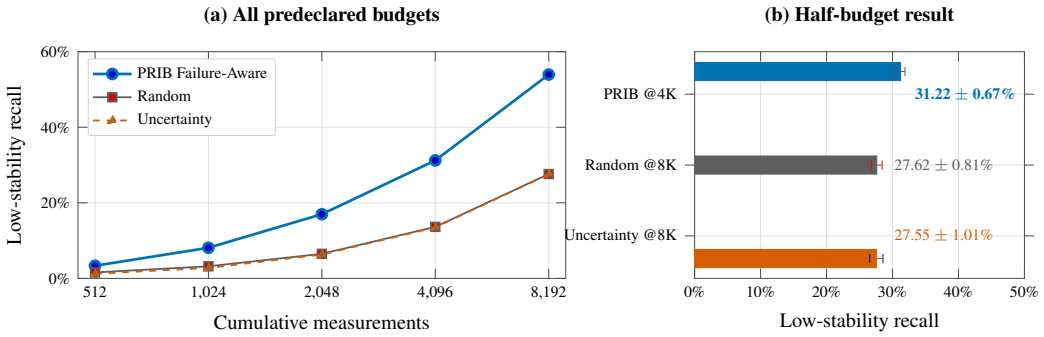


Figure 2: Primary outcome and label-efficiency comparison. **Left:** cumulative low-stability recall over all pre-specified budgets; markers and error bars are mean ± sample standard deviation across five algorithmic seeds. **Right:** at 4,096 labels, PRIB reaches $31.22 \pm 0.67\%$ recall, exceeding both 8,192-label baselines ($27.62 \pm 0.81\%$ Random; $27.55 \pm 1.01\%$ Uncertainty). All plotted values are frozen in `F2_data.csv`.

Table 1: Compact result summary (mean ± sample standard deviation over five algorithmic seeds). Higher recall and enrichment indicate more observed low-stability sequences. Validation and final-checkpoint test Spearman are predictive diagnostics; within the production replay, the test set is evaluated only at 8,192 labels.

Policy	R@512	R@8K	Enrich.@8K	Val. ρ@8K	Test ρ@8K
Random	$1.58 \pm 0.20\%$	$27.62 \pm 0.81\%$	1.011 ± 0.030	0.661 ± 0.041	0.582 ± 0.090
Uncertainty	$1.22 \pm 0.22\%$	$27.55 \pm 1.01\%$	1.009 ± 0.037	0.653 ± 0.015	0.520 ± 0.043
PRIB Failure-Aware	$3.35 \pm 0.37\%$	$53.95 \pm 0.77\%$	1.976 ± 0.028	0.640 ± 0.030	0.641 ± 0.010

validation Spearman is 0.640 for PRIB, 0.661 for Random, and 0.653 for Uncertainty (Table 1); the final held-out test diagnostic is 0.641, 0.582, and 0.520, respectively. The combination of strong recall gains with comparable global-regression diagnostics is consistent with the policy acting primarily through evidence allocation toward the operational low-stability region.

5.4 Label efficiency on the tested grid

At each adjacent budget doubling, PRIB at the smaller tested budget has greater mean low-stability recall than both baselines at the next doubled budget. For example, at 4,096 labels PRIB reaches 31.22% recall, compared with 27.62% for Random and 27.55% for Uncertainty at 8,192 labels. We use this fixed-grid comparison as a descriptive measure of label efficiency; intermediate budgets were not evaluated.

6 Discussion and Limitations

The central result is a decision-level effect. Under the frozen target definition, pool, representation, learner, and budget schedule, failure-aware acquisition directs substantially more of the measurement budget toward the low-stability region than Random or generic ensemble uncertainty. At the final budget, this produces approximately $1.95\times$ the recall of either baseline while global validation Spearman remains comparable across policies. Within this replay, optimizing the allocation objective therefore changes which low-stability cases are resolved under a finite experimental budget, with no corresponding increase in global validation correlation.

The present evidence covers one experiment-derived folding-stability endpoint, one fixed 30,000-sequence pool, five algorithmic seeds, and two reference policies. Mechanistic attribution remains incomplete: a failure-relevance-only policy is needed to isolate the incremental value of ensemble disagreement, and a diversity-aware batch policy would test whether redundancy matters as acquisition batches grow (Kirsch et al., 2019). Generalization further requires independently constructed pools,

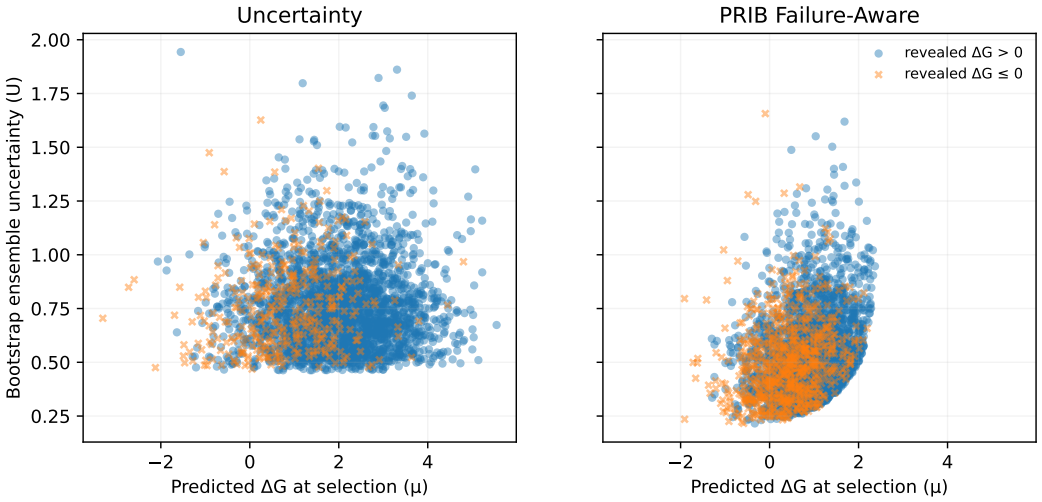


Figure 3: Selection-time diagnostic for the active policies. Each point stores the pre-reveal predicted stability and ensemble disagreement associated with an acquired candidate; color/marker indicates the subsequently revealed ΔG outcome. For legibility, the frozen plotting script samples up to 3,000 acquired candidates per policy with random state 20260923. PRIB concentrates acquisition in the joint low-predicted-stability and uncertain region targeted by its fixed rank-sum score.

protein families, and assay endpoints. Prospective assays would test the same decision rule under real experimental latency, noise, and operational constraints.

These next experiments follow directly from the hypothesis tested here. The hypothesis predicts the largest benefit when physical measurements are scarce, model disagreement is heterogeneous, and unresolved candidates occupy a failure-relevant region of the phenotype space. Testing that prediction across stability, expression, aggregation, and condition-specific stress endpoints would move reliability assessment from a single-target replay toward a broader experimental decision framework for protein design.

7 Conclusion

We formulated protein reliability as a sequential evidence-allocation problem and tested a failure-aware acquisition rule in an auditable retrospective closed loop. On the fixed MGnify Stability pool and budget grid, the combined rank objective recovered substantially more low-stability sequences than Random or generic uncertainty sampling, reaching 53.95% recall at 8,192 measurements compared with about 27.6% for both baselines. The mixed global-regression diagnostics indicate that the observed advantage is concentrated in evidence allocation. The broader design principle is that, once candidate generation outpaces measurement capacity, reliability assessment should optimize which physical evidence to acquire. Independent pools, additional reliability endpoints, and prospective assays provide the next tests of this principle.

AI Use Statement

For this submission, generative AI tools assisted with drafting and editing prose and with LaTeX figure layout. Experimental data, numerical results, and frozen scientific artifacts were produced by the recorded computational workflow and retained in the evidence bundle. The authors independently checked all reported formulae, citations, numerical claims, and figures against those records and take responsibility for the final content, including text and artifacts produced with generative-AI assistance.

Reproducibility Statement

Sections A–D specify the oracle, target transform, split reconciliation, fixed pool, representation, model hyperparameters, deterministic replay rules, metrics, and access guard. Appendix B reports the complete five-budget result grid and final test diagnostics. The frozen handoff records contain the machine-readable configuration, metric table, selection registry, oracle-access log, artifact manifest, and a numerical-claim audit; the appendix identifies the exact fields that support each reported result.

References

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, 2021. doi: 10.1038/s41592-021-01100-y. URL https://doi.org/10.1038/s41592-021-01100-y.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, et al. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020. doi: 10.1038/s41586-020-2442-2. URL https://doi.org/10.1038/s41586-020-2442-2.

Yehlin Cho, Kotaro Tsuboyama, Theodore J. Litberg, Michelle D. Jung, Adunoluwa Obisesan, Qian Wang, Claire M. Phoumyvong, Jane Thibeault, Sergey Ovchinnikov, and Gabriel J. Rocklin. Accurate protein stability prediction for small domains using mega-scale experiments. *bioRxiv*, 2026a. doi: 10.64898/2026.05.19.726285. URL https://doi.org/10.64898/2026.05.19.726285. Preprint.

Yehlin Cho, Kotaro Tsuboyama, Theodore J. Litberg, Michelle D. Jung, Adunoluwa Obisesan, Qian Wang, Claire M. Phoumyvong, Jane Thibeault, Sergey Ovchinnikov, and Gabriel J. Rocklin. Accurate protein stability prediction for small domains using mega-scale experiments: data and source files, 2026b. URL https://doi.org/10.5281/zenodo.19411306.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. doi: 10.1613/jair.295. URL https://doi.org/10.1613/jair.295.

Kevin P. Greenman, Ava P. Amini, and Kevin K. Yang. Benchmarking uncertainty quantification for protein engineering. *PLOS Computational Biology*, 21(1):e1012639, 2025. doi: 10.1371/journal.pcbi.1012639. URL https://doi.org/10.1371/journal.pcbi.1012639.

Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1(3):282–291, 2019. doi: 10.1016/j.trechm.2019.02.007. URL https://doi.org/10.1016/j.trechm.2019.02.007.

Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R. Sgrizzi, Lukas Villiger, Alisan Kayabolen, B. J. Kim, Josephine K. Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, Jonathan S. Gootenberg, and Omar O. Abudayyeh. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science*, 387(6732):eadr6006, 2025. doi: 10.1126/science.adr6006. URL https://doi.org/10.1126/science.adr6006.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison Department of Computer Sciences, 2009. URL https://minds.wisconsin.edu/handle/1793/60660.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023. doi: 10.1038/s41586-023-06328-6. URL https://doi.org/10.1038/s41586-023-06328-6.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Jason Yang, Ravi G. Lal, James C. Bowden, Raul Astudillo, Mikhail A. Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H. Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16:714, 2025. doi: 10.1038/s41467-025-55987-8. URL https://doi.org/10.1038/s41467-025-55987-8.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, 2019. doi: 10.1038/s41592-019-0496-6. URL https://doi.org/10.1038/s41592-019-0496-6.

A Detailed Frozen Experimental Specification

A.1 Scope of this appendix

This appendix documents the completed frozen Q17 replay in full. The retained archive contains the configuration, replay implementation, output metrics, selection registry, oracle-access log, and a manifest of artifact hashes. These records make the protocol inspectable from the target definition through each acquisition decision and final metric.

A.2 Canonical records, split reconciliation, and oracle

The official `mgnify_training_index.csv` contains 965,856 canonical rows with columns `name`, `split`, `seq`, `dG`, and `PDB_name`. The canonical split counts are 960,216 training, 2,357 validation, and 3,283 test records. The source labels `validation` (1,761 rows) and `validation_online` (596 rows) are reconciled to the single validation split only after the exact canonical-total check; the reconciliation preserves all 965,856 records exactly once.

The oracle is the official-index `dG` value after clipping to $[-1, 5]$ kcal/mol. The event $e_i = \mathbb{1}[y_i \leq 0]$ is unchanged by the clip. The target-provenance audit fixes this convention because the clipped index reproduces all 3,283 released test-benchmark targets exactly and matches the released benchmark/evaluation convention, while the retained recalibrated full table has non-unique final sequence-level targets for 14,149 indexed entities. Appendix C gives the full comparison. The recalibrated table is retained as a sensitivity and provenance source.

A.3 Pool, representation, and fixed learner

One pool of 30,000 canonical training sequences is sampled with seed 20260922. It contains 3,865 low-stability members (12.883%); this fixed denominator is used in every recall calculation. For each algorithmic seed $s \in \{0, \ldots, 4\}$, all policies begin from the same 256 revealed pool labels. The initial identifiers are sampled without replacement from the lexicographically sorted pool with NumPy generator seed $100000 + s$.

Sequences are represented by the final hidden state of frozen `facebook/esm2_t36_3B_UR50D`, snapshot `476b639933c8baad5ad09a60ac1a87f987b656fc`, mean-pooled over non-special amino-acid tokens. The embedding dimension is 2,560 and the replay uses the unadapted base checkpoint throughout. Stored embeddings use `float16`, inference uses `bfloat16`, and the retained run configuration records batch size 32 and chunk size 256. The learner is a three-member bootstrap Ridge ensemble with $\alpha = 1$, `lsqr` solver, tolerance 10^{-3}, and an intercept. Before each fit, the revealed examples are sorted by stable identifier. Each ensemble member draws a bootstrap sample of size equal to the current revealed set, with replacement. At a refit with cumulative budget b, member $m \in \{0, 1, 2\}$ uses bootstrap-generator seed

$$7000003 + 100003s + 17b + m. \tag{5}$$

Table 2: Complete frozen validation-grid results. Each cell is mean ± sample standard deviation over five algorithmic seeds. *Found* is the cumulative number of low-stability pool members; recall and enrichment use the fixed pool denominator and prevalence. Validation Spearman is reported as a diagnostic alongside the acquisition objective.

Policy	Budget	Recall (%)	Found	Enrichment	Validation ρ
PRIB	512	3.35 ± 0.37	129.6 ± 14.2	1.965 ± 0.215	0.569 ± 0.064
PRIB	1,024	8.10 ± 0.42	313.2 ± 16.1	2.374 ± 0.122	0.597 ± 0.041
PRIB	2,048	16.99 ± 0.62	656.8 ± 24.1	2.489 ± 0.091	0.609 ± 0.036
PRIB	4,096	31.22 ± 0.67	1206.8 ± 26.0	2.287 ± 0.049	0.635 ± 0.025
PRIB	8,192	53.95 ± 0.77	2085.0 ± 29.7	1.976 ± 0.028	0.640 ± 0.030
Random	512	1.58 ± 0.20	61.0 ± 7.6	0.925 ± 0.115	0.549 ± 0.082
Random	1,024	3.20 ± 0.28	123.8 ± 10.9	0.938 ± 0.083	0.576 ± 0.128
Random	2,048	6.50 ± 0.53	251.4 ± 20.3	0.953 ± 0.077	0.608 ± 0.061
Random	4,096	13.64 ± 0.53	527.2 ± 20.4	0.999 ± 0.039	0.632 ± 0.040
Random	8,192	27.62 ± 0.81	1067.4 ± 31.5	1.011 ± 0.030	0.661 ± 0.041
Uncertainty	512	1.22 ± 0.22	47.0 ± 8.4	0.713 ± 0.128	0.568 ± 0.040
Uncertainty	1,024	2.78 ± 0.58	107.6 ± 22.4	0.816 ± 0.170	0.586 ± 0.046
Uncertainty	2,048	6.31 ± 0.73	244.0 ± 28.3	0.925 ± 0.107	0.622 ± 0.037
Uncertainty	4,096	13.50 ± 0.93	521.8 ± 35.9	0.989 ± 0.068	0.660 ± 0.031
Uncertainty	8,192	27.55 ± 1.01	1064.8 ± 38.9	1.009 ± 0.037	0.653 ± 0.015

The seed is policy-independent, so policies with the same revealed set receive the same bootstrap randomness.

A.4 Exact acquisition and persistence order

At each target budget $b \in \{512, 1024, 2048, 4096, 8192\}$, the batch size is the difference between b and the currently revealed count. Random uses a separate deterministic generator with seed $9000011 + 100003s + b$. Uncertainty chooses the largest $U_t(x)$. PRIB uses

$$R_t(s; x) = \mathrm{pctRank}_{\mathcal{U}_t}^{\mathrm{average}}(s(x)), \qquad A_{\mathrm{PRIB},t}(x) = R_t(U_t; x) + R_t(-\mu_t; x), \tag{6}$$

where `pctRank` is the ascending average-tie percentile rank implemented over the current unrevealed pool. The largest value receives rank one. Candidates are sorted by descending score; exact score ties are broken by ascending `q17_id`. Thus the score is deterministic conditional on the revealed set and bootstrap draws.

For every acquired candidate, the registry records `q17_id`, policy, seed, source and target budgets, origin, selected rank, pre-reveal predicted mean, pre-reveal uncertainty, acquisition score, revealed ΔG, and low-stability indicator. The registry is persisted before checkpoint metrics are evaluated; it is the recovery source of truth after an interrupted run. The oracle log records the identifier, split, purpose, policy, seed, and budget for every label access. These design choices make it possible to distinguish a selection-time quantity from its subsequently revealed label.

B Metrics, Complete Results, and Test Isolation

B.1 Aggregation conventions

For seed s and budget b, let $f_{s,b}$ be the number of low-stability members in the cumulative revealed set and let $F = 3865$ be the pool total. We report $f_{s,b}/F$ as recall and $(f_{s,b}/b)/(F/30000)$ as cumulative enrichment. Each table entry is the arithmetic mean across the five algorithmic seeds plus or minus the sample standard deviation

$$\mathrm{sd}_{\mathrm{seed}}(z) = \sqrt{\frac{1}{4}\sum_{s=0}^{4}\left(z_s - \frac{1}{5}\sum_{j=0}^{4} z_j\right)^2}. \tag{7}$$

This sample standard deviation is distinct from the population memberwise dispersion U_t in Section 3.2. It summarizes algorithmic-seed variation on the fixed pool; we use it descriptively throughout.

Table 3: Final held-out test diagnostics at the sole test checkpoint (8,192 labels). Entries are mean ± sample standard deviation over five algorithmic seeds.

Policy	Test Spearman ρ	Test RMSE
Random	0.582 ± 0.090	1.310 ± 0.103
Uncertainty	0.520 ± 0.043	1.375 ± 0.054
PRIB Failure-Aware	0.641 ± 0.010	1.259 ± 0.023

Table 4: Published-test target audit. Raw refers to the official-index `dG` before clipping; clipped is the primary oracle used in the replay.

Quantity	Value
Published-test coverage	3,283 / 3,283
Raw-index rows differing from benchmark	321
Raw values below -1 / above 5	266 / 55
Clipped-index MAE / RMSE / maximum error	0 / 0 / 0
Raw-index MAE / RMSE / maximum error	0.295 / 1.231 / 13.838
Indexed entities without a unique recalibrated target	14,149

B.2 Paired-grid and held-out diagnostics

The retained metric table contains 75 policy–seed–budget rows: three policies, five seeds, and five budgets. On all 25 paired seed-by-budget comparisons, PRIB has greater recall than Random; the same is true against Uncertainty. This is a complete descriptive enumeration of the frozen grid. At every adjacent doubling, the PRIB mean at the smaller budget exceeds the corresponding baseline mean at the next budget; the eight direct checks are all true.

Validation labels are accessed at each fixed checkpoint to compute the diagnostic in Table 2. Within production replay, predictive test access is restricted to the final 8,192-label checkpoint. The final test access log contains 49,245 rows, exactly 3 policies $\times 5$ seeds $\times 3283$ test examples, with zero earlier predictive test accesses. The separate provenance audit uses the released benchmark target before replay to resolve the published target convention. The held-out predictive quantities below therefore describe the final frozen model state and play no role in policy, weight, or checkpoint selection.

C Target-Provenance Audit

The target audit resolves a concrete ambiguity in the released data products. The retained full recalibrated table has non-unique final `deltaG` values for 14,149 indexed entities, so simple replicate aggregation leaves the primary sequence-level target ambiguous. Clipping the official index target to the published $[-1, 5]$ range exactly reproduces the released 3,283-example test benchmark. Table 4 records the audit quantities. Because the event threshold is zero, clipping leaves low-stability membership unchanged.

D Integrity and Claim Traceability

The final validation report declares a complete 75-row metric grid, matching artifact-manifest hashes, and a passed final-test access guard. The numerical-claim audit maps every rendered project-specific value to a retained source artifact and locator, providing end-to-end traceability from frozen configuration to rendered claims. The following line-wrapped values identify the frozen archive, configuration, and official index, respectively:

```
c860675e3b494b39c4463a72253c47e
6194a51cab43ab384ecbea1c5d8547408
6738871646ad0f9f430ca1064c7c4a
167fd08bd794e274f2a897778c7feb079e
60ba064fd4c25052f8230833508e7df
9ddb178b7e42080290773c86d441f2c48
```