Does the Prior Pay Off? Scaling a Protein Language Model Lets Reinforcement Learning Beat Directed Evolution on GB1

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Reinforcement learning (RL) fine-tuning of protein language models (PLMs) is increasingly proposed for sequence design, but reported gains are usually measured against weak baselines on learned, gameable oracles. We run a deliberately fair comparison on the GB1 four-site combinatorial fitness landscape: an *exact-lookup* oracle (no surrogate to exploit), a *matched unique-query budget*, a strong classical baseline (simulated annealing), and a novelty floor, averaged over five seeds. We first show the PLM prior is a poor fitness predictor: ESM-2 masked-marginal score correlates with GB1 fitness at Spearman 0.04 (35M) and 0.15 (150M), and its top-ranked variants have near-zero fitness, so a no-RL masked-marginal proposer (greedy-MLM) is weak. Simulated annealing is a strong baseline that nearly solves the landscape. Group Relative Policy Optimization (GRPO) collapses by default (spending only 19 of 1000 queries) and requires discovery-shaped reward, an entropy bonus, and a low learning rate to be viable. Even tuned, GRPO over a 35M prior only matches annealing. Scaling the prior to 150M, however, changes the outcome: GRPO then *beats* simulated annealing by about 10% in the sample-limited regime where oracle queries are scarce, with the advantage narrowing as the budget grows and both methods approach the global optimum. Strikingly, this gain is *decoupled* from the prior's fitness-ranking quality, which remains poor at 150M, suggesting the larger PLM helps as an *initialization / inductive bias* rather than as a fitness oracle. The practical message: scaling the protein language model lets reinforcement learning overtake a strong classical optimizer precisely where queries are expensive.

1 Introduction

Machine-guided protein design searches an astronomically large sequence space using a fitness oracle (a wet-lab assay or simulator) that is expensive to query. The dominant recent narrative is that fine-tuning a pretrained protein language model (PLM) with reinforcement learning (RL) yields better proposals than classical search [1, 2]. Yet most such claims are benchmarked against weak baselines, on *learned* oracles that an optimizer can exploit, and without controlling the number of oracle queries, precisely the axis that matters in practice [3, 4].

We ask a sharp, controlled question: *at a matched unique-query budget, on an exact (ungameable) oracle, does GRPO [5] fine-tuning of ESM-2 [6] propose higher-fitness variants than cheap classical search?* We use the GB1 four-site combinatorial landscape [7, 8]: every one of the 160,000 variants

*Corresponding author: deep@lab.cloud

has (or lacks) a measured fitness, so the oracle is an exact lookup and the entire landscape is enumerable for evaluation. We enforce a matched budget (one query = one unique measured variant), a strong baseline (simulated annealing), a novelty floor, and five seeds.

Contributions.

1. A *prior-informativeness probe* showing ESM-2's masked-marginal score is nearly uninformative about GB1 fitness (ρ=0.04 at 35M, 0.15 at 150M), so "the prior already knows the answer" is false here (§5).
2. A fair matched-budget benchmark in which, at 35M, tuned GRPO only *ties* simulated annealing and never meets the pre-registered $\geq 10\%$ bar (§5).
3. A scale finding: a 150M prior makes GRPO *beat* directed evolution by $\sim$10% in the sample-limited regime, decoupled from the prior's (still poor) fitness-ranking, consistent with the PLM helping as an initialization / inductive bias rather than as a fitness oracle (a correlational link we do not isolate; §5, §6).
4. A reproducible account of the engineering required to make GRPO viable on a tiny epistatic landscape (anti-collapse reward shaping, entropy, learning rate) (§3).

2 Related Work

PLMs as fitness predictors. ESM-1v [9] and ESM-2 [6] score mutations by masked-marginal log-odds; our greedy-MLM baseline is exactly this no-RL use of the prior. The FLIP benchmark [8] already reports that pretrained embeddings "do not outperform simpler models on mutational landscapes (GB1, AAV)"; we quantify this as a calibration probe. **Model-based optimization for sequences.** CbAS [10], GFlowNets [11], AdaLead/FLEXS [4], and PEX [12] formalize budgeted, oracle-metered design; notably Sinai et al. [4] find a trivial greedy search out-competes the RL method DyNA-PPO [1], and Jain et al. [11] argue RL maximizers mode-collapse. These motivate our strong-baseline, novelty-aware protocol. **RL fine-tuning of PLMs.** GRPO [5] is our optimizer; theory shows it amplifies modes the base model already samples and cannot reach zero-prior modes [13]. Closest to us, Cao et al. [2] ask whether RL on PLMs explores beyond the prior and find it largely contracts the explored space, but they compare RL only to the base model's sampling distribution, never to a no-RL proposer at a matched unique-query budget on an exact oracle. Concurrent work points the same way: Kmicikiewicz et al. [14] report that frozen-prior search can match or beat RL fine-tuning, and Wang et al. [15] find a frozen language-model proposer competitive with evolutionary search (only tying it on GB1), both consistent with our 35M result. We add the exact-oracle, matched-budget, and scale axes.

3 Method

Task and oracle. A method is a search procedure that, given a budgeted oracle, proposes length-4 variants over GB1 positions $\{39, 40, 41, 54\}$. The oracle returns the measured fitness (wild-type = 1.0, non-binding ≈ 0, max 8.76); one *unique measured* variant costs one budget unit, repeats and unmeasured cells are free. The evaluator (fixed, never optimized) computes all metrics on the method's queried set.

Baselines. Random mutagenesis; directed evolution (greedy hill-climb and Metropolis / simulated annealing); greedy-MLM (ESM-2 masked-marginal additive scoring, ranking all combos and querying the top measured ones, the no-RL prior control); ESM-2 used only as a prior, never trained on GB1.

GRPO policy. The policy is a small autoregressive network over the four sites (each residue conditioned on previous choices, so it can represent epistasis), with per-position logits biased by the frozen ESM-2 masked-marginal prior; that is, the policy is *initialized at the prior* and KL-anchored to a frozen reference. For a group of G sampled variants with rewards r_i, GRPO uses the group-relative advantage $\hat{A}_i = (r_i - \text{mean}(\mathbf{r}))/(\text{std}(\mathbf{r}) + \epsilon)$ and a Schulman KL penalty to the reference, with no value network. Three choices prove necessary to avoid collapse (§5): a *discovery-shaped* reward (an already-queried or unmeasured proposal yields reward 0, so the policy is not rewarded for re-proposing known peaks), an entropy bonus, and a low learning rate, which the sweep identifies as the single most important hyperparameter.

4 Experimental Setup

Data. GB1 from FLIP [8] (Wu et al. [7]): 149,361 measured of 160,000 variants. A fixed pool of 96 low-fitness, leakage-checked seeds (no peak or near-peak) is the shared cold start. **Protocol.** Budgets $\{100, 1\text{k}, 1.28\text{k}, 5\text{k}, 20\text{k}\}$ unique queries (plus 2k/3k for the win-region map); 5 seeds for the headline; primary metric is the mean ground-truth fitness of the top-100 unique variants found (*top100-mean*, higher is better). Error bars and the "±" notation throughout denote one standard deviation over the five seeds. All runs execute on a single NVIDIA A10; the matched budget equalizes *oracle queries*, not compute (see §6). The pre-registered success criterion was a $\geq 10\%$ relative top100-mean win over the strongest non-RL baseline at a 5k budget *with the 35M backbone*, stable over ≥ 3 seeds, with a median-Hamming-≥ 2 novelty floor. We call a budget a *tie* when the relative gap is below this 10% margin and a Welch two-sample t-test over the five seeds is non-significant ($p > 0.05$).

5 Results

The prior is uninformative about fitness. ESM-2 masked-marginal score correlates with GB1 fitness at Spearman ρ=0.04 (35M) and 0.15 (150M), both negligible despite $N \approx 149$k measured variants, so neither is practically informative. The top-100 model-ranked variants average fitness 0.12 (35M) and 0.02 (150M), and the global peak ranks below the median. Note the non-monotonicity: rank correlation rises slightly with scale while the absolute fitness of the top-ranked picks does *not*, so the larger prior is not a usefully better fitness predictor. Consequently greedy-MLM is weak at both scales (Table 1, Fig. 1a).

Directed evolution is strong; 35M GRPO ties it. Simulated annealing nearly solves GB1 (success rate 1.0 by 1k queries; the global peak by 5k). A sweep produced two GRPO configurations; we report the sweep-selected config (B: entropy 0.2, learning rate 0.02), and note the higher-learning-rate config (A) underperforms (e.g. 4.70 vs. 5.51 at 5k). The tuned 35M policy matches annealing for budgets $\geq$ 1k within five-seed noise but loses at 100 and *never* beats it by $\geq 10\%$ (Table 1). The pre-registered primary criterion (a $\geq 10\%$ win at 5k with the 35M backbone) is therefore **not met**: a null we accept.

Scaling the prior makes RL win in the sample-limited regime. With a 150M prior, GRPO *beats* annealing across the query-scarce regime (by $+10.1\%/+6.8\%/+4.7\%/+2.1\%$ at

Table 1: Top-100 mean GB1 fitness (mean ± 1 s.d. over 5 seeds) at matched unique-query budgets. Bold marks the best method at each budget. gMLM (greedy masked-marginal, no RL) is deterministic, so no s.d. is shown. The last column is the GRPO@150M-vs-annealing relative gap; * denotes a significant difference (Welch t-test over 5 seeds, $p < 0.05$); "tie" follows the definition in §4 ($|\text{gap}| < 10\%$ and $p > 0.05$). Annealing is the strongest classical baseline: it dominates random mutagenesis and greedy hill-climbing at every budget (not shown), so we contrast against it alone. greedy-MLM@35M (the 35M prior control) is weaker still than greedy-MLM@150M shown here (e.g. 1.78 vs. 1.18 at 1k).

Budget	gMLM@150M	anneal	GRPO@35M	GRPO@150M	RL150 vs anneal
100	0.02	**0.93** ±0.44	0.22 ±0.14	0.36 ±0.28	loses
1,000	1.18	4.34 ±0.31	4.49 ±0.14	**4.77** ±0.07	+10.1%*
1,280	1.46	4.61 ±0.26	4.74 ±0.11	**4.92** ±0.13	+6.8%
5,000	3.80	**5.53** ±0.06	5.51 ±0.04	5.50 ±0.03	tie
20,000	4.93	**5.72** ±0.01	5.69 ±0.03	5.67 ±0.03	tie

1k/1.28k/2k/3k queries), with the largest margin where queries are scarcest, the gap narrowing as both methods approach the global optimum and converging to a tie at saturation (5k–20k) (Fig. 1b). The 1k margin is statistically significant (Welch $t(4.4)$=3.07, p=0.033, Cohen's d=1.9, 95% CI of the gap $[0.06, 0.82]$); significance softens as the budgets grow and the methods converge (marginal at 1.28k–2k, $p \approx 0.05$–0.07; not significant by 3k). We therefore read the *trend* (a clear, monotone RL advantage that peaks where samples are scarce) as the result, more than any single cell. Because greedy-MLM@150M stays weak, this advantage is decoupled from the prior's (still poor) fitness-ranking quality: the larger prior helps RL search, not by ranking fitness better (it does not), but as a starting point (§6). As this lives off the pre-registered operating point, we treat it as an exploratory finding.

GRPO requires shaping to avoid collapse. With an unshaped reward the policy collapses, spending only 19 of 1000 queries before stalling (it is rewarded for re-proposing found peaks). The final config jointly uses a discovery-shaped reward, an entropy bonus, and a low learning rate; the sweep identifies learning rate (0.02 vs. a collapsing 0.1) as the dominant hyperparameter. We did not run a full one-factor-at-a-time ablation, so we report these as the jointly-used choices rather than independently certified necessary conditions.

6 Discussion

The result is two-sided. On the cautionary side: on a fair benchmark the PLM prior carries essentially no fitness signal, and a carefully tuned RL policy only ties classical directed evolution at small scale, echoing Sinai et al. [4] and Cao et al. [2]. On the constructive side: scaling the prior buys a seed-consistent advantage in the sample-limited regime, and it does so without the prior becoming a usefully better fitness predictor: its top-ranked variants remain near-zero fitness even at 150M. Because the gain appears while ranking quality stays poor, the most parsimonious reading is that the larger prior helps as an initialization / inductive bias rather than as a fitness oracle. We stress this is a correlational decoupling across a single 35M→150M contrast: we do not isolate the channel, and alternatives we cannot exclude include a different initial-policy entropy or a smoother logit landscape that aids exploration, or top-region ranking structure that a global Spearman over

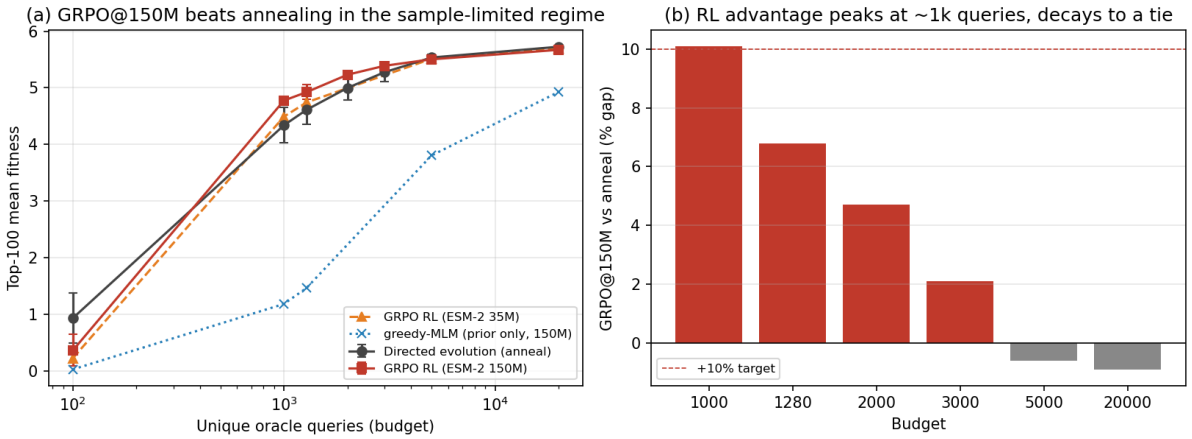

Figure 1: (a) Top-100 mean GB1 fitness (the primary metric; y-axis) versus unique-query budget (x-axis, log scale). Error bars are ± 1 standard deviation over five seeds. The prior-only proposer (greedy-MLM, deterministic) is weak; GRPO over a 150M prior overtakes simulated annealing in the sample-limited regime and converges to it at saturation. (b) Relative advantage (%) of GRPO@150M over annealing versus budget; the gap peaks near 1k queries (+10.1%, significant at p=0.03) and decays monotonically to a tie by 5k. The dashed line marks the pre-registered +10% target.

all 160k variants washes out. Isolating the mechanism (e.g. ablating or shuffling the 150M prior logits under an otherwise-identical policy) is left to future work. On this landscape, scale appears to help by providing a better starting policy rather than better fitness knowledge.

7 Limitations

Construct (novelty). The median-Hamming-≥ 2 floor is met by *no* method, including baselines, because GB1's design space has diameter 4 and all good methods converge on the same clustered optima, consistent with the floor being mis-calibrated for a diameter-4 landscape rather than with memorization. We flag that this is a *post-hoc* relaxation of a pre-registered gate: because the floor is unmet, the 150M advantage is defended geometrically (diameter 4) and not by a passed anti-memorization test. A full-length landscape (e.g. GFP) is needed for a meaningful novelty axis. **Compute (fairness).** The matched budget equalizes *oracle queries*, not computation: GRPO@150M performs group sampling and policy forward/backward passes on a 150M backbone per step, whereas annealing performs a single lookup plus a Metropolis test. On GB1's exact-*lookup* oracle a query is cheap, so the query-efficiency advantage in the sample-limited regime is *not* a wall-clock or FLOP advantage; it is relevant only where one oracle query (e.g. a wet-lab assay) far outweighs a GRPO step. We do not claim a practical-cost win on this benchmark. **External.** A single, small, fully-enumerable landscape, and a two-point scaling contrast (35M $\rightarrow$ 150M): two points do not establish monotone scaling. The effect should be confirmed on a second landscape and across more model sizes before it is treated as confirmatory rather than exploratory. **Internal.** The 150M win lives off the pre-registered operating point (5k, 35M backbone); we label it exploratory to avoid a garden-of-forking-paths reading, and the reported 35M "tie" uses the sweep-selected config (B). The RL policy (an autoregressive head biased by the prior) and the greedy-MLM control

(additive masked-marginal scoring) use the prior through *different* functional forms, so part of the RL-vs-prior gap may reflect the autoregressive head's capacity to represent epistasis rather than RL per se; an ESM-2-as-policy variant and a no-RL policy-at-init proposer are untested.

8 Availability

Artifacts (code, configurations, the exact GB1 oracle hash, per-seed result logs, and the project record) are available from the authors on request. The GB1 data is public via FLIP [8].

9 Conclusion and Future Work

On a deliberately fair benchmark, scaling the protein language model lets RL fine-tuning *overtake* a strong classical optimizer for protein fitness design: with a 150M prior, GRPO beats simulated annealing by about 10% in the query-scarce regime where it matters most. The effect is the more striking because the prior remains an uninformative fitness predictor, so the benefit appears to come from initialization rather than from the model knowing which proteins are good. At small scale (35M) RL only matches annealing, and the advantage narrows at large query budgets where both methods saturate. Next steps: confirm the scale effect on a second landscape (GFP) with a held-out surrogate oracle and a meaningful novelty axis; test larger priors and an ESM-2-as-policy variant; and characterize the budget at which the RL advantage appears as a function of model scale.

References

[1] 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 (ICLR)*, 2020.

[2] Hanqun Cao, Hongrui Zhang, Junde Xu, Zhou Zhang, Lingdong Shen, Minghao Sun, Ge Liu, Jinbo Xu, Wu-Jun Li, Jinren Ni, Cesar de la Fuente-Nunez, Tianfan Fu, Yejin Choi, Pheng-Ann Heng, and Fang Wu. From Supervision to Exploration: What Does a Protein Language Model Learn During Reinforcement Learning? 2025. URL https://arxiv.org/abs/2510.01571.

[3] Shikha Surana, Nathan Grinsztajn, Timothy Atkinson, Paul Duckworth, and Thomas D. Barrett. Overconfident Oracles: Limitations of In Silico Sequence Design Benchmarking. 2025. URL https://arxiv.org/abs/2502.17246.

[4] 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. 2020. URL https://arxiv.org/abs/2010.02141.

[5] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2024. URL https://arxiv.org/abs/2402.03300.

[6] 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.

[7] 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.

[8] 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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[9] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[10] David H. Brookes, Hahnbeom Park, and Jennifer Listgarten. Conditioning by adaptive sampling for robust design. In *International Conference on Machine Learning (ICML)*, 2019.

[11] Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological Sequence Design with GFlowNets. In *International Conference on Machine Learning (ICML)*, 2022.

[12] 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 (ICML)*, 2022.

[13] Youssef Mroueh. Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification. 2025. URL https://arxiv.org/abs/2503.06639.

[14] Michal Kmicikiewicz, Vincent Fortuin, and Ewa Szczurek. ProSpero: Active Learning for Robust Protein Design Beyond Wild-Type Neighborhoods. 2025. URL https://arxiv.org/abs/2505.22494.

[15] Yinkai Wang, Jiaxing He, Yuanqi Du, Xiaohui Chen, Jianan Canal Li, Li-Ping Liu, Xiaolin Xu, and Soha Hassoun. Large Language Model is Secretly a Protein Sequence Optimizer. 2025. URL https://arxiv.org/abs/2501.09274.