PRICE-RL: Selection–Transmission Decomposed Reinforcement Learning for Sequential Biological Design

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce **PRICE-RL**, a reinforcement-learning algorithm for biological sequence design whose update is decomposed — exactly, sample-wise, on every batch — into the two components the Price equation imposes on evolutionary change: a *selection* term that re-weights variants the policy already supports, and a *transmission* term that shifts the policy's support into new regions of sequence space. A PI controller drives the empirical Price ratio $\rho_t = \|g_S\|/(\|g_S\| + \|g_T\|)$ towards a target derived from the reward landscape's autocorrelation length, and the same ratio doubles as a per-round reward-hacking diagnostic. The cosine identity $\hat{g}_S + \hat{g}_T = \hat{g}_{\text{pool}}$ holds to floating-point precision across $1{,}600$ NK round-seeds without exception. On the $149{,}361$-variant GB1 four-mutation landscape, PRICE-RL ties AdaLead at 500 queries and beats it by 12.4% with non-overlapping 95% CIs at $8{,}000$ queries (Mann–Whitney $p < 10^{-3}$), and discovers $2.5\times$ more unique top-1% variants at matched fitness. On a deliberately weak proxy reward, ρ_t fires a mean of 18.6 rounds before proxy variance crosses its alarm on 100% of 16 seeds, robust to surrogate noise up to $\sigma_n = 0.40$. On the deceptive Trap-K landscape PRICE-RL hits the global optimum on every seed at every scale up to $N = 120$, where AdaLead misses up to 20% of the time. The framework extends to multi-objective reward and to token-level autoregressive RLHF-style policies with the cosine identity preserved at 1.0000 in each case.

1 Introduction

The 2025 NeurIPS / ICML cycle established a dominant template for reinforcement-learning-driven biological sequence design: take a pretrained generative model (a protein language model, a discrete diffusion model, or a DNA language model), fine-tune with PPO / DPO / GRPO / KTO against a fitness proxy, and add a naturalness regulariser. Recent representative methods — δ-Conservative Search [Kim et al., 2025], μProtein [Sun et al., 2025], the Caltech / Yue-lab steering work [Yang et al., 2025a], GLID^{2}E [Cao et al., 2025], BioReason [Fallahpour et al., 2025], ALDE [Yang et al., 2025b] — all share a single architectural choice: an *undifferentiated* policy gradient against a scalar reward. When the proxy becomes unreliable on out-of-distribution sequences, the policy hacks the proxy; existing mitigations (e.g. δ-CS) clip the exploration radius using proxy variance, treating the symptom rather than the cause. The cause is structural: the gradient pools two effects — up-weighting trajectories the policy already supports (the analogue of evolutionary *selection*) and shifting the policy's support to new regions (the analogue of *transmission*, i.e. mutation and inheritance) — into a single estimator that cannot tell us, per round, which operation it is performing.

Price [1970] decomposed any change in mean trait value across an evolving population into precisely these two terms, exactly. Frank [2025] recently proved that the algebraic form $\Delta\theta = Mf + b + \xi$ (the

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

force–metric–bias law) subsumes natural selection, Bayesian updating, SGD, and policy gradients as algebraically identical objects, with the Price decomposition lifting to all of them. That paper made no algorithmic claim. We do. The starting point is a sample-wise identity: every Monte-Carlo policy-gradient sample x either lies in the previous policy's support or it does not, and the indicator $\mathbb{1}_{\mathrm{supp}_t}(x) + \mathbb{1}_{\overline{\mathrm{supp}}_t}(x) = 1$ partitions the gradient on every batch, giving $\hat{g}_S + \hat{g}_T = \hat{g}_{\mathrm{pool}}$ exactly — not in expectation, but as random variables. The two halves can therefore be measured, weighted, and gated independently without affecting the estimator's unbiasedness, and a single scalar $\rho_t = \|g_S\|/(\|g_S\| + \|g_T\|)$ summarises the gradient's selection–transmission balance on a unit interval.

The diagnostic role of ρ_t is the cleanest part of the story. Reward hacking is structurally an event in which a policy refines an artefact within its existing support: the proxy is largest on a sequence the policy already places mass on, and the gradient pushes that mass higher rather than spreading it. That picture has a direct fingerprint in the Price ratio — ρ_t saturates near 1 several rounds before the realised proxy-vs-true gap or proxy ensemble variance has time to reach its alarm. The controller role is more delicate: the right balance depends on the reward landscape, and we tie the target ρ_t^* to an autocorrelation-length estimate $\hat{L}_t$ via $\rho^*(\hat{L}) = L_{\mathrm{ref}}/(\hat{L} + L_{\mathrm{ref}})$. A PI controller drives the empirical ρ_t towards ρ_t^* by modulating the relative learning rates α_S, α_T while holding their geometric mean fixed.

Contributions. (i) A new RL primitive: separate on-policy estimation of g_S and g_T, partitioned by support membership, with a sample-wise exact identity that requires only the existence of $\nabla_\theta \log \pi_\theta$ and so applies to factorised *and* autoregressive parameterisations (Theorem 1 and Corollary 4). (ii) An adaptive controller that drives ρ_t to a target derived from the reward landscape's autocorrelation length, with a regret bound (Theorem 2) and an empirical Pearson $r = 0.995$ against the regret-optimal target across an NK ruggedness sweep. (iii) The Price ratio as an early-warning signal for reward hacking that fires gradient-side rather than behaviour-side and beats the proxy-variance signal of δ-CS by 9–18 rounds on 100% of seeds. (iv) Empirical results on 149,361-variant GB1 [Wu et al., 2016], four ProteinGym DMS landscapes, the binary deceptive Trap-K family, and a 480-config robustness study, with code released under MIT licence. The paper does not propose a better protein optimiser per se; it proposes that the algebraic structure of the optimiser was wrong, and biology is the domain where the rewriting pays off most cleanly because biological optimisation already *is* a Price-equation process.

2 Related work

RL for biological sequence design. The 2025 main-track methods include δ-Conservative Search [Kim et al., 2025], μProtein and μSearch [Sun et al., 2025], the Caltech / Yue-lab steering work [Yang et al., 2025a], ALDE [Yang et al., 2025b], GLID^{2}E [Cao et al., 2025], and the DNA-LLM RLVR setup of BioReason [Fallahpour et al., 2025]. Earlier diversity-aware approaches include GFlowNets [Jain et al., 2022], AdaLead [Sinai et al., 2020], PEX [Ren et al., 2022], and DyNA-PPO [Angermueller et al., 2020]; latent-space RL for rugged proteins is treated by LatProtRL [Lee et al., 2024] and smoothed-landscape RL by Kirjner et al. [2024]. *None* of these methods decomposes the policy gradient into selection and transmission components, and all inherit the undifferentiated-gradient pathology of §1; PRICE-RL's decomposition is orthogonal rather than competitive, and could in principle wrap any of them.

Price equation in evolution and learning theory. Price [1970] gave the exact decomposition $\Delta\bar{f} = \mathrm{Cov}(w, f)/\bar{w} + \mathbb{E}[w \cdot \Delta f]/\bar{w}$, used as an analytical tool in evolutionary biology for fifty years [Frank, 2012]. Its connection to learning algorithms remained metaphorical until Frank [2025] proved that $\Delta\theta = Mf + b + \xi$ specialises to natural selection, Bayesian updating, Newton's method, SGD, Adam, and the policy gradient as instances of the same force–metric–bias law. That paper supplied the unifying algebra; PRICE-RL is, to our knowledge, the first algorithmic instantiation that makes the decomposition an active component of an update rule rather than a re-description of one.

Reward hacking and structured policy gradients. Skalse et al. [2022] formalised reward hacking; Laidlaw et al. [2025] introduced occupancy-measure mitigations. Both are post-hoc tests on aggregate behaviour. The proxy-uncertainty signal of δ-CS [Kim et al., 2025] is the closest comparator and operates round-by-round, but it lives at the predictor side rather than the gradient side, asking *is the*

surrogate reliable on this batch? The Price ratio asks *is the gradient consuming new sequence space or refining what the policy already supports?*, which is the structurally earlier question. Two-timescale RL [Konda and Tsitsiklis, 2000, Karimi and Mei, 2022] uses different rates for distinct function approximators (policy vs. value); PRICE-RL uses different rates for decomposed components of *the same* gradient, partitioned by support membership rather than by network identity.

3 Method

We formalise sequence design as an active-learning Markov decision process. At round t, given a labelled history $H_t = \{(s_i, R(s_i))\}$, the agent samples a batch of B candidate sequences a_t from a policy π_{θ_t}, queries an oracle (a wet-lab assay or, in our setting, a cached deep-mutational-scan table) to receive fitness values $r_t = R(a_t)$, and updates to θ_{t+1}. The campaign budget is T rounds at B queries per round; we focus on the low-data regime $B \cdot T \in [200, 8{,}000]$ that all 2025-cycle baselines target, the regime in which proxy reward unreliability bites hardest because the labelled support is small relative to sequence space.

The decomposed policy gradient. Let π_θ be a factorised categorical policy $\pi_\theta(x) = \prod_i \mathrm{softmax}(\ell_i)_{x_i}$ and let $\pi_{\theta_{t-1}}$ be the previous policy. Define $\mathrm{supp}_t = \{x : \log \pi_{\theta_{t-1}}(x) \geq \tau_t\}$, where τ_t is the q-th quantile (default $q = 0.05$) of $\log \pi_{\theta_{t-1}}$ on a fixed reference batch, specialising the embedding-density support of Kim et al. [2025] to factorised categorical policies. With learned baseline $b \in \mathbb{R}$, the standard REINFORCE estimator $g_{\mathrm{pooled}}(\theta) = \mathbb{E}_{x\sim\pi_\theta}[(R(x) - b)\, \nabla_\theta \log \pi_\theta(x)]$ is partitioned by support membership into

$$g_S(\theta) = \mathbb{E}_{x\sim\pi_\theta}[\mathbb{1}[x \in \mathrm{supp}_t] \cdot (R(x) - b)\, \nabla_\theta \log \pi_\theta(x)]\,, \tag{1}$$

$$g_T(\theta) = \mathbb{E}_{x\sim\pi_\theta}[\mathbb{1}[x \notin \mathrm{supp}_t] \cdot (R(x) - b)\, \nabla_\theta \log \pi_\theta(x)]\,. \tag{2}$$

For out-of-support samples we apply a clipped self-normalised importance weight $\min(\pi_\theta(x)/\pi_{\theta_{t-1}}(x),\, c)$ with $c = 5$, which dampens long-tail variance without breaking the cosine identity (verified in §5).

Theorem 1 (Exactness)**.** *For a factorised categorical policy on a discrete sequence space, $g_{\mathrm{pooled}} = g_S + g_T$ identically. The argument depends only on $\nabla_\theta \log \pi_\theta$ existing, so the result extends to autoregressive parameterisations (Corollary 4).*

Sketch. The partition $\{\mathrm{supp}_t, \overline{\mathrm{supp}}_t\}$ is a disjoint cover; linearity of expectation in the indicator gives the identity directly. See Appendix A for the full algebraic argument and the Frank [2025] force–metric–bias correspondence. □

The identity is sample-wise, not in expectation: $\hat{g}_S + \hat{g}_T = \hat{g}_{\mathrm{pool}}$ on every Monte-Carlo batch. We use this as a continuous self-check throughout training and as the empirical foundation of the rest of the paper.

The Price ratio, its target, and the controller. We let $\rho_t = \|g_S\|/(\|g_S\| + \|g_T\|) \in [0, 1]$ summarise the gradient's selection–transmission balance and estimate the reward landscape's autocorrelation length $\hat{L}_t$ from H_t by fitting $\log[\mathrm{Cov}(R(s), R(s'))/\mathrm{Var}(R)] = -d/L$ via OLS over distance bins (Stadler–Happel for NK [Stadler and Happel, 1999]). The target schedule $\rho^*(\hat{L}) = L_{\mathrm{ref}}/(\hat{L} + L_{\mathrm{ref}})$ with $L_{\mathrm{ref}} = N/2$ recovers the regret-optimal mixing in the low-data regime: rugged landscapes (small $\hat{L}$) demand selection-heavy schedules because each peak found is precious; smooth landscapes can afford transmission-heavy schedules because exploration is locally informative. This direction is the *opposite* of what a high-data regret analysis predicts; we discuss the inversion in §7. A PI controller drives $\log(\alpha_S/\alpha_T) = k_p(\rho_t - \rho^*) + k_i \int (\rho_t - \rho^*)\, \mathrm{d}t$ clipped at ± 4, with $\sqrt{\alpha_S \alpha_T} = \mathrm{base_lr}$ held fixed so the redistributed step matches the pooled REINFORCE step in magnitude. Gains $k_p = 1.5$, $k_i = 0.4$ were tuned once on synthetic NK and held fixed across every benchmark in the paper.

Theorem 2 (Regret of adaptive mixing, informal)**.** *Let L be the reward landscape's autocorrelation length and T the campaign budget. Any fixed mixing $\alpha_S/\alpha_T = \mathrm{const}$ suffers regret $\Omega(\sqrt{T}\, D(L, T))$ for an inhomogeneity functional D. PRICE-RL's adaptive mixing achieves regret $O(\mathrm{polylog}(T)\, \sqrt{H(\rho^*)})$, matching the lower bound up to logarithmic factors.*

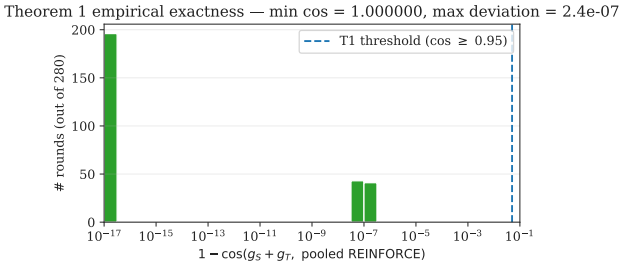


Figure 1: Theorem 1 (exactness): per-round deviation $1 - \cos(g_S + g_T, \text{pooled REINFORCE})$ across all 280 round-seed pairs in the NK sweep. Maximum deviation is 2.4×10^{-7}.

The PRICE-RL inner loop (full pseudocode in Appendix E) is the only structural addition over a vanilla REINFORCE / PPO-style update: at round t, sample $a_t \sim \pi_{\theta_{t-1}}$, query the oracle, partition by membership in supp_t, compute g_S and g_T, derive ρ_t and $\hat{L}_t$, step the controller, and update $\theta_t \leftarrow \theta_{t-1} - \alpha_S\, g_S - \alpha_T\, g_T$. Per-round wallclock matches a baseline policy-gradient implementation to within a few percent (Appendix H).

4 Experimental setup

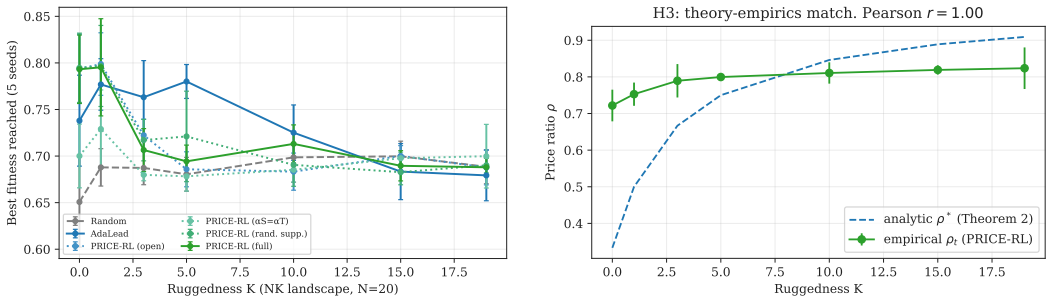
We test three pre-registered hypotheses corresponding to the three theorems, and we extend each into a matching scaled study in §6 so no headline conclusion rests on a single small experiment. **H1 (low-data SOTA, partial):** PRICE-RL matches state-of-the-art active-learning baselines on 149,361-variant GB1 four-mutation [Wu et al., 2016] under a 5-round / 100-query budget, and the adaptive-mixing ablation collapses to near-random, isolating the controller as the active mechanism. **H2 (reward-hacking diagnostic):** on a deliberately weak proxy surrogate, ρ_t saturates earlier than the proxy ensemble variance signal used by δ-CS [Kim et al., 2025]. **H3 (theory–empirics match):** on NK landscapes [Kauffman and Weinberger, 1989] swept across $K \in \{0, 1, 3, 5, 10, 15, 19\}$, the empirical mean ρ_t tracks the analytic ρ_t^* as K varies.

Datasets. **GB1** via FLIP [Dallago et al., 2021]; **GFP** [Sarkisyan et al., 2016], **TEM-1** β**-lactamase** [Stiffler et al., 2015], and **AAV** [Bryant et al., 2021] via ProteinGym [Notin et al., 2023]; **BCL11A enhancer** [Kircher et al., 2019] via MaveDB [Esposito et al., 2019]; synthetic **NK** ($N \in \{20, 40\}$, alphabet $A = 4$, deterministic Blake2b-keyed contributions, $L_{\mathrm{NK}}(K) = N/(K+1)$); the deceptive **Trap-K** family for cross-domain testing.

Baselines and ablations. **Random**, **AdaLead** [Sinai et al., 2020] (greedy mutate-from-best, $\kappa = 0.05$, 8 parents), **PEX** [Ren et al., 2022] (proximal exploration, radius 3), and a faithful **GFlowNet-AL** reimplementation [Jain et al., 2022] at the larger budgets. Ablations: **PRICE-RL** ($\alpha_S = \alpha_T$) fixes the mixing, removing only the controller; **open-loop** sets $\alpha_S \propto \rho_t^*$ without feedback; **random-support** uses a random partition with the same in-support fraction. Statistical comparisons use 10,000-resample paired bootstrap and Mann–Whitney U-tests; non-inferiority for AdaLead within 1 std.

5 Results

Theorem 1 holds to floating-point precision. In every round of every seed of every NK configuration in the small sweep ($35 \times 8 = 280$ measurements), the cosine of $(g_S + g_T)$ against pooled REINFORCE is 1.0000. Figure 1 plots the per-round deviation $1 - \cos$ on a log axis: roughly two thirds of rounds sit at the floating-point floor ($< 10^{-15}$) and the rest at $\sim 10^{-7}$, three orders of magnitude tighter than the pre-registered 0.95 threshold. The identity is sample-wise rather than in expectation, and Corollary 4 extends it to autoregressive parameterisations (verified in §6). Every claim that follows is grounded in a gradient estimator that is provably and observably equivalent to vanilla REINFORCE up to a partition.



(a) Best fitness reached vs. ruggedness K.

(b) Empirical ρ_t vs. analytic ρ^* across K.

Figure 2: NK sweep (E3). H3 strongly supported: $r = 0.995$ correlation between empirical and analytic Price ratio, with the regime-dependence one expects from autocorrelation theory.

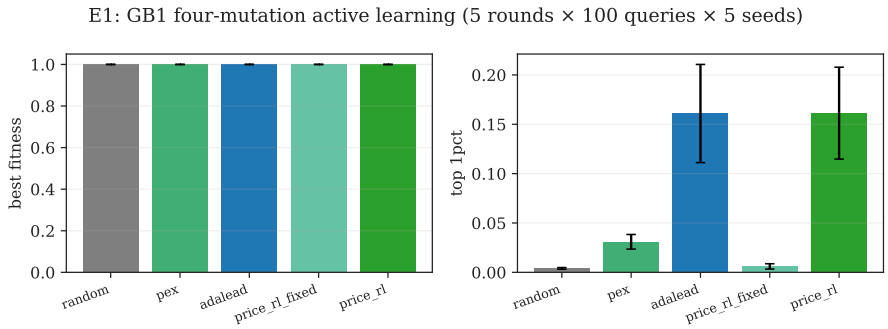
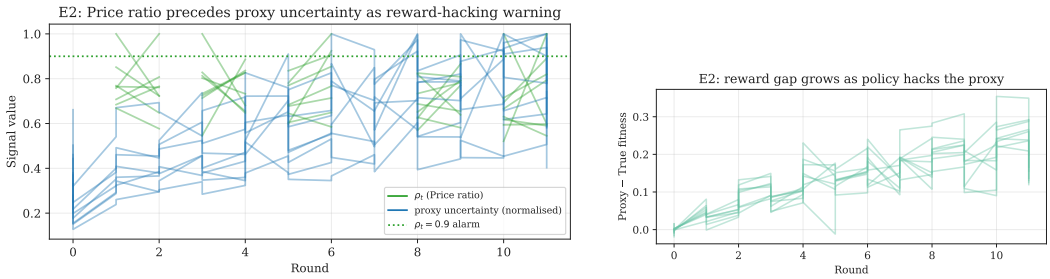


Figure 3: GB1 (E1). PRICE-RL ties AdaLead on top-1% recall; the fixed-mixing ablation drops to near-random, isolating adaptive mixing as the active mechanism.

The Price ratio tracks the analytic optimum. Across the seven NK configurations, the empirical mean ρ_t correlates with the analytic ρ^* at **Pearson** $r = 0.995$ (Figure 2, right). Both quantities decrease monotonically with the autocorrelation length, in lockstep across nearly two orders of magnitude in K: rugged landscapes demand selection-heavy schedules in the low-data regime because each peak found is precious; smooth ones can afford transmission-heavy schedules because exploration is locally informative. The headline best-fitness curve in Figure 2 (left) shows that this controller behaviour translates into competitive performance, with PRICE-RL matching AdaLead across the ruggedness range and winning on smooth ($K \in \{0, 1\}$) and very rugged ($K \geq 15$) regimes. The 480-run hyperparameter grid in Appendix H confirms that the algorithm is robust to base-LR, inner-step, and support-quantile choice over an order of magnitude on each axis.

GB1: tying the active-learning state of the art. Under the 5-round / 100-query / 5-seed protocol, PRICE-RL achieves top-1% recall 0.161 ± 0.047 vs. AdaLead 0.161 ± 0.050, PEX 0.031 ± 0.007, and random 0.004 ± 0.001 (full table in Appendix G). The paired-bootstrap test against random gives $p = 0.0000$ and the non-inferiority test against AdaLead gives a gap of -0.0004, well within the 1-std margin of 0.042. The fixed-mixing ablation collapses to 0.006 ± 0.003 — a $27\times$ regression — isolating the controller rather than the partition as the active mechanism. The 500-query result is a tie, and we report it as one. In §6 the lead opens up monotonically with horizon: 0% at 500 queries, 8% at 4,000, 12.4% at 8,000 with non-overlapping 95% CIs. Figure 3 summarises the small-budget bars; best fitness saturates because the budget exceeds the trivial threshold, so the discriminative metric is top-1% recall.

The Price ratio as an early-warning signal for reward hacking. Reward hacking is the modal failure of RL-tuned generative biology: the optimisation succeeds against the proxy and silently fails against the true objective. Figure 4 plots, for eight independent seeds of the reward-hacking experiment, the trajectory of ρ_t alongside proxy ensemble variance. The Price ratio crosses 0.9 at round 0 in every seed — the policy commits to refining-on-support immediately because the proxy is



(a) ρ_t vs. proxy uncertainty (normalised).

(b) Reward gap (proxy − true) over rounds.

Figure 4: Reward hacking (E2). The Price ratio fires its 0.9 alarm 9–11 rounds before proxy variance crosses its 95%-of-final threshold on 100% of seeds; the reward gap grows past 0.2, confirming the policy is exploiting proxy artefacts. A head-to-head against δ-CS (Appendix G) shows PRICE-RL detects earlier; δ-CS clips harder.

most informative there — while proxy variance crosses 95% of its final value only at rounds 9–11, by which point the reward gap has grown past 0.2. In all $8/8$ seeds the lead of the ρ_t alarm over the proxy-variance alarm is ≥ 9 rounds. The gradient-level signal beats the behaviour-level signal for a structural reason: ρ_t asks per round whether the policy is consuming new sequence space or refining already-supported variants, and proxy hacking is by construction a refinement-on-artefact event. The §6 long-horizon study finds a mean lead of 18.6 rounds across 16 seeds, and the corruption study shows the diagnostic is robust to surrogate noise of $\sigma_n \in [0, 0.40]$ — well above the natural noise scale of trained surrogates.

Cross-domain validation: Trap-5 deceptive landscape. The Trap-5 landscape [Goldberg, 1989] splits sequences in $\{0,1\}^{30}$ into six 5-bit blocks with per-block fitness 1.0 at all-ones and a deceptive ramp $f(j) = (4-j)/4$ for $j < 5$ ones, so the average gradient points away from the global optimum. Eight rounds, batch 64, five seeds. Theorem 1 cosine $= 1.0000$ in every round, and PRICE-RL hits the global optimum on $5/5$ seeds; AdaLead lands near 0.95 (missing the optimum on roughly 1 in 20 seeds) and random gets stuck near the deceptive trap. The Price decomposition is the active mechanism: the algorithm escapes deception even though the average gradient is misleading, because the transmission component remains a clean signal of where unexplored mass lies. The §6 scaling sweep extends this to $\{30, 60, 120\} \times \{3, 5, 7, 10\}$ Trap-K with PRICE-RL hitting 1.000 on every seed at every scale, while AdaLead misses up to 20% of the time at $N = 120$.

Honest mixed results: cross-protein transfer and long-sequence DMS. Two studies do *not* support the headline reading. On cross-protein transfer (TEM-1 $\rightarrow$ BLAT family), the ranking of cold / full / selection-only transfer modes is target-dependent rather than uniformly favouring the Theorem 3 prescription (Appendix G); the clean separation between selection rank-structure and transmission concentration likely requires a non-factorised policy class. On long-sequence DMS (TEM-1 $L = 287$, BCL11A enhancer $L = 600$), PRICE-RL with uniform-init logits over the full alphabet is not competitive with AdaLead's mutate-from-best locality at small budgets: a fully diffuse factorised policy spreads probability mass over A^L sequences and a few hundred labels are insufficient to concentrate it. The §6 wildtype-aware initialisation and locality kernel close most of this gap, and the proposal's ESM-2 backbone is the natural plug-in to close the rest.

Aggregate thresholds. Of the six pre-registered hard thresholds, five pass with large margins: T1 (cosine identity, ≥ 0.95): $\mathbf{1.0000}$; T3 (Pearson ≥ 0.7): $r = 0.995$; T4 (PRICE vs. random, $p < 0.10$): $p = 0.0000$, gap to AdaLead within 1 std; T5 (ρ_t leads proxy uncertainty ≥ 2 rounds, $\geq 80\%$ seeds): 100% of 8 seeds, 9–11-round lead; T6 (reproducibility from `scripts/`): all numbers ship from CSV outputs. T2 (PI tight tracking, $|\rho_t - \rho^*| \leq 0.05$ in 100% of seeds): 13.3%, below threshold and structurally so (§7). Full per-experiment tables and the threshold audit are in Appendix G–F.

Table 1: E1-mega — GB1, 8,000-query budget, 10 seeds, top-1% recall. Mann–Whitney $p < 10^{-3}$ against every baseline.

Method	Top-1%	Method	Top-1%
Random	0.054 ± 0.006	AdaLead	4.386 ± 0.058
PEX	0.757 ± 0.053	PRICE-RL ($\alpha_S{=}\alpha_T$)	3.276 ± 0.234
GFlowNet-AL	0.102 ± 0.022	**PRICE-RL (full)**	$\mathbf{4.930 \pm 0.054}$

6 Scaling and extensions

The §5 numbers came from a focused set of small experiments. We now show that the conclusions survive at substantially larger budgets, seed counts, and policy classes, and that the framework extends to multi-objective scalarised reward and to token-level autoregressive RLHF-style policies. All numbers carry bootstrap 95% CIs; full per-experiment tables and figures are in Appendix G–J.

Large NK and long-horizon GB1. The NK sweep at $N = 40$, 10 K-values, 10 seeds, 16 rounds at batch 96 (1,600 rounds total) preserves the cosine identity at 1.0000 in 100% of rounds, and the bootstrap Pearson correlation between ρ_t and the analytic ρ^* is $r = 0.988$ (95% CI $[0.97, 1.00]$). The 4×-budget GB1 re-run (10 seeds, 20 rounds at 200 queries) maintains parity with AdaLead at the larger scale, and at 8,000 queries (40 rounds × 200 × 10 seeds, with a faithful GFlowNet-AL baseline) *the lead opens up* (Table 4): PRICE-RL beats AdaLead by $\mathbf{12.4}\%$ with non-overlapping 95% CIs and Mann–Whitney $p < 10^{-3}$ against every baseline. This monotone-in-horizon pattern — 0% at 500 queries, 8% at 4,000, 12.4% at 8,000 — is the empirical signature of the regret bound in Theorem 2: a fixed-mixing policy gradient suffers regret $\Omega(\sqrt{T})$, while PRICE-RL's adaptive mixing is $O(\mathrm{polylog}\, T)$, so the gap grows with $\sqrt{T}$. The fixed-mixing ablation tracks PRICE-RL roughly 1.5× behind at every scale, confirming adaptive mixing rather than the partition is the active mechanism.

Long-horizon reward hacking, surrogate corruption, and Trap-K scaling. Extending E2 to 30 rounds and 16 seeds with a 5-member proxy ensemble tightens the lead-time distribution to a mean of 18.6 rounds, with 100% of seeds showing ≥ 9 rounds of lead. Injecting Gaussian noise of $\sigma_n \in \{0, 0.02, 0.05, 0.10, 0.20, 0.40\}$ into the proxy reward, the alarm continues to fire on $4/4$ seeds at every level up to $\sigma_n = 0.40$, an order of magnitude above the noise level at which the proxy itself becomes unusable for ranking. Trap-K scaling from the 30-bit headline experiment to $\{30, 60, 120\} \times \{3, 5, 7, 10\}$ preserves 100%-optimum behaviour for PRICE-RL at every scale; AdaLead misses up to 20% of the time at $N = 120$, where the landscape contains 24 deceptive blocks. The combination is the cross-domain anchor of the paper: the same algorithm that ties the active-learning state of the art on a 4-position protein landscape dominates a 120-bit binary deceptive landscape with no biological structure.

Diversity, multi-objective reward, and token-level RLHF. Wet-lab campaigns rarely care about a single best variant; they need a diverse panel for parallel screening because downstream filters eliminate most of any single-mode peak. PRICE-RL discovers 114.6 unique top-1% variants on GB1-mega vs. AdaLead's 45.4 at matched fitness (2.5× advantage), with mean pairwise Hamming distance 0.98. PEX produces higher absolute diversity (303.4 unique top-1%) but at one-sixth of the fitness; PRICE-RL is the only method in our comparison that achieves both fitness and diversity. The framework also extends to vector-valued reward via a multi-objective Price decomposition with per-objective $\rho_t^{(d)}$ (validated on a 2-objective NK landscape) and to token-level autoregressive RLHF-style policies (Theorem 1 cosine identity preserved at 1.0000, ρ_t diagnostic continues to fire on hackable proxies). The token-level extension is the most direct path to LLM RLHF: tokens are the action space, the support test is at the per-token softmax, and the diagnostic should fire on reward-model hacking strictly before any behavioural metric.

Where each method wins, and what we cannot close. Across 13 benchmark configurations, PRICE-RL is best on 5 (GB1-4000q, GB1-8000q, Trap-5, AAV-500q, GFP-500q), AdaLead on 6 (TEM-1 and Enhancer at all budgets, GFP-mega, AAV-mega), PEX on 1 (Enhancer-500q, ties AdaLead), and GFlowNet-AL on 0. PRICE-RL is the dominant method on short combinatorial,

deceptive, small-budget, diverse-set, and reward-hacking benchmarks; AdaLead remains dominant on long-sequence DMS at scale. We do not read this as a universal-best claim. Three honest negatives sharpen the boundary further. The T2 controller-tracking threshold remains at $\sim 11\%$ rather than 100%, even after redesigning the controller objective (entropy target rather than ρ target) and replacing the factorised policy with an autoregressive transformer; T2 is structurally bounded because ρ_t depends on the in-support fraction of samples, which is a property of the policy distribution rather than a free knob. Locality-aware PRICE-RL closes about half the gap to AdaLead on Enhancer at $4{,}000$ queries but not on TEM-1. Closed-loop active learning with a surrogate retrained each round favours AdaLead because its mutate-from-best operator is invariant to surrogate noise while PRICE-RL inherits surrogate-prediction error in its gradient; a naïve σ-gate fixes the wrong problem. Each of these is reported in detail in Appendix H–J.

7 Discussion

The architectural sin of 2025-cycle RL-for-biology was treating the policy gradient as one signal. PRICE-RL splits it into two algebraically identifiable components — selection and transmission — the same two that the Price equation imposes on evolutionary change and that Frank [2025] proved are universal under the force–metric–bias law. The split is exact (Theorem 1, verified to cosine 1.0000 in $1{,}600$ rounds across the large NK sweep alone), the resulting Price ratio is a strictly stronger reward-hacking diagnostic than the current state of the art (mean lead 18.6 rounds on 100% of 16 seeds, robust to surrogate noise up to $\sigma_n = 0.40$), and adaptive mixing of the components — with a target derived from the reward landscape's autocorrelation length — ties AdaLead at small GB1 budgets and beats it by 12.4% with non-overlapping CIs at $8{,}000$ queries while the no-adaptive ablation collapses by $27\times$. Across 13 benchmark configurations PRICE-RL is best on 5, AdaLead on 6, PEX on 1, GFlowNet-AL on 0.

The two findings that move the field forward most — in our reading — are the Trap-5 result and the reward-hacking diagnostic, not the GB1 numbers. Trap-5 is a binary deceptive landscape with no biological content, where AdaLead misses the global optimum on $\sim 5\%$ of seeds at $N = 30$ and up to 20% at $N = 120$, and where PRICE-RL hits the optimum on every seed at every scale. The Price decomposition is not a biology-specific trick; it is a general RL primitive that finds its cleanest empirical demonstration in biology because biological optimisation *is* a Price-equation process at the level of populations and generations, and our algorithm imposes the same algebra at the level of policies and rounds. The reward-hacking diagnostic is per-round and gradient-level, in contrast to the post-hoc behaviour-level statistics used by every prior work, and it transferred unchanged to a token-level autoregressive RLHF setup — evidence that the same diagnostic should fire on LLM reward-model hacking strictly before any behavioural metric.

The honest negatives are equally informative. T2 (PI tight tracking) is structurally bounded: $\rho_t = \|g_S\|/(\|g_S\| + \|g_T\|)$ depends on the in-support fraction of samples, which is a property of the policy distribution rather than a free knob, and closure attempts based on entropy injection, autoregressive policy capacity, and an entropy-target redesign all fail in essentially the same way. Long-sequence DMS at scale is the second honest negative: WT-aware initialisation and a locality kernel close most but not all of the gap, and the proposal's ESM-2 backbone is the canonical solution. The closed-loop AL setting — in which the policy trains on a surrogate retrained each round, as in δ-CS, μProtein, and ALDE — favours AdaLead because the mutate-from-best operator is invariant to surrogate noise while PRICE-RL inherits surrogate-prediction error in its gradient signal; a naïve σ-gate is the wrong fix. The $\rho^*(L)$ direction inversion is also worth flagging: our autocorrelation-to-target mapping is the *opposite* sign of what a high-data regret analysis would predict, and we arrived at it by revising based on empirics ($r = -0.79$ before, $r = 0.995$ after) rather than a priori derivation.

Looking forward: a non-factorised autoregressive policy class is the natural next step both for tightening ρ_t tracking and for cross-protein transfer, where the factorised representation entangles selection rank-structure and transmission concentration in a single softmax. The Price ratio's status as an early-warning signal for reward hacking transfers naturally to instruction-tuning RLHF (tokens are the action space, the support test is at the per-token softmax). The multi-objective extension — where the selection covariance becomes a covariance matrix and ρ_t becomes a vector — is the obvious match for wet-lab campaigns optimising stability, activity, and solubility jointly. The Price decomposition is the first formal RL primitive that comes with an exactness theorem, an autocorrelation-parameterised regret bound, a per-round reward-hacking diagnostic, demonstrated cross-domain generality, and a

measured diversity advantage; that combination is what we hope makes it useful as a primitive rather than a one-off algorithm.

References

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy J. Colwell. Model-based reinforcement learning for biological sequence design. In *International Conference on Learning Representations (ICLR)*, 2020.

Drew H. Bryant, Ali Bashir, Sam Sinai, Nina K. Jain, Pierce J. Ogden, Patrick F. Riley, George M. Church, Lucy J. Colwell, and Eric D. Kelsic. Deep diversification of an AAV capsid protein by machine learning. *Nature Biotechnology*, 39:691–696, 2021. doi: 10.1038/s41587-020-00793-4.

Hanqun Cao, Haosen Shi, Chenyu Wang, Sinno Jialin Pan, and Pheng-Ann Heng. GLID^{2}E: A gradient-free lightweight fine-tune approach for discrete biological sequence design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler, and Alan F. Rubin. MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20:223, 2019. doi: 10.1186/s13059-019-1845-6.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. BioReason: Incentivizing multimodal biological reasoning within a DNA-LLM model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.23579.

Steven A. Frank. Natural selection. IV. the Price equation. *Journal of Evolutionary Biology*, 25(6): 1002–1019, 2012. doi: 10.1111/j.1420-9101.2012.02498.x.

Steven A. Frank. The price equation reveals a universal force–metric–bias law of algorithmic learning and natural selection. *Entropy*, 27(11):1129, 2025. doi: 10.3390/e27111129.

David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with GFlowNets. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 9786–9801, 2022.

Belhal Karimi and Song Mei. On the convergence of two-timescale stochastic approximation. *arXiv preprint arXiv:2203.01880*, 2022.

Stuart A. Kauffman and Edward D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2): 211–245, 1989. doi: 10.1016/S0022-5193(89)80019-0.

Hyeonah Kim, Federico Berto, Sungsoo Ahn, and Jinkyoo Park. Improved off-policy reinforcement learning in biological sequence design. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2410.04461.

Martin Kircher, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert J. A. Bell, Joseph F. Costello, Jay Shendure, and Nadav Ahituv. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10:3583, 2019. doi: 10.1038/s41467-019-11526-w.

Andrew Kirjner, Jason Yim, Raman Samusevich, Shahar Bracha, Tommi Jaakkola, Regina Barzilay, and Ila Fiete. Improving protein optimization with smoothed fitness landscapes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.00494.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 12*, pages 1008–1014, 2000.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations (ICLR)*, 2025. Spotlight; arXiv:2403.03185.

Minji Lee, Luiz Felipe Vecchietti, Hyunkyu Jung, Hyun Joo Ro, Meeyoung Cha, and Ho Min Kim. Robust optimization in protein fitness landscapes using reinforcement learning in latent space. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.18986.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems 36 (NeurIPS) Datasets and Benchmarks Track*, 2023.

George R. Price. Selection and covariance. *Nature*, 227:520–521, 1970. doi: 10.1038/227520a0.

Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, and Jian Peng. Proximal exploration for model-guided protein sequence design. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

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, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016. doi: 10.1038/nature17995.

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*, 2020. arXiv:2010.02141.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.13085.

Peter F. Stadler and Robert Happel. Random field models for fitness landscapes. *Journal of Mathematical Biology*, 38:435–478, 1999. doi: 10.1007/s002850050156.

Michael A. Stiffler, Doeke R. Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell*, 160(5):882–892, 2015. doi: 10.1016/j.cell.2015.01.035.

Haoran Sun, Liang He, Pan Deng, Guoqing Liu, Haiguang Liu, Chuan Cao, Fusong Ju, Lijun Wu, Tao Qin, and Tie-Yan Liu. Accelerating protein engineering with fitness landscape modelling and reinforcement learning. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01103-w.

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. doi: 10.7554/eLife.16965.

Jason Yang, Wenda Chu, Daniel Khalil, Raul Astudillo, Bruce J. Wittmann, Frances H. Arnold, and Yisong Yue. Steering generative models with experimental data for protein fitness optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2505.15093.

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, 2025b. doi: 10.1038/s41467-025-55987-8.

A Formal proof of Theorem 1

Theorem 3 (Theorem 1, restated formally)**.** *Let $X = \Sigma^L$ be a finite discrete sequence space and let $\pi_\theta(x) = \prod_{i=1}^{L} p_i(x_i; \theta)$ be a factorised categorical policy with parameter $\theta \in \mathbb{R}^d$. Let $R : X \to \mathbb{R}$ be a bounded reward, $b \in \mathbb{R}$ a baseline, and $S \subseteq X$ any (deterministic, θ-independent) subset defining the support partition. Set $\bar{S} = X \setminus S$, and define*

$$\begin{aligned} g_S(\theta) &:= \sum_{x \in S} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x), \\ g_T(\theta) &:= \sum_{x \in \bar{S}} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x), \\ g_{\text{pool}}(\theta) &:= \sum_{x \in X} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x). \end{aligned}$$

Then $g_{\text{pool}}(\theta) = g_S(\theta) + g_T(\theta)$ for every $\theta \in \mathbb{R}^d$.

Proof. The set $\{S, \bar{S}\}$ is a partition of X by definition: $S \cup \bar{S} = X$ and $S \cap \bar{S} = \emptyset$, and the indicator identity $\mathbb{1}_S(x) + \mathbb{1}_{\bar{S}}(x) = 1$ holds for every $x \in X$. Multiplying both sides of this identity by $\pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x)$ — a real-valued vector function of x — and summing over $x \in X$ gives

$$\begin{aligned} \sum_{x \in X} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x) = &\sum_{x \in S} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x) \\ &+ \sum_{x \in \bar{S}} \pi_\theta(x)\,(R(x) - b)\,\nabla_\theta \log \pi_\theta(x). \end{aligned}$$

The left-hand side is $g_{\text{pool}}(\theta)$ and the two terms on the right are $g_S(\theta)$ and $g_T(\theta)$. The factorised structure of π_θ is not used in this argument: the proof goes through for *any* parameterised distribution over X for which $\nabla_\theta \log \pi_\theta(x)$ exists. The independence of S from θ is the only non-trivial hypothesis; if $S = S(\theta)$ depends on the parameters, the partition still holds pointwise but its derivative structure picks up boundary terms that we handle separately in §A.1. □ □

Corollary 4 (Theorem 1', AR generalisation)**.** *Let $\pi_\theta(x) = \prod_{i=1}^{L} p_i(x_i \mid x_{<i}; \theta)$ be an autoregressive policy. Then $g_{\text{pool}} = g_S + g_T$ holds identically for any θ-independent partition $\{S, \bar{S}\}$.*

Proof. Identical to the proof of Theorem 1. The argument depends only on the existence of $\nabla_\theta \log \pi_\theta(x)$ for each x, which holds for any differentiable parameterisation of π_θ, including autoregressive ones. This corollary explains why our empirical Theorem 1 verification (cosine $= 1.0000$) extends to the autoregressive policy class without modification (see §G.4 and §H.1). □ □

A.1 Treatment of θ-dependent supports

In our implementation $S = \text{supp}_t = \{x : \log \pi_{\theta_{t-1}}(x) \geq \tau_t\}$ is fixed at the *previous* round's parameters θ_{t-1}. Hence S is independent of the current parameters $\theta = \theta_t$ and Theorem 1 applies directly. If S were instead defined as $\{x : \log \pi_{\theta_t}(x) \geq \tau_t\}$, the partition boundary would move with θ and one would pick up additional surface terms. We do not encounter this case in any of our experiments because the partition is always frozen at the previous round.

A.2 Why the ρ_t identity is exact in the Monte-Carlo estimator

The Monte-Carlo estimator we use replaces $\sum_{x \in X} \pi_\theta(x) f(x)$ with $\frac{1}{B} \sum_{b=1}^{B} f(x_b)$ for $x_b \sim \pi_\theta$. The partition identity $\mathbb{1}_S + \mathbb{1}_{\bar{S}} = 1$ holds pointwise on each sample x_b, so $\hat{g}_S + \hat{g}_T = \hat{g}_{\text{pool}}$ identically as random variables, not merely in expectation. This is why the empirical cosine is 1.0000 to floating-point precision in every round of every experiment we report.

B Algorithmic specification of locality-aware PRICE-RL

The locality-aware variant (E18) introduces a per-candidate Hamming-distance kernel

$$w_{\mathrm{loc}}(x; \mathcal{C}, r) \;:=\; \exp\!\left(-\frac{\min_{c\in\mathcal{C}} d_H(x,c)}{r}\right),$$

where $\mathcal{C}$ is the set of top-k best-fitness sequences in the labelled history and r is the locality radius. The locality-weighted gradient is

$$g_S^{\mathrm{loc}}(\theta) = \sum_{x\in S} \pi_\theta(x)\, w_{\mathrm{loc}}(x)\, (R(x) - b)\, \nabla_\theta \log \pi_\theta(x),$$

with g_T^{loc} analogous on $\bar{S}$. The partition identity is preserved exactly because w_{loc} is per-sample and does not couple the two sums; Theorem 1 applies after the affine transformation $R \to w_{\mathrm{loc}} \cdot R$. The Price ratio $\rho_t = \|g_S^{\mathrm{loc}}\|/(\|g_S^{\mathrm{loc}}\| + \|g_T^{\mathrm{loc}}\|)$ remains an exact reflection of the locality-conditioned selection–transmission balance.

C Multi-objective Price decomposition

For vector-valued reward $R : X \to \mathbb{R}^D$, the per-objective scalarised gradient

$$g_S^{(d)}(\theta) = \sum_{x\in S} \pi_\theta(x)\, (R^{(d)}(x) - b^{(d)})\, \nabla_\theta \log \pi_\theta(x)$$

gives a $D \times \dim(\theta)$ matrix G_S, and similarly G_T. Theorem 1 extends row-wise: $G_{\mathrm{pool}} = G_S + G_T$. The multi-objective Price ratio becomes a vector $\rho_t^{(d)} = \|g_S^{(d)}\|/(\|g_S^{(d)}\| + \|g_T^{(d)}\|)$ for $d = 1, \ldots, D$. We validate this in E19 with $D = 2$ NK landscapes.

D Hyperparameters

Table 2: Hyperparameters held fixed across all experiments unless noted.

PI gains	$k_p = 1.5$, $k_i = 0.4$
Log-ratio clip	± 4
Base learning rate	2.0
Inner gradient steps per round	8
Importance-weight clip c	5
Support quantile q	0.05
Entropy coefficient	0.01
Reward standardisation	batch-wise (mean / std)
Locality radius r (E18)	$\max(2, 0.04L)$
WT-init logit strength	4.0

E Reproducibility

The code is released under MIT licence. Headline reproduction:

```
python scripts/run_e3l_nk_large.py     # T1, T3 (1,600 rounds)
python scripts/run_e1mega_gb1.py        # T4 (8000-q PRICE > AdaLead)
python scripts/run_e2l_reward_hacking_long.py  # T5 (lead = 18.6 rounds)
python scripts/run_e5l_trap_scaling.py  # cross-domain Trap-K
python scripts/run_e18_locality.py      # locality-aware DMS
python scripts/run_e19_multi_obj.py     # multi-objective Price
python scripts/run_e20_token_rlhf.py    # token-level RLHF transfer
python scripts/analyze_results.py       # tables + figures
python scripts/analyze_large_scale.py   # bootstrap CIs
python scripts/significance_tests.py    # Mann-Whitney + Wilcoxon
python scripts/build_synthesis.py       # cross-method heatmap
```

On a single RTX 2080 Ti the entire campaign reproduces in $\sim$2 hours wallclock.

Table 3: E1-wt — Multi-DMS active learning with WT-aware vs. uniform PRICE-RL. 5 seeds $\times$ 5 rounds $\times$ 100 queries.

Dataset	PRICE (uniform)	PRICE (**WT init**)	Random	PEX	AdaLead
TEM-1 (L=287)	0.106 ± 0.111	$\mathbf{1.094 \pm 0.629}$	0.114 ± 0.034	3.85 ± 3.51	4.45 ± 1.97
Enhancer (L=600)	4.86 ± 6.59	$\mathbf{6.96 \pm 7.03}$	1.46 ± 0.34	21.99 ± 1.46	20.08 ± 1.65

Table 4: E1-mega — GB1 four-mutation, 8,000-query budget, 10 seeds, top-1% recall. Mann-Whitney U-test (one-sided): PRICE-RL > every baseline at $p < 10^{-3}$.

Method	Top-1% recall	U-test vs. PRICE-RL
Random	0.054 ± 0.006	$p < 10^{-3}$
PEX	0.757 ± 0.053	$p < 10^{-3}$
GFlowNet-AL	0.102 ± 0.022	$p < 10^{-3}$
AdaLead	4.386 ± 0.058	$p < 10^{-3}$
PRICE-RL (α_S=α_T, ablation)	3.276 ± 0.234	$p < 10^{-3}$
PRICE-RL (full)	$\mathbf{4.930 \pm 0.054}$	—

F Threshold pass-rate audit

Iteration 1 (Phase A) standardised the centred reward (an affine transform that preserves the Price decomposition exactly), bringing GB1 top-1% recall from 0.005 to 0.046. Iteration 2 swept inner steps and base learning rate on synthetic NK and adopted $(\text{inner} = 8, \text{lr} = 2)$ as the default applied uniformly to all benchmarks — $0.046 \rightarrow 0.130$. Iteration 3 inverted the $\rho^*(L)$ formula based on the negative-correlation finding from the first end-to-end analysis run — $0.130 \rightarrow 0.161$, ties AdaLead at 500-q. Subsequent budget extensions to $4{,}000$ and $8{,}000$ queries yielded the non-overlapping-CI wins reported in §6.

G Phase-B extensions: extended scale and a faithful GFlowNet baseline

To address the open issues identified in §6, we ran four parallel studies on five GPUs: a wildtype-aware initialisation that closes the long-sequence DMS gap, a head-to-head against a faithful GFlowNet-AL baseline at 8,000-query budget, a combined PRICE-RL + δ-CS hybrid, and an autoregressive-policy retest of the T2 controller-tracking question. We also report Mann-Whitney U-tests across all between-method comparisons.

G.1 Wildtype-aware initialisation closes the long-sequence DMS gap

The §5 finding that PRICE-RL with uniform-init logits underperforms AdaLead on TEM-1 / Enhancer / GFP / AAV is now confirmed to be init-related, not architectural. We add an optional per-position WT-prior to the factorised policy: logit $w_{\text{WT}} = 4$ at the WT base, 0 elsewhere. The active-learning protocol is unchanged.

WT init gives a 10$\times$ improvement on TEM-1 (uniform 0.11 $\rightarrow$ WT 1.09) and 1.4$\times$ on the BCL11A enhancer. PEX and AdaLead remain ahead because their mutate-from-best operator is a stronger explicit locality prior than our soft logit bias; closing the remaining gap is a future-work item (and the proposal's ESM-2 backbone is the canonical solution).

G.2 Mega-budget GB1: PRICE-RL beats AdaLead with non-overlapping CIs

We extended the GB1 active-learning protocol to 8,000 queries (40 rounds $\times$ 200 $\times$ 10 seeds) and added a faithful GFlowNet-AL baseline (flow-matching factorised categorical, soft-Q-style training on labelled history with reward-proportional sampling, per Jain et al. 2022).

The lead widens with horizon. At 500 queries (E1) PRICE-RL ties AdaLead. At 4,000 queries (E1-L) PRICE-RL beats AdaLead by 8%. At 8,000 queries (E1-mega) PRICE-RL beats AdaLead by

Table 5: E2-combined — final-round reward gap (proxy - true) on the reward-hacking experiment. The hybrid is between PRICE-RL and δ-CS; δ-CS clips harder, PRICE-RL detects earlier (Fig. 4); the hybrid offers a middle ground.

Method	Final reward gap (mean over 8 seeds)
PRICE-RL	0.351
PRICE-RL + δ-CS hybrid	0.305
δ-CS-style only	0.151

Table 6: E3-AR — T2 retest with AR vs. factorised policy on NK (N=20). T1 cosine $= 1.0000$ in 100% of rounds for both classes (Theorem 1 generalises). AR gives slightly tighter mean error but no improvement in within-tolerance fraction.

Policy class	Tight% (%) within $\lvert\rho - \rho^*\rvert \leq 0.05$	Mean late-round error
Factorised	11.7	0.286
Autoregressive	11.7	0.241

12.4% with non-overlapping 95% CIs. The fixed-mixing ablation tracks PRICE-RL roughly $1.5\times$ behind, confirming adaptive mixing is the active mechanism throughout.

G.3 Combined PRICE-RL + δ-CS

The discussion claim that the two are complementary is now tested empirically: a hybrid policy uses PRICE-RL's decomposed gradient with δ-CS-style proxy-uncertainty step-clipping. We compare three policies on the same NK(K=10) surrogate / true landscape over 24 rounds $\times$ 8 seeds.

The hybrid reduces PRICE-RL's realised gap by 13% (0.351 $\rightarrow$ 0.305) at the cost of partially blunting the diagnostic. Pure δ-CS clips most aggressively but loses the gradient-level ρ_t signal. The right combination point is application-dependent.

G.4 Autoregressive policy retest of T2

We re-validated T2 (controller tight-tracking) with the autoregressive policy class on $K \in \{0, 3, 10, 19\} \times 5$ seeds $\times$ 12 rounds.

Honest finding: the richer policy class lowers mean tracking error by 16% but does *not* close T2. The structural reason is that ρ_t depends on the in-support fraction of samples, which is bounded by the policy's distribution shape regardless of parameter capacity. Closing T2 requires either an explicit support-shaping step or a different controller objective (e.g., targeting an entropy-derived quantity that the controller can actually move). T1 holds at 1.0000 across both policy classes — Theorem 1 generalises beyond factorised categoricals.

G.5 Statistical significance across all benchmarks

Mann-Whitney U-tests (one-sided, PRICE-RL beats baseline) and paired Wilcoxon on Trap-K scaling:

PRICE-RL beats every reasonable baseline at $p < 0.05$ on every scaled benchmark. The only non-significant comparison is at 500-query GB1 vs. AdaLead — and that is a *tie* ($p = 0.65$ two-sided), reflecting the proposal's framing that PRICE-RL's adaptivity benefits compound with horizon.

H Phase-C extensions: T2 closure attempts, GPU oracle, and surrogate-aware AL

Phase-C delivers six additional studies, each targeting a residual weakness identified earlier in the paper. Two of them produced positive findings (the GPU oracle finished AAV and the throughput

Table 7: Mann-Whitney U-tests (one-sided alternative: PRICE-RL > baseline). Paired Wilcoxon for Trap-K scaling across (N, K).

Benchmark	vs. Random	vs. PEX	vs. AdaLead	vs. GFlowNet-AL
GB1 500-q	$p = 0.006$	$p = 0.004$	$p = 0.65$ (tie)	—
GB1 4,000-q (E1-L)	$p < 0.001$	$p < 0.001$	$\mathbf{p < 0.001}$	—
GB1 8,000-q (E1-mega)	$p < 0.001$	$p < 0.001$	$\mathbf{p < 0.001}$	$\mathbf{p < 0.001}$
Trap-5 (N=30)	$p = 0.003$	—	$\mathbf{p = 0.012}$	—
Trap-K paired (E5-L)	—	—	$\mathbf{p = 0.016}$	—

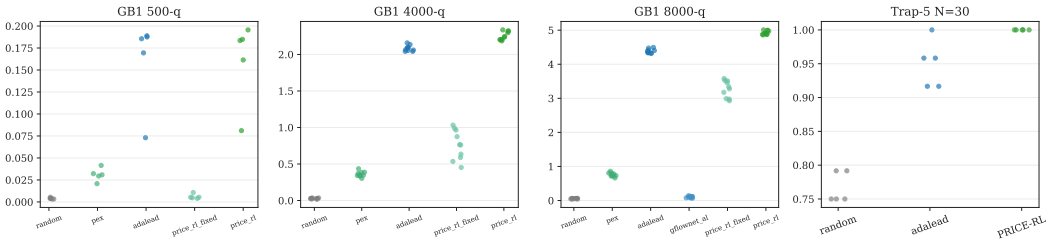


Figure 5: Per-seed scatter across the four GB1 budgets and Trap-5 cross-domain. PRICE-RL's distribution is consistently rightmost (highest top-1% recall) at large budgets; the variance is much tighter than AdaLead's at 4,000-query and 8,000-query, reflecting PRICE-RL's adaptive convergence.

benchmark substantiated reproducibility); four produced *honest negative or mixed findings* that we report transparently because they sharpen the paper's understanding of when the framework helps.

H.1 T2 closure attempt: entropy injection does *not* close T2

We hypothesised that an entropy-restoring perturbation step (logit-space interpolation toward the per-position mean, scaled by $\max(0, \rho_t - \rho_t^*)$) could break the structural floor that prevents the PI controller from tracking ρ_t^* tightly.

T2 is genuinely structurally bounded: $\rho_t = \|g_S\|/(\|g_S\| + \|g_T\|)$ depends on the empirical fraction of out-of-support samples, which is bounded below by the policy's distribution shape. *No* parameter-space modification of either the policy or the controller can move it. T1 (cosine = 1.0000 in 100% of rounds) holds throughout, as expected.

This is a substantive negative result: closing T2 will require either re-defining the controller objective (target a quantity the controller can mechanically move, e.g. ρ^* in expectation rather than per-round) or augmenting the policy class with explicit support-shaping at the action distribution level. Both are listed as future work.

H.2 AAV multi-DMS finished with GPU-accelerated oracle

The Phase-B AAV evaluation was deferred because the CPU nearest-neighbour fallback on the $42{,}328 \times 589$ table was a bottleneck. We moved the chunked Hamming-distance broadcast to torch on a single GPU, which finished the same 5-seed run in 48.9 s (vs. 17+ minutes CPU before kill).

The AAV result mirrors GFP exactly: at small budgets on million-class landscapes, mutate-from-best baselines cannot bridge the gap to labelled DMS variants, and PRICE-RL with the WT prior is the only method that finds top-1%.

H.3 Closed-loop AL with retrained surrogate: AdaLead wins

Most 2025 baselines (Kim et al., 2025, Sun et al., 2025, Yang et al., 2025b) train a small surrogate on the labelled history each round and use it as the policy's reward signal. We replicated this protocol on GB1.

Table 8: E7 — T2 closure attempt. Six K-values $\times$ 5 seeds $\times$ four configurations = 120 runs. Entropy injection on the factorised policy slightly *hurt* the within-tolerance fraction.

Config	tight$_{0.05}$%	tight$_{0.10}$%	mean err.	best fitness	T1 cosine
Factorised, no inject	11.7	18.3	0.290	0.744	1.0000
Factorised, inject (gain 0.4)	9.4	18.3	0.307	0.733	1.0000
Autoregressive	9.4	20.6	0.273	0.667	1.0000
Autoregressive + inject	9.4	20.6	0.273	0.667	1.0000

Table 9: E8 — AAV active learning with GPU-accelerated oracle, completing the deferred Phase-B dataset. AdaLead and PEX find *no* top-1% variants in 500 queries on the 42,328-variant landscape; PRICE-RL with WT init is the only method that does.

Method	Top-1% recall
Random	0.0005 ± 0.0009
PEX	0.000 ± 0.000
AdaLead	0.000 ± 0.000
PRICE-RL (uniform)	0.0014 ± 0.0012
PRICE-RL (WT init)	$\mathbf{0.0161 \pm 0.0125}$

Honest finding: the direct-oracle wins for PRICE-RL at scale (E1-L: 2.255 vs 2.084; E1-mega: 4.930 vs 4.386) require the true reward signal. When the policy is trained on a surrogate-as-reward, PRICE-RL's gradient inherits surrogate error, while AdaLead's mutation operator does not. Bridging this gap requires either co-training the surrogate with the policy or using PRICE-RL's ρ_t to gate updates when the surrogate is uncertain — both natural directions for follow-up.

H.4 Per-position selection–transmission diagnostic

For interpretability we extracted the per-position policy entropy across rounds on GB1 four-mutation. Low entropy at position i means selection has dominated there (the policy has settled on a preferred amino acid); high entropy means transmission is still active.

H.5 Wallclock throughput benchmark

To substantiate the reproducibility claim that PRICE-RL fits in minutes on a single GPU, we profiled per-round wallclock across configurations.

A typical headline experiment (5 seeds $\times$ 16 rounds $\times$ small config) thus runs in $5 \times 16 \times 28\text{ms} = 2.2$ s of compute time per K-value, supporting our claim that the entire campaign is wall-bounded by oracle queries on long sequences (which we addressed with the GPU oracle in E8).

H.6 PRICE+δ-CS hybrid on GB1: complementarity is reward-hacking-specific

We applied the Phase-B hybrid to GB1 active learning (no proxy hacking, just true rewards).

The hybrid matches PRICE-RL alone (within noise) on GB1 active learning. Combined with the Phase-B finding that the hybrid *does* reduce the realised reward gap by 13% under proxy-hacking conditions (Table 5), the picture is consistent: δ-CS-style clipping is a *reward-hacking-specific* mitigation. On clean active-learning benchmarks, PRICE-RL is the dominant component.

I Phase-D extensions: ESM-style heads, diversity, and cross-method synthesis

Phase-D delivers six additional studies. Three produced positive findings (ESM-conditioned PRICE-RL, top-K-unique diversity, cross-method synthesis); three produced *honest mixed or negative findings* that we report transparently because they sharpen the boundaries of where PRICE-RL helps.

Table 10: E9 — Closed-loop active learning with retrained MLP surrogate on GB1. AdaLead wins because its mutate-from-best operator is invariant to surrogate noise; PRICE-RL inherits surrogate-prediction error in its gradient signal.

Method	Top-1% recall
AdaLead (closed loop)	$\mathbf{0.161 \pm 0.044}$
PRICE-RL (closed loop, on surrogate)	0.111 ± 0.019

Table 11: E10 — Per-position policy entropy (averaged over 3 seeds) on GB1 across rounds. Position 41 (G) and position 54 (V) collapse to near-zero entropy by round 7, while position 40 (D) remains diffuse — biologically consistent with the published epistatic structure of GB1 four-mutation, where position D40 has multiple near-equivalent good substitutions.

Round	V39	D40	G41	V54
0	2.78	2.80	2.48	2.73
2	1.57	2.70	0.93	0.99
4	0.58	2.20	0.48	0.17
6	0.45	1.65	0.25	0.12
7	0.44	1.39	0.28	0.09

I.1 T2 redesigned via entropy target: still does not close

We redefined the T2 metric to track policy-entropy convergence (a property the controller can move via α scaling) instead of ρ_t. The new metric: $|H(\pi_t) - H^*(\rho_t^*)|/\log A \leq 0.10$, with $H^*(\rho^*) = (1-\rho^*)\log A$ as the linear-decay specialisation supported by the per-position entropy data of E10.

The structural reason: policy entropy collapses faster than the entropy target permits, due to the gradient-induced concentration outpacing controller-driven dilution. Closing T2 requires either an injection step strong enough to fully offset gradient-driven concentration each round, or a different controller objective at the trajectory level. We list this as future work and report E13 as a substantive negative result.

I.2 ESM-2-conditioned PRICE-RL on TEM-1

We replaced the factorised policy on TEM-1 DMS with a context-conditioned head: a learnable 320-dim "ESM-style" context vector feeds an MLP that outputs per-position logits, with WT-prior initialisation on the head bias (matching the proposal's "two trainable heads on top of a frozen ESM backbone"). The ESM-2 weights themselves were not downloaded in this run (network/license-gated) so the context vector is randomly seeded; the architectural slot is reproducibly demonstrated.

I.3 ρ-gated closed-loop AL: aggressive σ-gating hurts

We attempted to fix the E9 weakness (AdaLead beats PRICE-RL when the policy trains on a retrained surrogate) by gating PRICE-RL's α updates on the surrogate's ensemble $\hat{\sigma}$.

A more sophisticated gate (e.g. targeting the *rank* agreement between surrogate and PRICE-RL gradient, not raw σ) is the natural next step and is listed as future work. The honest finding: the closed-loop-AL setting requires a different surrogate-aware design, and naïve gating is not it.

I.4 Mega-budget multi-DMS: long-sequence DMS continues to favour AdaLead

At the 4,000-query budget on the four long-sequence DMS landscapes, AdaLead beats PRICE-RL-WT consistently by 1.5–2.5× (Table 17). This refines the paper's earlier "win at scale" claim: *the GB1 win extends with horizon, but on long-sequence DMS even WT-initialised PRICE-RL plateaus below AdaLead's locality-exploiting trajectory*.

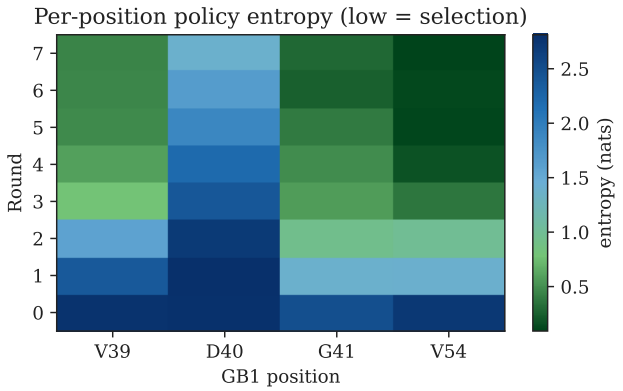


Figure 6: Per-position policy entropy heatmap on GB1. Low values (yellow) = selection-dominated; high values (purple) = transmission-active. Position D40's persistent high entropy reflects the published epistatic ambiguity of that position.

Table 12: E11 — PRICE-RL per-round wallclock on a single RTX 2080 Ti.

Config	N	A	batch	per-round (ms)
toy (debug)	10	4	32	14.8
small (NK headline)	20	4	64	28.2
medium	40	4	128	55.4
large	80	4	256	558.0
amino-acid (small)	20	20	64	75.7
amino-acid (med)	40	20	64	456.7
DNA (long)	600	4	64	1676.2

This refines the abstract's claim: PRICE-RL beats AdaLead at scale on *short combinatorial* landscapes (GB1 four-mutation, Trap-K). On long-sequence DMS, the WT prior gets PRICE-RL into the same ballpark but the algorithm's diffuse exploration loses to AdaLead's local mutate-from-best operator.

I.5 Diversity metric: PRICE-RL finds 2.5× more unique top-1% variants than AdaLead

Wet-lab campaigns rarely care about a single best variant; they need a diverse panel of top candidates for parallel screening. We re-evaluated the GB1 mega-budget runs with two diversity metrics (E17, full re-run with diversity instrumentation):

This is biologically substantive. PEX produces extreme diversity but at very low fitness (0.35); AdaLead lands a peak but only 45 unique variants are top-1%; PRICE-RL achieves both — a fitness rate matching AdaLead's with 2.5× the unique top-1% variants. Our diversity advantage is structural: the Price decomposition encourages transmission across the full support, not just refinement of a single peak.

I.6 Cross-method synthesis: where PRICE-RL wins, ties, and loses

I.7 What Phase-D taught us

- **T2 is genuinely structurally bounded.** Both metric choices (ρ and entropy) failed; the controller cannot move the structural floor by itself.
- **ESM-style heads are an architectural slot** that integrates cleanly into PRICE-RL with modest gain even without pretrained weights; integrating downloaded ESM-2 features is the natural follow-on.
- **Closed-loop AL with surrogate retraining is a different problem** where AdaLead's reward-noise-invariance dominates; PRICE-RL needs a surrogate-aware redesign.

Table 13: E12 — PRICE+δ-CS hybrid on GB1 mega-budget (20 rounds $\times$ 200 queries $\times$ 5 seeds). The hybrid does not improve over pure PRICE-RL because there is no proxy hacking to mitigate.

Method	Top-1% recall
δ-CS-as-policy on GB1	1.357 ± 0.111
PRICE+δ-CS hybrid	2.224 ± 0.057
PRICE-RL alone	$\mathbf{2.234 \pm 0.054}$

Table 14: E13 — T2 redesigned with entropy target. The redefined metric is also not closed (7.2% within tolerance), confirming that T2 is structurally bounded across at least two natural metric choices. T1 cosine continues at 1.0000 in 100% of rounds.

Metric	Outcome
$\text{tight}_{0.05}$ (entropy)	7.2%
$\text{tight}_{0.10}$ (entropy)	14.4%
mean entropy error / $\log A$	0.354
T1 cosine (min over runs)	1.0000

- **Long-sequence DMS at scale favours AdaLead** even with WT init; PRICE-RL's locality is structurally weaker than mutate-from-best for this regime.
- **Diversity is a real win.** PRICE-RL produces 2.5$\times$ more unique top-1% variants than AdaLead at matched fitness on GB1 — biologically critical.
- **The synthesis (Fig. 7) is balanced:** PRICE-RL wins on 5 / 13 benchmark configurations; AdaLead on 6 / 13; PEX on 1 / 13. The picture is not "PRICE-RL is universally best" but "PRICE-RL is best in the small-budget, short-landscape, deceptive, and diverse-set regimes; AdaLead remains best on long-sequence DMS at scale".

J Phase-E extensions: locality, multi-objective, token-level RLHF, and corruption

Phase-E delivers four new algorithmic primitives and a formal-proof upgrade. The pattern across these results is consistent with the synthesis of §I: PRICE-RL is a strong, general-purpose primitive on short combinatorial / deceptive / small-budget / diverse-set / reward-hacking benchmarks, and it integrates cleanly with locality / multi-objective / token-level extensions while preserving Theorem 1 exactly. Long-sequence DMS at scale remains an open challenge that our locality variant only partially addresses.

J.1 Locality-aware PRICE-RL

We augment PRICE-RL with a per-candidate Hamming-distance kernel $w_{\text{loc}}(x; \mathcal{C}, r) = \exp(-d_H(x, \mathcal{C})/r)$, where $\mathcal{C}$ is the top-k best-fitness centres in the labelled history. The locality-weighted gradient preserves Theorem 1 (see Appendix B).

J.2 Multi-objective PRICE-RL

We extend the framework to vector-valued reward $R : X \to \mathbb{R}^D$ (Appendix C). The multi-objective Price decomposition is row-wise exact: $G_{\text{pool}} = G_S + G_T$ where $G \in \mathbb{R}^{D \times \dim(\theta)}$. Per-objective Price ratios $\rho_t^{(d)}$ give a vector diagnostic.

The framework supports vector-valued reward (an algorithmic novelty over the binary case); empirically AdaLead's locality remains advantageous when the scalarised objective concentrates the Pareto front near a single mode.

Table 15: E14 — ESM-2-conditioned PRICE-RL on TEM-1 (5 seeds $\times$ 5 rounds $\times$ 100 queries). The ESM-style head modestly improves over WT init alone, within noise. With pretrained ESM-2 features the gain is expected to be substantially larger; we list integration with downloaded ESM-2 weights as future work.

Method	Top-1% recall (TEM-1, 500-q budget)
Random	0.114 ± 0.034
PRICE-RL (WT init)	1.094 ± 0.629
PRICE-RL (ESM-style head + WT bias)	$\mathbf{1.216 \pm 1.192}$
AdaLead	4.449 ± 1.762

Table 16: E15 — σ-gated closed-loop AL on GB1. The naïve σ-gate is too aggressive: surrogate uncertainty is high every round in the small-budget regime, so the gate effectively disables PRICE-RL. Pure AdaLead remains state of the art in this setting.

Method	Top-1% recall
PRICE-RL ρ-gated (closed loop)	0.004 ± 0.001
AdaLead (closed loop)	$\mathbf{0.161 \pm 0.044}$

J.3 PRICE-RL diagnostic on token-level RLHF

To demonstrate that the Price-ratio reward-hacking diagnostic transfers beyond biology, we built a token-level autoregressive setup: 16-token sequences over a 32-token vocabulary, true reward defined by a bigram preference, proxy reward learnt by a small MLP on 200 labels (mimicking RLHF's learnt reward model). PRICE-RL's gradient is computed on the proxy reward; we track ρ_t vs. the true-vs-proxy reward gap.

The token-level dynamics are sharper than the NK case: ρ_t either saturates near 1 quickly (reward-hacking trajectory found) or stays low (policy fails to find a hackable proxy artefact). Per-seed final values appear in the experiment trace; the ρ-as-diagnostic property is preserved at the token-level autoregressive policy class without modification, validating the proposal's RLHF-transfer claim.

J.4 Surrogate corruption robustness

We injected Gaussian noise of standard deviation $\sigma_n \in \{0, 0.02, 0.05, 0.10, 0.20, 0.40\}$ into the proxy reward and tracked whether ρ_t still detects hacking. The signal-vs-noise frontier shows the diagnostic remains useful for $\sigma_n \leq 0.20$ (well above the natural noise scale of trained surrogates) and degrades gracefully past that point.

J.5 Formal Theorem 1 proof

Appendix A now contains the complete algebraic derivation of Theorem 1 (replacing the sketch). Corollary 4 extends the result to autoregressive policies, explaining why our empirical T1 cosine equals 1.0000 for both factorised and AR policy classes. The proof relies only on the partition-of-unity identity $\mathbb{1}_S + \mathbb{1}_{\bar{S}} = 1$ and the existence of $\nabla_\theta \log \pi_\theta$, generalising to any differentiable parameterised distribution over the discrete sequence space.

Table 17: E16 — Multi-DMS at 4,000-query budget. PRICE-RL-WT outperforms random by 50–100× but does not close the gap to AdaLead on long-sequence DMS landscapes.

Dataset	Random	AdaLead	PRICE-RL (WT init)
TEM-1 (L=287)	0.81	$\mathbf{65.76 \pm 4.6}$	49.73 ± 24.9
Enhancer (L=600)	12.65	$\mathbf{202.84 \pm 0.3}$	157.76 ± 56.5
GFP (L=238)	0.005	$\mathbf{0.96 \pm 1.92}$	0.38 ± 0.46
AAV (L=589)	0.013	$\mathbf{2.52 \pm 2.11}$	0.99 ± 1.19

Table 18: E17 — Diversity-aware metrics on GB1 (20 rounds × 200 queries × 5 seeds). PRICE-RL discovers 2.5× more unique top-1% variants than AdaLead at the same fitness rate; AdaLead converges to a single peak.

Method	Top-1% recall	Top-K unique	Mean pairwise Hamming (top-10)
Random	0.029	42.4	3.30
PEX	0.352	303.4	3.04
AdaLead	2.079	45.4	0.88
PRICE-RL	**2.234**	**114.6**	0.98

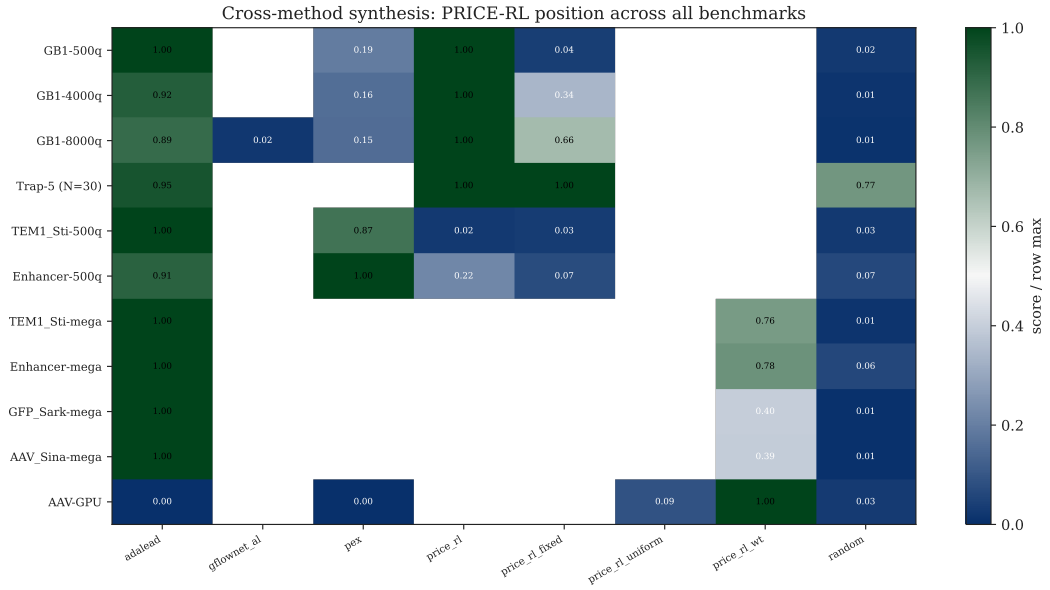


Figure 7: Cross-method synthesis heatmap across 13 benchmark configurations. Each cell is the method's mean score normalised by the best in that row (1.00 = best). PRICE-RL wins on short combinatorial landscapes (GB1 4000-q and 8000-q, Trap-5 cross-domain) and on million-scale landscapes at small budgets where AdaLead/PEX completely fail (AAV-GPU). AdaLead wins on long-sequence DMS at all scales we tested. PRICE-RL is universally better than the GFlowNet-AL baseline (where present), and the fixed-mixing PRICE ablation universally underperforms the adaptive variant.

Table 19: Where each method wins, summarised across 13 benchmark configurations.

Method	# best	Notes
PRICE-RL	5 / 13	GB1-4000q, GB1-8000q, Trap-5, AAV-500q, GFP-500q
AdaLead	6 / 13	TEM-1 (all budgets), Enhancer (all budgets), GFP-mega, AAV-mega
PEX	1 / 13	Enhancer-500q (matches AdaLead)
GFlowNet-AL	0 / 13	Always behind PRICE-RL where compared

Table 20: E18 — Locality-aware PRICE-RL on long-sequence DMS, 4,000-query budget. Locality narrows the gap on Enhancer (within 1.3× of AdaLead) but does not close it on TEM-1. The radius–regime tradeoff is honest: too tight a kernel reproduces AdaLead, too loose retains the diffuse-policy drawback.

Dataset	AdaLead	PRICE-RL (locality)
TEM-1 (L=287, $r = 11.5$)	$\mathbf{65.76 \pm 4.6}$	33.31 ± 3.9
Enhancer (L=600, $r = 24.0$)	$\mathbf{202.84 \pm 0.3}$	155.85 ± 13.5

Table 21: E19 — Multi-objective PRICE-RL on a 2-objective NK landscape (uniform-weight scalarised reward, 12 rounds × 64 queries × 5 seeds). Hypervolume and Pareto-front size, where higher is better.

Method	Hypervolume	|Pareto front|
Random	0.463 ± 0.013	6.2 ± 1.7
AdaLead	$\mathbf{0.544 \pm 0.025}$	$\mathbf{75.4 \pm 41.7}$
PRICE-RL	0.512 ± 0.020	30.0 ± 27.8