ARIADNE: REGULATORY DNA DESIGN WITH ADAPTIVE REWARD SHAPING AND REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cis-regulatory elements (CREs), including enhancers, regulate gene expression through the composition and organization of transcription factor binding sites (TFBSs) within their sequence context. The vast sequence space makes exhaustive experimental screening infeasible, motivating computational methods to design enhancers with high activity in a target cell type. A central challenge is to improve predicted activity while determining whether optimization narrows the generated sequence distribution and whether gains extend beyond the predictor used for training. We present Ariadne, a reinforcement-learning framework that combines an enhancer sequence prior with target-cell-directed optimization to address this challenge. Ariadne first adapts an autoregressive DNA language model to massively parallel reporter assay (MPRA)-derived enhancer sequences through a single shared supervised fine-tuning (SFT) stage and then trains separate target-cell policies with proximal policy optimization (PPO), using the corresponding Enformer prediction as the reward signal. To account for changing reward distributions during optimization, the evaluated reward combines exponential moving average (EMA) normalization, within-batch ranking, and a clipped bonus above an EMA-tracked upper-quantile threshold. Normalization controls favor an evolving reward reference over fixed SFT statistics, but do not establish a universal advantage of EMA over batch normalization. Component ablations support modest incremental Enformer gains from rank, whereas tail has cell-dependent returns that are not uniformly reproduced by Malinois. PPO improves the reported Enformer and Malinois summaries relative to the shared SFT policy, but Hamming diversity decreases from SFT and Enformer-embedding concentration increases. These results support Ariadne as a computational pipeline for improving predicted enhancer activity and characterizing the accompanying distributional trade-offs. They do not establish universal superiority of the complete EMA–rank–tail reward, experimental activity, or cell-type selectivity.

1 INTRODUCTION

Cis-regulatory elements (CREs) control spatiotemporal gene expression through the composition and arrangement of transcription factor binding sites (Shlyueva et al., 2014; Georgakopoulos-Soares et al., 2023). A central goal of regulatory sequence design is to identify CREs with high activity in a chosen cell type. Massively parallel reporter assays (MPRAs) support sequence–activity models (Melnikov et al., 2012; Inoue & Ahituv, 2015), but sample only a minute fraction of the 4^{200} possible 200-bp sequences. This gap motivates computational methods that use measured sequences to guide the search for new candidates for experimental testing in regulatory genomics, synthetic biology, and transgene control.

Sequence-to-function models provide surrogate objectives for this search. Enformer learns associations between sequence and functional measurements (Avsec et al., 2021); coupled to generative models, such predictors guide discrete design. TACO and DREAM-DNA demonstrate RL-based and iterative flow-based regulatory design, respectively (Yang et al., 2025; Wu et al., 2026). Learning an enhancer sequence distribution, however, does not by itself direct generation toward higher predicted activity in a chosen cell type. Proximal policy optimization (PPO) provides a mechanism

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

for this additional optimization by accommodating non-differentiable sequence-level rewards while clipping each policy update (Schulman et al., 2017).

The resulting scientific question is not only whether predictor-guided optimization can increase predicted activity, but also how it changes the generated sequence distribution and whether its gains extend beyond the training predictor. As policies leave the supervised fine-tuning (SFT) distribution, predictor scale may shift and outputs may concentrate despite remaining exactly unique. Uniqueness alone misses near-neighbor sequences and convergence in oracle representation space, while gains under one oracle may not transfer to another predictor or assay. Optimizing a learned predictor therefore does not establish biological function or cell-type selectivity. Addressing this question requires evaluating activity alongside novelty, diversity, representation concentration, and naturalness.

We introduce *Ariadne*, an oracle-guided PPO pipeline for enhancer design that combines a shared enhancer sequence prior with target-cell-directed optimization. The two-stage design uses MPRA-derived sequences to establish the initial generative distribution, then applies predictor feedback to increase predicted activity in each target cell type. Ariadne trains one shared, unconditional SFT model by fine-tuning the 1.2-billion-parameter GENERATOR-V2 model with beginning-of-sequence (BOS)-only generation (Wu et al., 2025; Li et al., 2026). The retained checkpoint initializes separate HepG2, K562, and SK-N-SH PPO policies; cell dependence enters only through their target-specific Enformer rewards.

To account for changing reward distributions during PPO, the reward combines EMA-normalized activity, within-rollout percentile rank, and a bounded bonus above an EMA-tracked upper quantile. EMA supplies a history-aware scale, rank preserves local ordering, and tail emphasizes recent upper-tail candidates. We test their incremental value rather than assuming the complete combination is best. Controlled reward trials use ten 100-step PPO stages without early stopping; Enformer activity selects checkpoints, while diversity is reporting-only.

2 Related Work

2.1 Predictor-Guided Reinforcement Learning

Gradient-based activation maximization can suffer from vanishing gradients, predictor pathologies, and distributional drift, motivating discrete generative optimization (Linder et al., 2020; Linder & Seelig, 2021). TACO applies RL to a pretrained DNA model using min–max-normalized fitness and TFBS-derived rewards (Yang et al., 2025), but its fixed scale does not adapt as the policy-induced score distribution shifts or concentrates near its upper tail.

Activity-conditioned regLM pairs an autoregressive generator with a supervised sequence-to-function model (Lal et al., 2024), while D3 conditions discrete diffusion on desired activity and examines functional and regulatory-composition properties (Sarkar et al., 2026). Both are alternatives to Ariadne's shared-prior PPO design. Iterative model-guided design with MPRA feedback can also improve cell-type-specific enhancers (Yin et al., 2025); Ariadne does not include this prospective wet-lab loop.

DNA-CRAFT combines class-conditioned diffusion with Monte Carlo tree guidance to optimize differential activity between desired and undesired cells (Awasthi et al., 2026), unlike Ariadne's single-target objective without an off-target penalty. Active learning instead alternates measurement, predictor fitting, and sequence selection rather than optimizing a frozen oracle (Shen et al., 2025). Reviews and reasoning-informed predictors provide broader context (De Winter et al., 2025; Duan et al., 2026).

2.2 Conditional Diffusion and Experimental Design

DNA-Diffusion uses a cell-conditioned diffusion model and classifier-free guidance to balance activity, specificity, and diversity, with prospective validation (DaSilva et al., 2026). Ariadne instead adapts one unconditional autoregressive prior to frozen, potentially non-differentiable predictors through PPO without retraining a conditional generator. This modularity is useful when targets or scoring functions change, but it is not evidence of empirical superiority. DNA-Diffusion pro-

vides stronger direct evidence for specificity and biological activity, and the methods have not been matched on data, oracle, or budget.

2.3 Flow Matching and Iterative Refinement

DREAM-DNA combines discrete flow matching, iterative masking and resampling, and reward backpropagation through relaxed trajectories (Wu et al., 2026). Ariadne instead reuses a left-to-right language-model policy with a black-box terminal reward, avoiding differentiability but forgoing bidirectional revision. This supports arbitrary frozen scorers and existing autoregressive checkpoints, whereas DREAM-DNA can revise earlier positions within a rollout. Different generators, search dynamics, oracles, and protocols preclude a head-to-head claim.

Relative to conditional or refinement-based generators, Ariadne focuses on reward construction within fixed-budget PPO: a running activity reference tracks recent performance, rollout-relative ordering compares current candidates, and bounded upper-tail emphasis prioritizes high-score improvement. We ablate these complementary components under identical initialization and training budgets, using activity-selected continuation without diversity-based early stopping. Their empirical value is tested rather than assumed from their distinct roles.

3 Methods

Ariadne adapts an autoregressive DNA model to enhancer sequences, then optimizes one policy per target cell using stagewise PPO (Figure 1).

3.1 Problem Formulation

Let $\mathcal{X} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^L$, with $L = 200$. A sequence is represented by non-overlapping 6-mer tokens $z_{1:T}$ and decoded to DNA by D_L. For target cell c, BOS-only generation and optimization are defined as

$$\begin{aligned} \pi_{\theta_c}(z_{1:T}) &= \prod_{t=1}^{T} \pi_{\theta_c}(z_t \mid \langle \mathrm{BOS} \rangle, z_{<t}), \\ \theta_c^* &= \arg\max_{\theta_c} \mathbb{E}_{z \sim \pi_{\theta_c}} \left[f_c(D_L(z)) \right], \end{aligned} \tag{1}$$

where f_c is a frozen Enformer-based target-cell MPRA regressor (Avsec et al., 2021). The shared SFT checkpoint contains no cell condition and initializes every target; cell dependence enters only through f_c during PPO, without an off-target penalty. This is the design objective; PPO optimizes a shaped surrogate with adaptive KL regularization, without a guarantee of the same optimum.

3.2 Shared Unconditional Supervised Fine-Tuning

We fine-tune the 1.2-billion-parameter GENERator-v2 model once on the shared MPRA-derived sequence corpus (Li et al., 2026), using the nucleotide-marginal objective in Equation equation 11 (Appendix C). SFT uses no cell or activity condition. One checkpoint with minimum validation loss among saved checkpoints initializes every PPO run; the exact vocabulary mapping and optimization settings are given in the appendix.

3.3 History-Aware Stagewise PPO

Ariadne combines exponential moving average (EMA) normalization, rank, and tail in the ERT (EMA+rank+tail) reward. These components address an evolving score distribution, relative quality within each rollout, and additional emphasis on candidates exceeding a recent high-score threshold, respectively. For candidate i in rollout b, let $a_{b,i} = f_c(x_{b,i})$. Using the preceding-batch EMA mean μ_{b-1}, standard deviation σ_{b-1}, EMA 90th percentile q_{b-1}, and within-batch rank $k_{b,i}$ among N candidates, Ariadne assigns

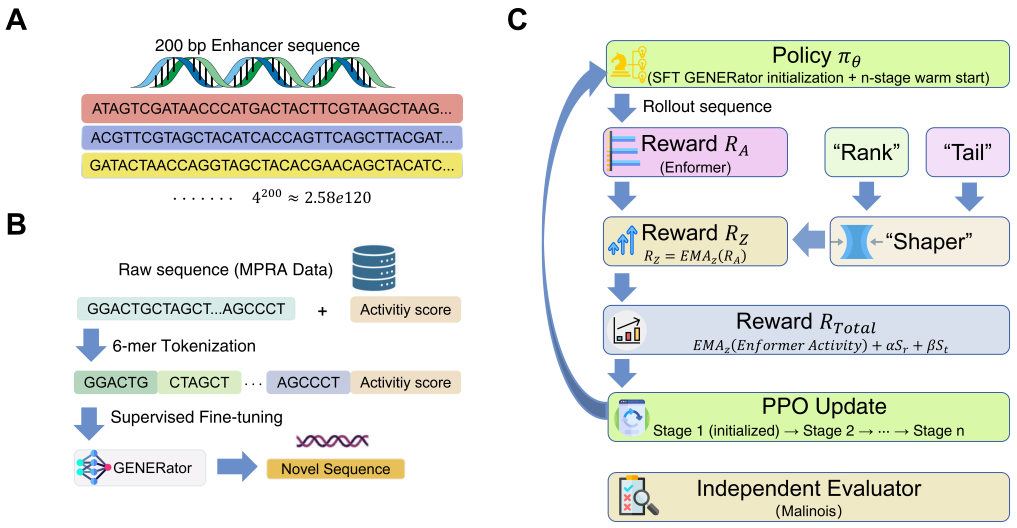


Figure 1: **Ariadne overview.** One shared, unconditional SFT model initializes target-cell Enformer-guided PPO; Malinois is evaluation only.

$$R_{b,i} = \text{clip}\left(\frac{a_{b,i} - \mu_{b-1}}{\sigma_{b-1} + \epsilon}, -5, 5\right) + 0.3\left(2\frac{k_{b,i} - 1}{N - 1} - 1\right) + 0.6\,\text{clip}\left(\frac{a_{b,i} - q_{b-1}}{\sigma_{b-1} + \epsilon}, 0, 3\right). \tag{2}$$

EMA: a running reference for activity. As PPO changes the policy, the location and spread of predicted scores can change across rollouts. A fixed reward scale does not track these changes. EMA normalization uses recent performance within each 100-step stage, with its state reset at stage boundaries, retaining standardized score differences within the unclipped range while adapting the reference mean and scale. Clipping bounds the resulting activity term. This motivates EMA as the base component, although normalization alone does not guarantee stable PPO updates or prevent distributional concentration.

Rank: relative quality within the current rollout. A historical reference may lag the current score distribution, and nearby scores can receive similar EMA rewards. The rank term adds a bounded ordering signal whose spacing does not depend on the numerical gaps between scores. It therefore complements EMA's magnitude-sensitive signal with relative quality within the current rollout. Rank alone, however, assigns the same reward to the highest-ranked candidate in a weak or a strong rollout and cannot quantify how much one candidate improves on another. Combining it with EMA retains a reference to recent performance as well as local ordering.

Tail: additional emphasis above a recent elite threshold. EMA already rewards larger standardized scores, but does not assign extra weight specifically to exceeding a high-score threshold. The tail term adds this preference: it is zero below the EMA-tracked 90th percentile and increases with exceedance above it, up to a fixed cap. Where both magnitude-based terms are unclipped, this raises the reward slope from $1/(\sigma_{b-1} + \epsilon)$ to $1.6/(\sigma_{b-1} + \epsilon)$ above the threshold, with the discrete rank contribution considered separately. Tail thus reweights upper-tail improvement rather than introducing magnitude information absent from EMA. Its historical threshold also distinguishes current batch rank from exceedance of recent high scores. The cap limits the additional reward from extreme predictions, without guaranteeing protection against predictor exploitation.

Together, EMA supplies a running reference, rank supplies local ordering, and tail emphasizes upper-tail improvement. Tail reflects the high-score focus of checkpoint selection without making

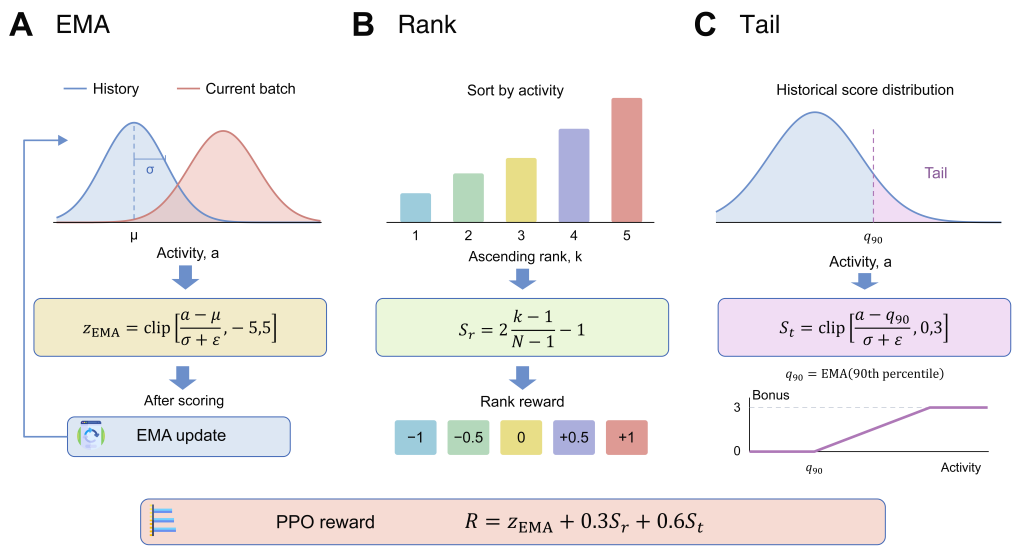


Figure 2: **Reward shaping.** **A**, EMA normalization; **B**, Within-batch rank; **C**, Bounded upper-tail bonus (schematic).

the two criteria equivalent. Their empirical contributions are tested in Section 4.3.1; Appendix D details EMA state and edge cases.

PPO uses Equation equation 2 as a terminal reward with token-level KL regularization (Schulman et al., 2017). Each stage retains the checkpoint balancing *Fitness* and *Top*:

$$\pi_s^* = \arg\max_{\pi \in \mathcal{C}_s} [0.7\ \mathrm{Fitness}(\pi) + 0.3\ \mathrm{Top}(\pi)], \qquad \pi_{s+1,0} = \pi_{s+1}^{\mathrm{ref}} = \pi_s^*. \tag{3}$$

Fitness is the median of the top 128 normalized Enformer scores and Top is the top-16 mean; shaping uses raw scores. Controlled reward trials complete ten 100-step stages, and diversity never controls training. Full PPO settings are in Appendix D.

3.4 Generation and Evaluation

Retained policies generate 42 six-mer tokens from a BOS-only prompt and retain 200 valid bases after decoding. Enformer defines the objective; Malinois and all sequence- and representation-level metrics are post hoc (Gosai et al., 2024). Decoding edge cases and evaluation details are in Appendices D and F.

4 Experiments

We evaluate predicted activity and distributional concentration in HepG2, K562, and SK-N-SH, followed by reward and backbone ablations. E and M denote Enformer and Malinois.

4.1 Experimental Setup

Dataset. We use 763,684 MPRA-derived 200-bp sequences with measurements for all three cell types. Chromosome 21 is reserved for validation, chromosomes 7, 13, and 22 for testing, and the remainder for training, yielding 669,233/10,583/83,868 train/validation/test sequences. All experiments share these chromosome-disjoint splits.

Shared SFT and PPO. The shared GENERATOR-v2 SFT checkpoint initializes all three target policies. Each uses its target-cell Enformer and ERT reward, with five training seeds. Controlled reward trials use ten 100-step stages without early stopping. Tokenization, sampling settings, and seed identifiers are given in Appendices C and D.

Table 1: **Before SFT, after shared SFT, and after target-cell PPO.** Entries are mean (sample SD); SFT rows use 59 generation seeds and Ariadne rows use five PPO training seeds.

Cell	Policy	E Top	E Fitness	M Top	Diversity
HepG2	Before SFT	0.55 (0.02)	0.40 (0.00)	0.56 (0.03)	112.05 (3.28)
	After SFT	0.66 (0.01)	0.50 (0.02)	0.67 (0.03)	**145.97 (3.07)**
	Ariadne	**0.93 (0.03)**	**0.92 (0.03)**	**0.98 (0.06)**	138.34 (3.41)
K562	Before SFT	0.68 (0.03)	0.40 (0.00)	0.57 (0.04)	112.05 (3.80)
	After SFT	0.81 (0.01)	0.53 (0.03)	0.66 (0.02)	**146.47 (2.55)**
	Ariadne	**0.99 (0.01)**	**0.97 (0.01)**	**0.89 (0.04)**	132.68 (1.34)
SK-N-SH	Before SFT	0.54 (0.02)	0.41 (0.00)	0.49 (0.03)	122.65 (6.51)
	After SFT	0.71 (0.02)	0.52 (0.02)	0.60 (0.03)	**142.36 (3.71)**
	Ariadne	**0.94 (0.02)**	**0.92 (0.02)**	**1.09 (0.05)**	138.48 (5.47)

Evaluation protocol and metrics. For each policy, 59 generation seeds produce 256 sequences each, measuring within-policy Monte Carlo variation rather than independent training replication. We average seed-level summaries within each policy, then report mean (sample SD) across five training seeds, rounded to two decimals.

Scores use fixed, unclipped dataset ranges. *Top* is the top-16 mean, *Fitness* the top-128 median, and *Diversity* their median pairwise Hamming distance. E and M independently rank the same pool, so their Top summaries may select different sequences. Malinois is a secondary surrogate (Gosai et al., 2024). Appendix F defines normalization, embedding similarity, motif, novelty, uniqueness, naturalness, and the post hoc target–off-target diagnostic. These measures probe sequence-space and representation-space concentration, but none enters reward or checkpoint selection.

4.2 Main Results

Table 1 reports the shared model before and after enhancer SFT, scored separately by each target-cell predictor, together with the Ariadne policies represented in the supplied figures. Each SFT row describes the same fixed checkpoint over 59 generation seeds, whereas Ariadne variation is across five independently trained PPO policies after within-policy 59-seed aggregation; the two types of uncertainty are therefore descriptive and should not be treated as a common statistical test. Aggregation and selection limitations are given in Appendix I.2.

Shared SFT improves predicted activity across cells, and target-cell PPO further increases both Enformer and Malinois summaries (Table 1). PPO also lowers Hamming diversity relative to SFT, revealing an activity–diversity trade-off; embedding concentration also increases (Appendix I.1). Agreement between predictors provides computational support but does not establish experimental activity.

Figure 3 shows that the Malinois-to-Enformer Top ratio increases after PPO in all three cells, with the largest shift in SK-N-SH. Because each predictor independently selects its top 16 candidates, this ratio compares summary scales rather than scoring one fixed selected set. The SFT points also describe one shared checkpoint, whereas the PPO points pool generation seeds from five trained policies; the plotted points therefore are not independent training replicates.

The before/after-SFT cross-Enformer analysis likewise shows that E Top rises under each target predictor after shared SFT, although target-minus-off-target gaps are not uniformly positive at either SFT stage. After target-specific PPO, all 30 training-run/off-target mean gaps are positive (Appendix I). Because the three evaluators belong to the same predictor family, these gaps remain computational diagnostics rather than biological validation of selectivity.

Published baselines provide external context but differ in initialization, budget, and protocol; Table 5 in Appendix E reports the numerical comparison.

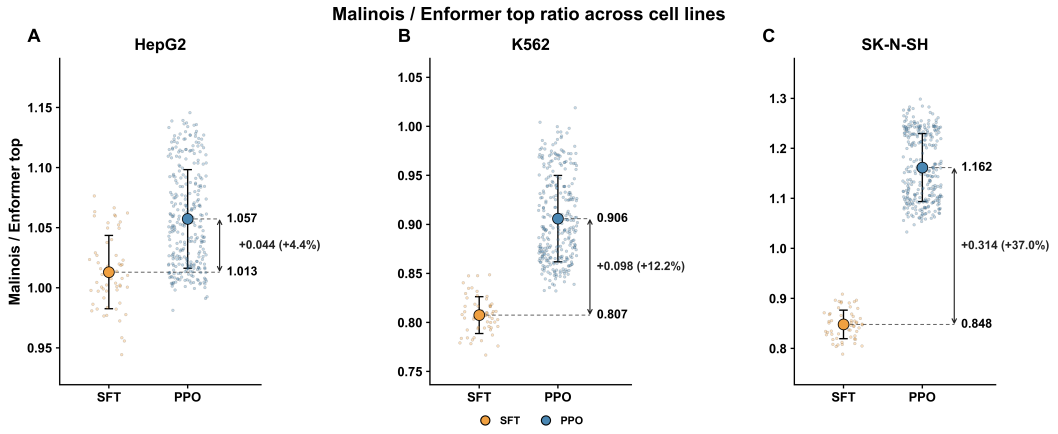


Figure 3: **Cross-predictor Top ratio.** Per-generation-seed Malinois Top divided by Enformer Top after shared SFT and target-cell PPO; large markers and bars show mean and sample SD after 1.5×IQR filtering.

Table 2: **Reward ablation.** Five-seed mean (sample SD); bold marks the best displayed mean per cell and metric.

Cell	Reward	E Top	E Fitness	M Top	Diversity
HepG2	EMA only	0.90 (0.02)	0.88 (0.02)	0.97 (0.06)	139.70 (3.00)
	EMA+rank	0.91 (0.03)	0.89 (0.03)	**0.99** (**0.03**)	139.80 (2.10)
	ERT	**0.93** (**0.03**)	**0.92** (**0.03**)	0.98 (0.06)	138.34 (3.41)
	Raw Enformer	0.81 (0.02)	0.79 (0.02)	0.93 (0.02)	**145.40** (**1.20**)
K562	EMA only	0.98 (0.00)	0.96 (0.00)	**0.93** (**0.02**)	135.60 (1.20)
	EMA+rank	**0.99** (**0.01**)	**0.97** (**0.00**)	0.92 (0.03)	132.80 (3.20)
	ERT	**0.99** (**0.01**)	**0.97** (**0.01**)	0.89 (0.04)	132.68 (1.34)
	Raw Enformer	0.95 (0.03)	0.93 (0.03)	0.89 (0.03)	**140.00** (**2.10**)
SK-N-SH	EMA only	0.92 (0.02)	0.90 (0.01)	**1.12** (**0.03**)	137.60 (3.10)
	EMA+rank	**0.95** (**0.02**)	**0.93** (**0.02**)	1.08 (0.05)	138.60 (2.10)
	ERT	0.94 (0.02)	0.92 (0.02)	1.09 (0.05)	138.48 (5.47)
	Raw Enformer	0.89 (0.01)	0.86 (0.01)	1.08 (0.01)	**149.00** (**0.00**)

4.3 Ablation Studies

4.3.1 EMA, Rank, and Tail Rewards

The ablation examines normalization, local ordering, and upper-tail emphasis. Seven shaped rewards, raw Enformer, fixed-z, and batch-z use the shared SFT initializer and five seed identifiers per cell. Table 2 summarizes the main configurations; Appendix I.6 adds normalization controls and paired rank increments. ERT summaries throughout use the same aggregation as the figures; selection limitations are stated in Appendix I.2.

EMA is the strongest single shaped component and improves Enformer Fitness over raw reward. Both EMA-only and batch-z outperform fixed-z across cells, supporting an evolving reference; EMA-only exceeds batch-z only in K562. Thus, the controls support adaptation but do not establish that historical smoothing is indispensable (Appendix I.6).

Adding rank yields modest Enformer gains in 4/5–5/5 seed pairs per cell, consistent with the value of local ordering. Tail-containing ERT has higher HepG2 summaries than EMA+rank, but comparable K562 and slightly lower SK-N-SH results. Malinois does not consistently reproduce these gains. These descriptive comparisons motivate rank as a useful addition and tail as a conditional upper-tail preference; they establish neither universal component necessity nor diversity preservation.

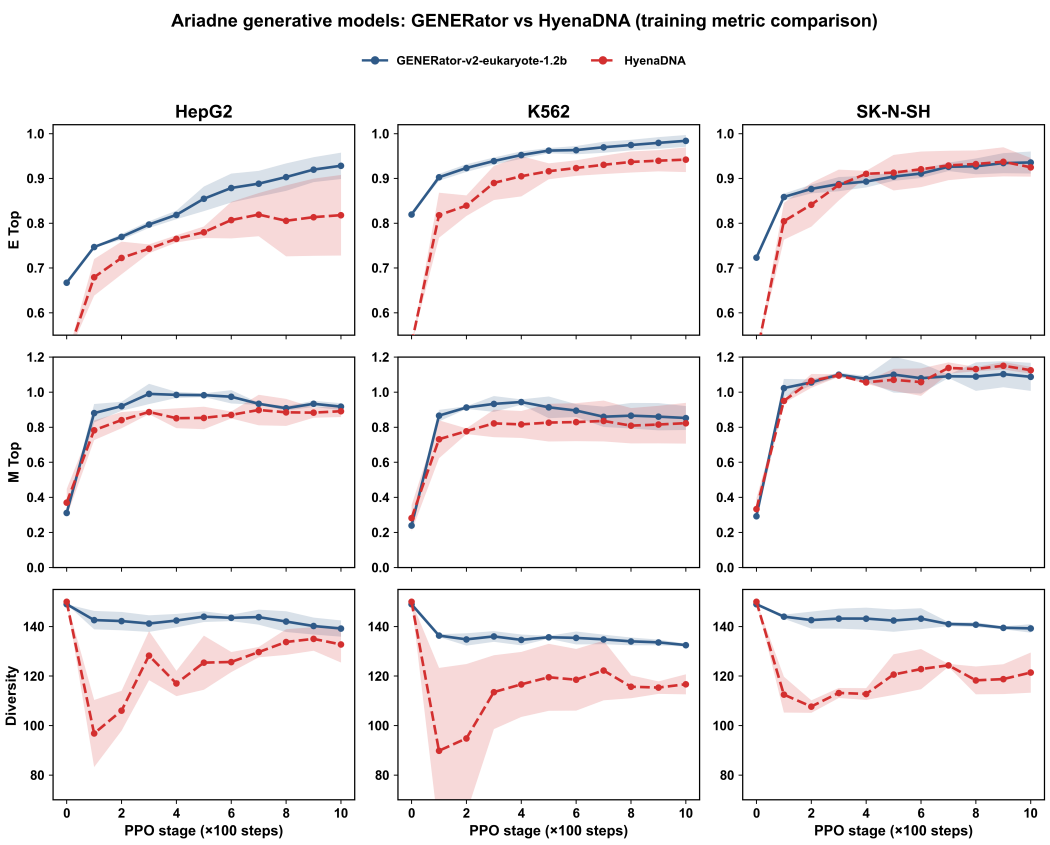


Figure 4: **Backbone training dynamics.** Rollout E Top, M Top, and diversity across five training seeds; curves show stagewise means with sample-SD bands after 1.5×IQR filtering. Stage 0 is the first pre-update evaluation; stages 1–10 contain 100 PPO updates each.

4.3.2 GENERator-v2 and HyenaDNA Configurations

With shared data and ERT shaping, both configurations improve rollout Enformer Top (Figure 4). On held-out evaluation, GENERator-v2 achieves higher activity in HepG2 and K562 and greater diversity across cells; HyenaDNA (Nguyen et al., 2023) is slightly higher on SK-N-SH Enformer activity, with Malinois Top tied after rounding (Tables 17–19). Because backbone, tokenization, and action horizon change together, this comparison remains descriptive; full settings are in Appendix H.

5 Limitations

All activities are computational predictions. Enformer supplies the reward, checkpoint selection, and primary evaluation; Malinois is a related secondary surrogate that selects its own top candidates. Ariadne also lacks an off-target objective, so its outputs are target-optimized rather than experimentally validated or biologically cell-type-selective.

Reward shaping changes its balance with adaptive KL regularization and the value loss. Fixed-z and batch-z narrow the normalization question, but matched controller settings do not imply matched realized KL or reward scale. The components' distinct roles do not establish that each is indispensable: tail is not uniformly beneficial, and higher activity accompanies lower diversity and greater embedding concentration. Five-seed comparisons remain descriptive; aggregation limitations are detailed in Appendix I.2.

External baselines use different protocols, while HyenaDNA also changes tokenization and action horizon. Fixed-set cross-predictor evaluation, scale-matched controls, explicit off-target optimiza-

tion, and prospective MPRA validation remain necessary; detailed statistical and protocol limitations are given in Appendix F.

6 CONCLUSION

Ariadne combines one shared, unconditional GENERATOR-V2 enhancer prior with target-cell Enformer-guided PPO. Its reward combines three complementary signals: EMA normalization measures activity relative to recent performance, rank expresses relative quality within a rollout, and tail adds bounded emphasis on improvement above a recent elite threshold. Under identical SFT initialization, normalization controls favor an evolving reference over fixed SFT statistics, and rank yields modest incremental Enformer gains. Tail improves the HepG2 Enformer summaries but has cell-dependent returns, supporting its role as an explicit upper-tail preference rather than a universally necessary component. Compared with its protocol-matched raw-reward control, Ariadne improves predicted activity but sacrifices Hamming diversity.

Relative to shared SFT, PPO increases predicted activity but reduces diversity and increases embedding concentration. Ariadne thus prioritizes candidates under an activity–diversity trade-off; neither universal superiority of the ERT reward nor biological activity or selectivity is established.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, English–Chinese translation, language and LaTeX editing, consistency checks between the manuscript and the author-provided repository, and review of methodology and experimental reporting. They also provided feedback on ablation and evaluation design and helped summarize and interpret author-provided experimental aggregates when revising tables and discussion. They were not used to generate measurements reported as experimental results, to run the reported experiments, or to alter the scientific content of the supplied figures. The authors manually verified all AI-assisted text, equations, tables, citations, code-related descriptions, and numerical transcriptions against the repository and source results. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This study performs computational analysis of previously released sequence–activity data and does not collect identifiable personal data or conduct new human-participant or wet-lab experiments. The generated regulatory sequences are unvalidated computational candidates and are not presented as ready for synthesis, delivery, or clinical use. Regulatory-sequence design may carry biosafety and dual-use risks; any downstream synthesis should therefore undergo commercial DNA-synthesis screening, applicable institutional biosafety review, and experiment-specific risk assessment. Release of complete high-scoring candidate sets should also be evaluated under an appropriate access policy. Enformer and Malinois are surrogate predictors, and prospective experiments would require separate assessment of off-target activity, genomic-integration risk, and the relevant containment and oversight measures.

REPRODUCIBILITY STATEMENT

The appendices and the anonymized supplementary archive provide materials supporting reproduction of the reported computational experiments. The archive contains source code for data preparation, SFT and PPO training, evaluation, and figure generation; executable pipeline scripts; unit tests; dependency specifications; and a machine-readable JSON manifest recording the principal hyperparameters, dataset splits, model and checkpoint identifiers, and software environment. The accompanying README documents the environment setup, principal commands, expected directory structure, and required external assets. Owing to file-size and redistribution constraints, the training data, pretrained model weights, trained checkpoints, and experiment logs are not included; their required versions and acquisition information are documented in the archive.

REFERENCES

Anne Auger and Nikolaus Hansen. Tutorial CMA-ES: Evolution strategies and covariance matrix adaptation. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, pp. 827–848. ACM, 2012. doi: 10.1145/2330784.2330919. URL https://doi.org/10.1145/2330784.2330919.

Ziga Avsec, Vikram Agarwal, Daniel Visentin, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Animesh Awasthi, Raphael Bednarsky, Moritz Schaefer, and Christoph Bock. Conditional monte carlo tree diffusion for designing cell-type-specific and biologically faithful regulatory DNA. In *MLGenX Workshop at ICLR*, 2026. URL https://openreview.net/forum?id=4DyPOqo2hs.

Xingyu Chen, Shihao Ma, Runsheng Lin, Jiecong Lin, and Bo Wang. Ctrl-DNA: Controllable cell-type-specific regulatory DNA design via constrained RL. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/14c018d2e72c521605b0567029ef0efb-Abstract-Conference.html.

Lucas Ferreira DaSilva, Simon Senan, Judith F. Kribelbauer-Swietek, et al. Designing synthetic regulatory elements using the generative AI framework DNA-Diffusion. *Nature Genetics*, 58:180–194, 2026. doi: 10.1038/s41588-025-02441-6. URL https://doi.org/10.1038/s41588-025-02441-6.

Seppe De Winter, Vasileios Konstantakos, and Stein Aerts. Modelling and design of transcriptional enhancers. *Nature Reviews Bioengineering*, 3:374–389, 2025. doi: 10.1038/s44222-025-00280-y.

Yi Duan, Zhao Yang, Jiwei Zhu, Ying Ba, Chuan Cao, and Bing Su. Biological reasoning-informed regression for interpretable regulatory DNA activity prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3818836.

Ilias Georgakopoulos-Soares, Chengyu Deng, Vikram Agarwal, Candace S. Y. Chan, Jingjing Zhao, Fumitaka Inoue, and Nadav Ahituv. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature Communications*, 14:2333, 2023. doi: 10.1038/s41467-023-37960-5.

Sager J. Gosai, Rodrigo I. Castro, Natalia Fuentes, John C. Butts, Kousuke Mouri, Michael Alasoadura, Susan Kales, Thanh Thanh L. Nguyen, Ramil R. Noche, Arya S. Rao, Mary T. Joy, Pardis C. Sabeti, Steven K. Reilly, and Ryan Tewhey. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 634:1211–1220, 2024. doi: 10.1038/s41586-024-08070-z.

Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011. doi: 10.1093/bioinformatics/btr064. URL https://doi.org/10.1093/bioinformatics/btr064.

Fumitaka Inoue and Nadav Ahituv. Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164, 2015. doi: 10.1016/j.ygeno.2015.06.005.

Avantika Lal, David Garfield, Tommaso Biancalani, and Gokcen Eraslan. Designing realistic regulatory DNA with autoregressive language models. *Genome Research*, 34(9):1411–1420, 2024. doi: 10.1101/gr.279142.124.

Qiuyi Li, Zhihao Zhan, Shikun Feng, Yiheng Zhu, Yuan He, Wei Wu, Zhenghang Shi, Shengjie Wang, Zongyong Hu, Zhao Yang, Jiaoyang Li, Jian Tang, Haiguang Liu, and Tao Qin. Generator-v2: Reconciling coarse tokenization with single-nucleotide resolution in genomic language modeling. *bioRxiv*, 2026. doi: 10.64898/2026.01.27.702015. URL https://www.biorxiv.org/content/early/2026/05/04/2026.01.27.702015.

Johannes Linder and Georg Seelig. Fast activation maximization for molecular sequence design. *BMC Bioinformatics*, 22:510, 2021. doi: 10.1186/s12859-021-04437-5.

Johannes Linder, Nicholas Bogard, Alexander B. Rosenberg, and Georg Seelig. A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences. *Cell Systems*, 11(1):49–62.e16, 2020. doi: 10.1016/j.cels.2020.05.007. URL https://doi.org/10.1016/j.cels.2020.05.007.

Mingqian Ma, Wanjuan Bu, Guoqing Liu, Yuxuan Liu, Sizhen Liu, Zhen Zhao, Shijie Yao, Qingru Hua, Yujie Zhang, Cuiting Zhong, Haitao Huang, Pan Deng, Peiran Jin, Qijin Yin, Chuan Cao, Haiguang Liu, Mo Xu, Yuan He, Tao Qin, and Zeyu Chen. Reconstructing sequence-grammar trajectories enables interpretable and tunable cis-regulatory element design. *bioRxiv*, 2026. doi: 10.64898/2026.07.10.737719. URL https://doi.org/10.64898/2026.07.10.737719.

Alexey Melnikov, Ashwin Murugan, Xinyi Zhang, Tiberiu Tesileanu, Lihua Wang, Petar Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, and Justin B. Kinney. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, 2012. doi: 10.1038/nbt.2137.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1872. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract.html.

Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A. Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 2024. doi: 10.1093/nar/gkad1059. URL https://doi.org/10.1093/nar/gkad1059.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18520–18536. PMLR, 2022. URL https://proceedings.mlr.press/v162/ren22a.html.

Anirban Sarkar, Alejandra Duran, Yiyang Yu, Da-Wei Lin, Yijie Kang, Nirali Somia, Pablo Mantilla, Jessica Zhou, Masayuki Nagai, Ziqi Tang, Kaarina Hanington, Kenneth Chang, and Peter K. Koo. Designing dna with tunable regulatory activity using discrete diffusion. *bioRxiv*, 2026. doi: 10.1101/2024.05.23.595630. URL https://www.biorxiv.org/content/early/2026/04/27/2024.05.23.595630.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yuxin Shen, Grzegorz Kudla, and Diego A. Oyarzún. Optimization of regulatory DNA with active learning. *Computational and Structural Biotechnology Journal*, 27:4384–4392, 2025. doi: 10.1016/j.csbj.2025.09.033.

Darya Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014. doi: 10.1038/nrg3682.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D. Kelsic. AdaLead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020. doi: 10.48550/arXiv.2010.02141. URL https://arxiv.org/abs/2010.02141.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2951–2959, 2012. URL https://papers.nips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Wei Wu, Qiuyi Li, Yuanyuan Zhang, Zhihao Zhan, Ruipu Chen, Mingyang Li, Kun Fu, Junyan Qi, Yongzhou Bao, Chao Wang, et al. Generator: a long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025.

Weimin Wu, Jiunhau Chen, Xuefeng Song, Jerry Yao-Chieh Hu, and Han Liu. DREAM-DNA: Controlled design via reasoning and matched-flows for DNA. In *Gen2 Workshop at ICLR*, 2026. URL https://openreview.net/pdf?id=Dk2LBHsbrn.

Zhao Yang, Bing Su, Chuan Cao, and Ji-Rong Wen. Regulatory DNA sequence design with reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://mlanthology.org/iclr/2025/yang2025iclr-regulatory/.

Christopher Yin, Sebastian M. Castillo-Hair, Gun Woo Byeon, Peter Bromley, Wouter Meuleman, and Georg Seelig. Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. *Cell Systems*, 16(7):101302, 2025. doi: 10.1016/j.cels.2025.101302.

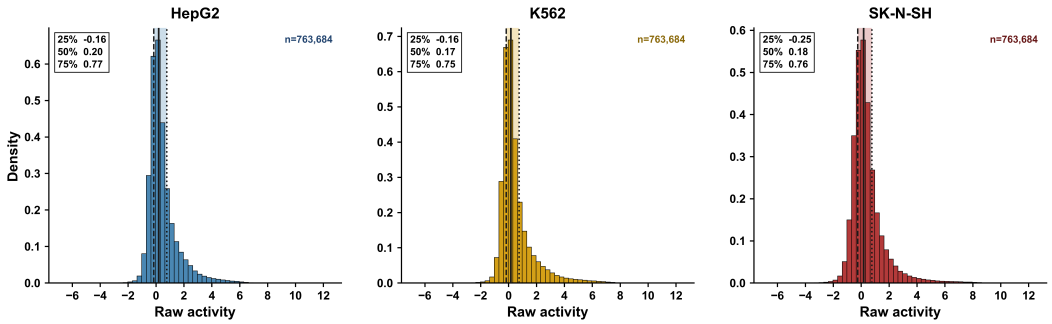


Figure 5: Distributions of experimentally measured MPRA activity across the 763,684 records retained in the Ariadne derivative. The vertical lines denote the 25th, 50th, and 75th percentiles.

A Dataset Details

A.1 MPRA regulatory-sequence dataset

Source and composition. The experimentally measured regulatory sequences used by Ariadne were derived from the MPRA collection assembled for training Malinois (Gosai et al., 2024). The source study aggregated 798,064 unique oligonucleotides from ten MPRA experiments in three projects: UK Biobank (OL27–OL33), GTEx (OL41–OL42), and OL15. Of these, 783,978 oligonucleotides were designed to assay human genetic variation, including 706,054 reference–alternative allele sequences and 77,924 sequences representing pairwise combinations of two variants. The remaining 14,086 sequences originated from DNase-I-hypersensitive or H3K27ac-marked regions.

The Malinois preprocessing removed oligonucleotides with fewer than 20 plasmid reads, with no RNA reads in any assayed cell type, non-natural OL15 sequences, and a small number of extreme low-activity observations. Duplicated sequences were resolved using the project hierarchy described by Gosai et al. (2024). The resulting Malinois training collection contains 776,474 unique 200-nt sequences with MPRA $\log_2$ fold-change measurements in K562, HepG2, and SK-N-SH.

Ariadne derivative. Ariadne uses a coordinate-annotated derivative table. After Ariadne-specific filtering, it contains 763,684 200-bp records. The official pipeline uses genomic coordinates, `Sequence`, and the three cell-line activity columns `K562_mean`, `HepG2_mean`, and `SKNSH_mean`.

The activity distributions in Figure 5 are included only to describe the retained data.

A.2 Transcription-factor motif assets

The motif library is a *Homo sapiens* JASPAR 2024 CORE MEME export (Rauluseviciute et al., 2024). The local MEME file contains 755 motif matrices. Ariadne reuses the Ctrl-DNA selection table `selected_ppms.csv`, which contains 464 selected motif identifiers, and the cell-specific reference tables `{cell}_tfbs_freq_all.csv` (Chen et al., 2025). Each reference table contains one sequence-identifier column and 464 motif-count columns.

For a generated sequence, Ariadne scans the selected motifs with FIMO at $P < 10^{-3}$ (Grant et al., 2011). Let $\mathbf{m}(x) \in \mathbb{R}^{464}$ be the per-sequence motif hit-count vector and let

$$\mathbf{m}_{\text{ref}}^{(c)} = \sum_{r \in \mathcal{R}_c} \mathbf{m}(r)$$

be the column-sum vector from the Ctrl-DNA reference table for cell line c. The reported motif statistic is

$$\text{MotifCorr}_c = \frac{1}{|\mathcal{G}|} \sum_{x \in \mathcal{G}} \text{Pearson}\left(\mathbf{m}(x), \mathbf{m}_{\text{ref}}^{(c)}\right). \tag{4}$$

Rows with a constant motif-count vector contribute zero under the implementation. Motif correlation is a reporting metric only; it does not contribute to the official SFT loss, PPO reward, or best-policy score.

B Data Preparation

Chromosome labels are normalized to the `chrN` form. Chromosome 21 forms the validation split, chromosomes 7, 13, and 22 form the test split, and all remaining chromosomes form the training split. The retained counts are 669,233 training records, 10,583 validation records, and 83,868 test records, totaling 763,684 records. This split is Ariadne-specific and should not be conflated with the validation/test chromosomes used to train the original Malinois model.

For each target-cell data export, records with non-finite target activity values are removed. Sequences are uppercased, must match `[ACGTN]+`, and are right-trimmed to a length divisible by six:

$$L_i' = 6 \left\lfloor \frac{L_i}{6} \right\rfloor, \tag{5}$$

$$x_i' = x_i[1 : L_i']. \tag{6}$$

Thus, each original 200-bp sequence contributes 198 bp, or 33 non-overlapping 6-mer tokens, to GENERator-v2 SFT. The official launcher reads only `trimmed_sequence`.

C Shared Unconditional Supervised Fine-Tuning

C.1 Training protocol

The official pipeline trains a single shared, unconditional SFT model with `--no-conditioned-input`. Consequently, neither cell identity nor an activity token is present in the SFT input. All 763,684 records share the same DNA sequence corpus, and the single retained SFT checkpoint initializes the PPO runs for all three targets. Any target-named artifact paths used by the downstream launcher point to identical retained model weights, not separate SFT models for different cells. Target-specific supervision is introduced only by the cell-specific Enformer predictor during PPO.

The tokenizer prepends one beginning-of-sequence token and partitions the 198-bp training string into 33 non-overlapping 6-mers. Before padding, record i is represented as

$$[\langle \mathrm{BOS} \rangle, z_{i,1}, \ldots, z_{i,33}] .$$

The model factorization is therefore

$$p_\theta(x_i') = \prod_{t=1}^{33} p_\theta \left(z_{i,t} \mid \langle \mathrm{BOS} \rangle, z_{i,<t} \right), \tag{7}$$

with no activity condition in the autoregressive context.

C.2 Nucleotide-marginal objective

Let $\mathcal{V}$ be the tokenizer vocabulary and $\mathcal{V}_{\mathrm{DNA}} \subset \mathcal{V}$ the valid A/C/G/T 6-mer tokens. At causal context $h_{i,t}$, the raw token distribution is

$$p_\theta(v \mid h_{i,t}) = \frac{\exp(\ell_{i,t,v})}{\sum_{u \in \mathcal{V}} \exp(\ell_{i,t,u})}. \tag{8}$$

For a DNA token v, let $g_j(v)$ be the nucleotide at within-token position $j \in \{1, \ldots, 6\}$. The implementation initializes the vocabulary-wide nucleotide lookup to zero, which corresponds to adenine, and overwrites the lookup only for valid DNA tokens. Hence

$$\widetilde{g}_j(v) = \begin{cases} g_j(v), & v \in \mathcal{V}_{\mathrm{DNA}}, \\ \mathrm{A}, & v \notin \mathcal{V}_{\mathrm{DNA}}, \end{cases} \tag{9}$$

Table 3: Shared unconditional GENERATOR-V2 SFT settings.

Hyperparameter	Value
Backbone	GENERATOR-V2 eukaryote 1.2B
Input	BOS + 198-bp DNA; no condition token
Tokenizer	Non-overlapping 6-mer tokenizer
Training / validation batch size	64 / 16
Gradient accumulation	1
Optimizer	AdamW
Peak learning rate	2×10^{-5}
Weight decay	10^{-2}
Learning-rate schedule	2,000-step warm-up + cosine
Minimum learning-rate ratio	0.1
Maximum epochs	2
Gradient clipping norm	1.0
Evaluation interval	500 steps
Checkpoint interval	5,000 steps
Maximum retained checkpoints	4
Model-selection criterion	Minimum validation loss
Attention / precision	SDPA / bfloat16
Gradient checkpointing	Enabled

and the implemented nucleotide marginal is

$$p_{\theta,j}(b \mid h_{i,t}) = \sum_{\substack{v \in \mathcal{V}: \\ \widetilde{g}_j(v)=b}} p_\theta(v \mid h_{i,t}). \tag{10}$$

This means that probability mass assigned to non-DNA vocabulary entries is included in the adenine marginal. This implementation detail is retained here because it affects the exact reproduced objective.

Let $\mathcal{I}$ contain shifted target positions whose labels are valid DNA 6-mers. Unknown, padding, ignored, and non-DNA targets are excluded. The official trainer uses only the base-pair marginal loss

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{6|\mathcal{I}|} \sum_{(i,t)\in\mathcal{I}} \sum_{j=1}^{6} \log\left[\max\left\{p_{\theta,j}(g_j(z_{i,t}) \mid h_{i,t}), 10^{-8}\right\}\right]. \tag{11}$$

No token-level cross-entropy term is added. The descriptive base-pair perplexity is $\mathrm{bpPPL} = \exp(\mathcal{L}_{\mathrm{SFT}})$.

Observed magnitude of the non-DNA contribution. A limited HepG2 validation audit reports mean special-token probability mass of approximately 0.26%, contributing approximately 0.17–1% of the adenine marginal depending on position. A short generation check reported no invalid 6-mers, discarded tokens, or N-padding. These observations bound the measured effect in that check; they do not establish zero non-DNA mass across the full corpus or equivalence to a DNA-only marginal objective. The reported checkpoints retain Equation equation 10; changing its vocabulary support would require SFT retraining and renewed PPO evaluation.

C.3 OPTIMIZATION SETTINGS

Figure 6 reports the corresponding SFT training losses for the two implemented backbones. The curves are shown once because the unconditional DNA corpus is shared across the three downstream target cells.

Both curves decrease rapidly during the initial updates and then reach a stable regime: GENERATOR-V2 changes from 1.144 to 1.131 over the plotted run, whereas HyenaDNA changes from 1.208 to 1.145. These absolute values are not a backbone ranking because GENERATOR-V2 uses the 6-mer nucleotide-marginal loss in Equation 11, whereas HyenaDNA uses a single-base

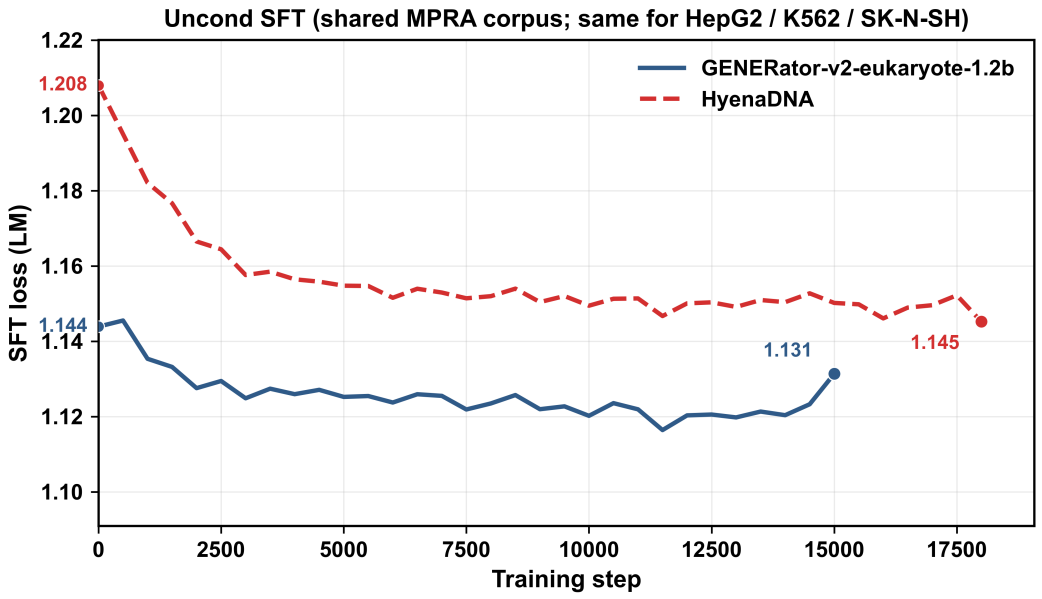


Figure 6: Unconditional SFT losses on the shared MPRA corpus over two epochs.

causal language-model objective. The figure therefore verifies optimization stability; downstream activity, diversity, and motif behavior are assessed with the held-out generation protocol rather than inferred from SFT loss.

C.4 Training algorithm

Algorithm 1 summarizes the shared unconditional SFT procedure.

Algorithm 1 Shared unconditional supervised fine-tuning in Ariadne

Require: Shared training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, pretrained parameters θ_0, 6-mer tokenizer τ

Ensure: Shared SFT parameters $\theta^{\star}_{\text{SFT}}$

Initialize $\theta \leftarrow \theta_0$, AdamW, and the warm-up/cosine schedule
for each training step s, for at most 2 epochs **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}_{\text{train}}$
  for all $x_i \in \mathcal{B}$ **do**
    $L'_i \leftarrow 6\lfloor |x_i|/6 \rfloor$; $x'_i \leftarrow x_i[1 : L'_i]$
    $\mathbf{u}_i \leftarrow [\langle \text{BOS} \rangle, \tau(x'_i)]$
    Mask unknown, padding, ignored, and non-DNA targets
  end for
  Compute causal logits and the nucleotide marginals in Equation 10
  Compute $\mathcal{L}_{\text{SFT}}$ using Equation 11
  Backpropagate, clip the gradient norm to 1.0, and update θ with AdamW
  if s is a multiple of 500 **then**
    Evaluate the loss on $\mathcal{D}_{\text{val}}$
  end if
  if s is a multiple of 5,000 **then**
    Save a checkpoint
  end if
end for
return the checkpoint with the lowest validation loss

C.5 BEFORE- AND AFTER-SFT EVALUATION CHECKPOINTS

The *Before SFT* rows in the main result table use the untouched `GENERator-v2-eukaryote-1.2b-base` checkpoint. The *After SFT* rows use the one retained shared checkpoint. For compatibility with target-specific PPO launch paths, the artifact is present under `outputs/generator_uncond/sft_*`, but the retained model files are identical copies of one shared checkpoint, not separate models for the three cells. Training runs the complete two-epoch schedule; after completion, the saved checkpoint with the minimum validation language-model loss is selected.

Both stages are evaluated with the unconditional BOS-only decoder used by the main pipeline: 42 new 6-mer tokens, terminal cropping to 200 bp, temperature 1.0, top-$p = 1.0$, 256 sequences per generation seed, and seeds 42–100. Thus, the main-table SFT rows summarize 59 generation seeds for one fixed checkpoint for the shared model under each target predictor. The cross-Enformer SFT analysis in Table 10 uses a materialized 15-seed subset of the same new before/after experiment; it is not an additional training split and is not pooled with the five PPO training seeds.

D CELL-SPECIFIC PPO OPTIMIZATION

D.1 BOS-ONLY ROLLOUT CONSTRUCTION

For each cell line, PPO stage 1 initializes a policy from the same shared unconditional SFT checkpoint; later stages use the stagewise warm start described below. Every stage loads the corresponding cell-specific Enformer regression checkpoint. The tokenizer receives an empty prompt and inserts exactly one BOS token. The official rollout and final evaluation both sample from the standard Transformers categorical distribution over raw 6-mer logits with temperature 1.0 and top-$p = 1.0$. The repository explicitly bypasses the checkpoint's custom base-pair sampler so that the sampling distribution agrees with the token log-probabilities recomputed by TRL during PPO.

Each rollout requests 42 new 6-mer tokens. Valid DNA tokens are flattened to bases; the implementation retains the terminal 200 bases. If fewer than 200 valid bases remain after non-DNA tokens are discarded, the sequence is left-padded with `N`. This decode-and-crop rule, rather than a claim that 42 tokens equal 200 bp, defines the exact PPO action-to-sequence mapping.

The complete sampled token response and the decoded 200-bp candidate have different roles. TRL computes the PPO trajectory and token log-probabilities from the complete response. Only the decoded candidate is passed to the activity predictor. Thus, source tokens for bases removed by terminal cropping remain in the PPO trajectory even though those bases are absent from the predictor input.

D.2 ENFORMER REWARD SHAPING

Let $a_{t,i} = f_c(x_{t,i})$ be the raw cell-specific Enformer score of sequence i in PPO batch t. The reward shaper maintains exponentially weighted estimates of the mean μ, second moment m_2, and batch 0.90 quantile q. Except for the first batch, the current batch is transformed using the state from the preceding batch. At the first batch of each stage, the current batch mean, second moment, and 0.90 quantile are used as the reference statistics for the transformation below; this is a batch-normalized cold start rather than a zero-valued fallback:

$$\sigma_{t-1} = \sqrt{\max\{m_{2,t-1} - \mu_{t-1}^2, 10^{-8}\}}, \tag{12}$$

$$z_{t,i} = \mathrm{clip}\left(\frac{a_{t,i} - \mu_{t-1}}{\sigma_{t-1} + 10^{-4}}, -5, 5\right). \tag{13}$$

For a batch of size N, stable ascending sorting assigns a zero-based rank $r_{t,i} \in \{0, \ldots, N-1\}$, and

$$s_{t,i}^{\mathrm{rank}} = 2\frac{r_{t,i}}{N-1} - 1. \tag{14}$$

The historical-tail term is

$$s_{t,i}^{\mathrm{tail}} = \mathrm{clip}\left(\max\left\{0, \frac{a_{t,i} - q_{t-1}}{\sigma_{t-1} + 10^{-4}}\right\}, 0, 3\right). \tag{15}$$

The terminal sequence reward passed to PPO is

$$R_{t,i} = z_{t,i} + 0.3\, s_{t,i}^{\mathrm{rank}} + 0.6\, s_{t,i}^{\mathrm{tail}}. \tag{16}$$

Define the current-batch statistics as $\bar{a}_t = N^{-1}\sum_i a_{t,i}$, $\overline{a^2}_t = N^{-1}\sum_i a_{t,i}^2$, and $\widehat{q}_t = Q_{0.90}(\{a_{t,i}\}_{i=1}^N)$. After computing the reward, the shaper updates its state exactly as

$$(\mu_t, m_{2,t}, q_t) = \begin{cases} (\bar{a}_t, \overline{a^2}_t, \widehat{q}_t), & t = 1, \\ 0.99(\mu_{t-1}, m_{2,t-1}, q_{t-1}) + 0.01(\bar{a}_t, \overline{a^2}_t, \widehat{q}_t), & t > 1. \end{cases} \tag{17}$$

Thus the first batch both supplies its own reference statistics and initializes the stored state, whereas later batches use decay 0.99. Importantly, for $t > 1$, batch t is transformed using the state available before this update.

All reward-state calculations use float32. The variance is floored at 10^{-8} and the standardized denominators include 10^{-4}. The threshold q_t is therefore an EMA of batch-level 0.90 quantiles, not the exact 0.90 quantile of all candidates seen so far. Stable sorting gives tied values consecutive ranks in their original order; a singleton batch receives rank zero, and an empty batch leaves the state unchanged. With the stated clips and weights, the activity reward in Equation 16 lies in $[-5.3, 7.1]$ before the token-level KL penalty is applied.

No motif score, motif grammar score, TFBS Lagrange multiplier, Malinois prediction, diversity term, naturalness term, or embedding-similarity term is included in Equation 16.

D.3 Token-level PPO objective

Ariadne uses TRL 0.11.4. For response token z_k sampled by the pre-update policy π_{old} during stage s, TRL constructs

$$r_k^{\mathrm{token}} = -\beta_{s,t} \log \frac{\pi_{\mathrm{old}}(z_k \mid h_k)}{\pi_s^{\mathrm{ref}}(z_k \mid h_k)} + \mathbf{1}\{k = K\} R(x), \tag{18}$$

where the sequence-level activity reward is placed only on the terminal token and $\beta_{s,t}$ is adapted at update t within stage s. A value head supplies advantage estimates. With $w_k(\theta) = \pi_\theta(z_k \mid h_k)/\pi_{\mathrm{old}}(z_k \mid h_k)$, the clipped policy objective is

$$\mathcal{J}_{\mathrm{clip}}(\theta) = \mathbb{E}_k\Big[\min\Big\{w_k(\theta)\widehat{A}_k, \mathrm{clip}(w_k(\theta), 1-\epsilon_{\mathrm{PPO}}, 1+\epsilon_{\mathrm{PPO}})\,\widehat{A}_k\Big\}\Big]. \tag{19}$$

The advantages already incorporate the KL-penalized token rewards in Equation 18; no second KL term is added to Equation 19. The trainer jointly optimizes its value loss. The rollout policy changes after an update, whereas the reference policy is fixed within a stage and refreshed at the next stage boundary. For exact reproduction, Ariadne overrides the fields listed in Table 4 and otherwise retains TRL 0.11.4 defaults: $\epsilon_{\mathrm{PPO}} = 0.2$, value clip range 0.2, $\gamma = 1.0$, GAE $\lambda = 0.95$, value-loss coefficient 0.1, and initial KL coefficient 0.2. TRL uses Adam without a repository-level override of its β values or weight decay, and reward whitening and score scaling are disabled.

D.4 PPO settings and stagewise warm start

The configured `target_kl`=0.004 and the adaptive-controller setpoint `target`=6.0 are distinct TRL fields. Controlled reward trials retain the fixed 10×100 schedule stated above.

The nominal training budget is therefore $10 \times 100 \times 256 = 256{,}000$ sampled candidates per target cell and training seed, excluding checkpoint evaluation and final generation. Controlled reward trials complete this full schedule; the selection scope of the figure-aligned endpoint summaries is stated in Appendix I.2.

Within a stage, the normalized Enformer scores are used to compute

$$\mathrm{Fitness} = \mathrm{median}(\text{highest 128 scores}), \tag{20}$$

$$\mathrm{Top} = \mathrm{mean}(\text{highest 16 scores}), \tag{21}$$

$$S_{\mathrm{select}} = 0.7\ \mathrm{Fitness} + 0.3\ \mathrm{Top}\,. \tag{22}$$

The policy with the largest observed S_{select} at a checkpoint evaluation is retained as that stage's `best_policy`.

The launcher realizes the ten 100-step stages as separate processes. The complete state-transition and reset logic is given in Algorithm 2.

Table 4: Official PPO and generation settings.

Hyperparameter	Value
Training budget	10 stages $\times$ 100 PPO steps
Implementation	TRL 0.11.4
Rollout batch / mini-batch	256 / 64
PPO epochs per rollout	4
Learning rate	10^{-5}
Policy / value clip range	0.2 / 0.2
GAE γ / λ	1.0 / 0.95
Value-loss / initial KL coefficient	0.1 / 0.2
Configured `target_kl`	0.004
Adaptive KL target	6.0
Importance-ratio threshold	5
Skipped-mini-batch safety fraction	0.95
Reward EMA decay / tail quantile	0.99 / 0.90
EMA-score / tail clips	5 / 3
Numerical epsilon	10^{-4}
Rollout / evaluation sampling	temperature 1.0, top-$p = 1.0$
Maximum new tokens	42 6-mer tokens
Policy-selection interval	5 steps, plus steps 1 and 100
Selection set size	top 128; top summary over 16
Reward	$1.0\, z_{\mathrm{EMA}} + 0.3\, s^{\mathrm{rank}} + 0.6\, s^{\mathrm{tail}}$
Training seeds	0, 1, 2, 42, 100

D.5 PPO ALGORITHM

Algorithm 2 summarizes the official stagewise PPO procedure. Reporting-only metrics do not enter the reward or policy update.

Algorithm 2 BOS-only cell-specific stagewise PPO

Require: Shared unconditional SFT policy θ_0, cell-specific Enformer f_c, exactly $S = 10$ stages and $T = 100$ steps per stage
Ensure: Selected policy $\theta_c^\star$

for $s = 1, \ldots, S$ **do**
  $\theta \leftarrow \theta_0$ if $s = 1$; otherwise $\theta \leftarrow \theta_{s-1}^{\mathrm{best}}$
  Set $\pi_s^{\mathrm{ref}} \leftarrow \mathrm{copy}(\theta)$; reinitialize the PPO optimizer, adaptive KL controller, reward EMA, and within-stage best-policy tracker
  for $t = 1, \ldots, T$ **do**
    From a BOS-only prompt, sample 256 sets of 42 6-mer tokens
    Drop non-DNA tokens, retain the terminal 200 bp, and left-pad short sequences with `N`
    Set $a_i \leftarrow f_c(x_i)$, compute R_i using Equation 16, and then update the EMA state
    if $t = 1$, $t = T$, or t is a multiple of 5 **then**
      Compute $S_{\mathrm{select}} = 0.7\,\mathrm{Fitness} + 0.3\,\mathrm{Top}$
      If the score strictly improves, save the current policy as θ_s^{best}
    end if
    Update θ with batch size 256, mini-batch size 64, and 4 PPO epochs
  end for
end for
return the stage-10 policy $\theta_{10}^{\mathrm{best}}$

Checkpoint scoring and saving occur before the PPO update associated with the same rollout. Therefore, the policy entering the last update and the final post-update policy are distinct candidates, and the latter is not selected automatically. Selection uses only normalized Enformer activity through

Equation 22; it does not use the shaped reward, Malinois, diversity, embedding similarity, or motif statistics.

E Compared Baselines and Related Work

E.1 Compared optimization baselines

SFT prior. The pre-PPO supervised checkpoint is the local generative control. It samples from the learned enhancer-sequence distribution without querying an activity predictor during generation. Comparisons against this checkpoint isolate the effect of predictor-guided PPO from the effect of domain SFT.

TACO. TACO fine-tunes an autoregressive regulatory-DNA generator with reinforcement learning and combines a predictor-derived activity reward with inferred TFBS roles, a high-scoring replay buffer, and entropy regularization (Yang et al., 2025). It is the closest published task-level reference because it reports the same three human cell lines and 200-bp enhancer setting, although its initializer, reward, number of optimization rounds, and conditioning protocol differ from Ariadne.

PEX. Proximal Exploration is a model-guided evolutionary method that favors high-scoring proposals requiring relatively few mutations from their parent or reference sequence (Ren et al., 2022). Its original application is protein design; the regulatory-DNA comparison uses the adaptation reported by TACO.

AdaLead. AdaLead is an adaptive greedy directed-evolution algorithm. It repeatedly mutates high-scoring leaders and continues local ascent while proposals improve under the learned landscape model (Sinai et al., 2020). This makes it a strong mutation-based control but does not provide an autoregressive sequence prior comparable to Ariadne's SFT policy.

BO. Bayesian optimization maintains a probabilistic surrogate and selects new queries through an acquisition rule that trades off exploitation and uncertainty-aware exploration (Snoek et al., 2012). The regulatory-DNA numbers used in prior comparison material come from the FLEXS-style implementation evaluated by TACO, rather than from a new Ariadne-side run (Sinai et al., 2020; Yang et al., 2025).

CMA-ES. The covariance matrix adaptation evolution strategy updates a multivariate search distribution using selected high-scoring samples (Auger & Hansen, 2012). TACO applies it to a continuous one-hot representation of DNA; projection or decoding back to discrete sequence space makes this protocol materially different from autoregressive token sampling.

HyenaDNA. HyenaDNA is a single-nucleotide autoregressive genomic foundation model based on the subquadratic Hyena operator (Nguyen et al., 2023). Ariadne reruns its ERT PPO recipe with this backbone as a local implementation comparison; the exact protocol and its confounders are given in Section H.

TACO's reported suite also contains DNARL, a sequence-mutation PPO baseline adapted from latent-space protein optimization. It is included in Table 5; TACO does not report its embedding similarity for this setting.

E.2 Relationship to regulatory-DNA design methods

Ariadne belongs to the family of predictor-guided sequence-design methods. Fast SeqProp performs differentiable activation maximization over relaxed sequence representations (Linder & Seelig, 2021), whereas GENERator-v2 supplies a long-context autoregressive genomic prior (Li et al., 2026). DNA-Diffusion uses conditional diffusion with classifier-free guidance and prospective validation (DaSilva et al., 2026); DREAM-DNA uses discrete flow matching with reward-guided masking and resampling (Wu et al., 2026). TACO adds inferred TFBS roles to an RL objective (Yang et al., 2025), Ctrl-DNA formulates motif-related requirements as constraints (Chen et al., 2025), and GO-CRE reconstructs sequence-grammar trajectories (Ma et al., 2026). Ariadne instead studies black-box PPO adaptation of one unconditional autoregressive prior and keeps motif agreement, Malinois activity, diversity, and naturalness as post hoc diagnostics.

Table 5: **External context.** Each cell shows Top/Fitness and Diversity. External five-run values are from the TACO human-enhancer hard setting (Yang et al., 2025); protocols differ.

Method	HepG2	K562	SK-N-SH
	0.76 (0.01)/0.74 (0.01)	0.93 (0.01)/0.91 (0.01)	0.92 (0.00)/0.88 (0.01)
TACO	131.80 (4.27)	125.60 (2.88)	137.40 (1.14)
	0.85 (0.04)/0.82 (0.02)	0.95 (0.01)/0.92 (0.02)	0.90 (0.02)/0.86 (0.03)
PEX	16.00 (2.65)	15.90 (1.34)	23.00 (2.74)
	0.74 (0.02)/0.73 (0.02)	0.85 (0.01)/0.84 (0.01)	0.79 (0.05)/0.78 (0.05)
AdaLead	8.00 (1.87)	8.80 (1.64)	14.40 (4.45)
	0.68 (0.04)/0.63 (0.08)	0.74 (0.03)/0.70 (0.04)	0.71 (0.06)/0.63 (0.04)
BO	39.80 (5.07)	37.00 (6.52)	39.90 (6.60)
	0.69 (0.03)/0.43 (0.02)	0.73 (0.05)/0.47 (0.05)	0.74 (0.02)/0.44 (0.03)
CMA-ES	77.20 (5.17)	76.80 (4.55)	76.00 (3.54)
	0.76 (0.04)/0.72 (0.01)	0.89 (0.01)/0.87 (0.02)	0.82 (0.01)/0.81 (0.01)
DNARL	20.00 (3.42)	17.50 (3.33)	18.70 (3.21)
	0.93 (0.03)/0.92 (0.03)	**0.99 (0.01)/0.97 (0.01)**	**0.94 (0.02)/0.92 (0.02)**
Ariadne (ERT)	**138.34 (3.41)**	**132.68 (1.34)**	**138.48 (5.47)**

E.3 Provenance and limits of external comparisons

Table 5 reports external TACO, PEX, AdaLead, BO, CMA-ES, and DNARL results transcribed from TACO's human-enhancer "hard" oracle-guided experiment, summarizing five reported runs (Yang et al., 2025). These values were not regenerated by Ariadne. The table reports Top, Fitness (called "Medium" by TACO), and Diversity from that setting, without mixing in embedding similarity from the separate offline-MBO experiment.

Ariadne combines high predicted activity with greater diversity than the mutation/search baselines, while TACO is the closest external diversity reference. Different pretrained models, budgets, checkpoint rules, and repeated-generation protocols make these contextual comparisons rather than a controlled benchmark. They do not establish state-of-the-art performance. A strict comparison would match the initializer or initial pool, oracle, query budget, sequence length, training seeds, candidate count, sampling, checkpoint selection, and evaluation implementation.

F Evaluation Protocol and Metrics

F.1 Evaluation unit and aggregation

For each trained policy, final evaluation uses generation seeds 42–100 (59 seeds) and samples 256 sequences per seed. A generation seed measures Monte Carlo variation for one fixed policy and is not an independent training replicate. Seed-level summaries are first averaged within a trained policy; the main error bars are then sample standard deviations across the five independent PPO training seeds 0, 1, 2, 42, and 100. Motif correlation is the exception: all generated sequences from one trained policy are pooled before the policy-level value is computed, and only then are the five policy values summarized.

F.2 Activity predictors and normalization

Enformer (E) is both the PPO optimization oracle and the primary activity evaluator (Avsec et al., 2021). Its scores are linearly normalized using the fixed Ctrl-DNA intervals $[-6.051336, 10.992575]$ for HepG2, $[-5.857445, 10.781755]$ for K562, and $[-7.283977, 12.888308]$ for SK-N-SH. Malinois (M) is a reward-independent secondary predictor trained from MPRA measurements (Gosai et al., 2024). Each 200-bp candidate is flanked by 200 bp of fixed MPRA vector sequence on each side to form its 600-bp Malinois input, and its output is normalized with fixed full-dataset, cell-specific minimum and maximum values. Neither transformation is clipped, so a normalized score can lie outside $[0, 1]$ and must not be interpreted as a probability.

The local Hugging Face checkpoint uses `MalinoisForSequencePrediction` with input length 600 and a three-output regression head. Output indices 0, 1, and 2 correspond to

Table 6: Malinois output mapping and fixed normalization constants.

Cell	Output index	$s^{(c)}_{\min}$	$s^{(c)}_{\max}$
HepG2	1	-1.92500138	8.15189648
K562	0	-1.41078162	10.60450172
SK-N-SH	2	-2.72447777	11.12477207

K562, HepG2, and SK-N-SH, respectively. If a sequence is shorter than 600 bp, the remaining length is divided between the left and right sides; the suffix of `MPRA_UPSTREAM` and prefix of `MPRA_DOWNSTREAM` in `src/config/constants.py` supply the fixed flanks. A longer input is center-cropped. The evaluator then selects `output.logits[:, index]` for the requested target cell.

For cell c, a raw Malinois logit s is transformed as

$$\widetilde{s}_c = \frac{s - s^{(c)}_{\min}}{s^{(c)}_{\max} - s^{(c)}_{\min}}. \tag{23}$$

The constants in Table 6 were computed once over all 763,684 retained sequences by `src/eval/compute_malinois_dataset_interval.py`. The implementation does not clip Equation 23. Unlike result-table entries, these calibration constants are shown at stored precision to permit exact reconstruction.

For a predictor, sort the $N = 256$ candidates as $v_{(1)} \geq \cdots \geq v_{(N)}$. Activity summaries are

$$\text{Top} = \frac{1}{16}\sum_{i=1}^{16} v_{(i)}, \tag{24}$$

$$\text{Fitness} = \text{median}\{v_{(1)}, \ldots, v_{(128)}\}. \tag{25}$$

Fitness is retained because it enters checkpoint selection in Equation 22 and is reported alongside Top. E Top and M Top independently rank the same candidate pool with their respective predictors. Hence they test different notions of high predicted activity and should be compared within, not across, predictor columns.

To assess predicted target-cell specificity, the same generated batch is also scored by all three cell-specific Enformer predictors. For target cell c and off-target cell c', we compute

$$\Delta R_{c \to c'} = \text{E Top}_c(\mathcal{B}_c) - \text{E Top}_{c'}(\mathcal{B}_c), \tag{26}$$

where $\mathcal{B}_c$ is generated by the policy optimized for c, and each predictor uses its own fixed normalization interval. Positive ΔR means that the batch scores higher under its target predictor than under the off-target predictor. This is a post hoc, within-family predictor diagnostic, not an independent biological validation of cell-type selectivity.

F.3 Sequence-space diagnostics

Let $\mathcal{S}_{128}$ be the predictor-ranked top-128 multiset, retaining duplicate sequences, and let d_H denote basewise Hamming distance. Then

$$\text{Diversity} = \text{median}\{d_H(x_i, x_j) : i < j,\ x_i, x_j \in \mathcal{S}_{128}\}, \tag{27}$$

$$\text{Novelty} = \text{median}_{x \in \mathcal{S}_{128}} \min_{r \in \mathcal{R}_{1000}} d_H(x, r). \tag{28}$$

Here $\mathcal{R}_{1000}$ is the first 1,000 sequences loaded from the reference table, so Novelty measures distance to this fixed subset rather than the entire training collection. The exact unique rate is the number of distinct strings divided by 256 and is computed on the complete candidate pool. Hamming distances are measured in base positions, not token positions.

Unless a metric name is explicitly prefixed by M, reported Diversity and Novelty use the Enformer-ranked $\mathcal{S}_{128}$. This convention is needed because Enformer and Malinois can select different top-128 subsets from the same 256 candidates.

F.4 Representation, motif, and naturalness diagnostics

For Enformer embeddings e_i of its top-K candidates, with $K = 128$, the representation-concentration statistic is

$$\mathrm{EmbSim} = \frac{2}{K(K-1)} \sum_{i<j} \frac{e_i^\top e_j}{\|e_i\|_2 \|e_j\|_2}. \tag{29}$$

Lower values indicate less concentration in Enformer representation space; this is complementary to, and not interchangeable with, Hamming Diversity. The implementation calls the paired Enformer checkpoint with `return_emb=True`. Its trunk output has shape $(N, L, 2d)$ and is mean pooled over the length axis to obtain one vector per sequence; the reported checkpoints use $d = 512$ and therefore 1,024-dimensional vectors. EmbSim is then computed from the off-diagonal entries of the pairwise cosine-similarity matrix of the target-Enformer top 128. Motif agreement is the Pearson-correlation statistic defined in Equation 4 and is computed only for reporting.

Naturalness diagnostics use the complete 256-sequence pool. For a sequence x, let $u(x)$ be the fraction of invalid bases, $g(x)$ its GC fraction among valid bases, and $h_7(x)$ the fraction of length-seven windows containing one repeated valid base. The implementation reports these components and

$$V(x) = \mathrm{clip}\left(u(x) + \frac{[0.25 - g(x)]_+ + [g(x) - 0.75]_+}{0.50} + h_7(x), 0, 1 \right). \tag{30}$$

Thus the diagnostic flags invalid symbols, GC fraction outside $[0.25, 0.75]$, and homopolymer runs longer than six bases. It is a heuristic quality-control measure, not a learned biological score, and it does not enter SFT, PPO reward, or policy selection.

F.5 Interpretation and statistical scope

Except for the SFT results, all PPO result tables report the mean (sample standard deviation) across five training seeds. Existing tables use two decimal places; supplementary normalization and embedding tables use three. Parentheses in the main before/after-SFT rows summarize generation-seed variation for one fixed checkpoint; the appendix before/after-SFT cross-Enformer table reports the available means over 15 generation seeds because its source summary does not include dispersion. No confidence interval, formal paired test, or multiple-comparison correction is applied. With $n = 5$ training seeds and many reward–cell–metric summaries, differences of 0.01–0.02 that are comparable to their sample SD are not treated as inferential evidence. The stagewise curves are descriptive and do not constitute a formal stability analysis.

ERT changes reward geometry and magnitude relative to adaptive KL regularization and the value loss. Raw Enformer is a protocol-matched but not scale-matched control. Fixed-z freezes SFT score statistics, whereas batch-z recomputes them from the current rollout (Appendix I.6). Neither establishes matched realized KL, a scale-independent EMA effect, or equivalence of the shaped surrogate and raw expected activity. Non-resetting EMA, a reward-scale–KL sweep, and an activity–diversity Pareto match remain untested.

Enformer is not an independent validator because it supplies the PPO reward, checkpoint score, and primary evaluation. Malinois is reward-independent but is trained from related MPRA data and independently selects its own top 16; it does not evaluate the same Enformer-selected candidates. Cross-predictor agreement therefore cannot exclude oracle exploitation without fixed-set scoring, rank correlations, or top-set overlap. The cell-specific normalization intervals also preclude interpreting absolute E–M or cross-cell differences on a common physical scale.

External TACO, PEX, AdaLead, and related values were not rerun under Ariadne's initializer, oracle, query budget, sampling, and checkpoint protocol. DNA-Diffusion and DREAM-DNA likewise have no matched head-to-head comparison, while HyenaDNA changes backbone, tokenization, and action horizon together. Finally, Novelty is measured only against the fixed 1,000-sequence subset $\mathcal{R}_{1000}$, not the full 669,233-sequence training split or the 763,684-record corpus, so it cannot exclude a close training-set neighbor or memorization. Prospective MPRA, explicit off-target optimization, and a true cell-selective assay remain necessary for biological and selectivity claims.

Table 7: Model, data, and evaluation assets.

Asset	Identifier or path
GENERator-v2 base	`GENERator-v2-eukaryote-1.2b-base/`
HyenaDNA base	`LongSafari/hyenadna-medium-160k-seqlen-hf`
Enformer predictors	`enformer/ckpt/human_regression_paired_CELL.ckpt`, where `CELL` is `hepg2`, `k562`, or `sknsh`
Malinois predictor	`Malinois/` (local Hugging Face checkpoint)
Retained enhancer table	Released 763,684-sequence coordinate-annotated MPRA table
Motif library	**JASPAR**; FIMO threshold 10^{-3}
Flank definitions	`MPRA_UPSTREAM` and `MPRA_DOWNSTREAM` in `src/config/constants.py`

G Reproducibility Manifest

Table 7 records the model and evaluation assets used by the reported experiments. Identifiers are repository-relative unless a public model identifier is shown.

The pinned training environment uses PyTorch 2.7.1+cu128, Transformers 5.7.0, TRL 0.11.4, and Accelerate 1.14.0. Training and evaluation use NVIDIA CUDA GPUs; the GPU model can vary across cluster nodes. Motif scans use a MEME Suite FIMO build compatible with the JASPAR file in Table 7. Repository paths and executable scripts are reported so that hardware-dependent throughput is not conflated with the fixed optimization budgets.

H HyenaDNA Comparison Protocol

The backbone comparison initializes from `LongSafari/hyenadna-medium-160k-seqlen-hf` (Nguyen et al., 2023). The reported HyenaDNA comparison follows an unconditional SFT objective on the same shared DNA corpus as the main GENERator-v2 pipeline. HyenaDNA likewise uses one shared SFT model with neither cell identity nor cell-specific measurements; target dependence begins with the Enformer-guided PPO runs. PPO rollout and final generation start from a BOS-only prompt without a cell-condition token. HyenaDNA tokenizes DNA at one base per token, uses a maximum input length of 256 during SFT, and generates 200 new tokens to obtain the 200-bp candidate without 6-mer terminal cropping. By contrast, GENERator-v2 samples 42 six-mer actions (up to 252 decoded bases), discards non-DNA tokens, retains the terminal 200 valid bases, and left-pads short outputs. Tokens whose bases are cropped can therefore remain in its PPO trajectory even though they are absent from the scored Enformer input. Its SFT schedule uses two epochs, per-device train/evaluation batch sizes 64/16, learning rate 2×10^{-5}, AdamW with weight decay 10^{-2}, a 2,000-step warm-up followed by cosine decay to 0.1 of the peak rate, gradient clipping at 1.0, bf16, and gradient checkpointing. Checkpoints are evaluated every 500 steps and the minimum-validation-loss checkpoint is selected after the complete two-epoch schedule. Its SFT optimization trace is compared with GENERator-v2 in Figure 6.

For PPO, the archived HyenaDNA runs use the same ERT reward coefficients, five training seeds (0, 1, 2, 42, and 100), fixed budget of ten stages by 100 steps, batch/mini-batch sizes 256/64, four PPO epochs, learning rate 10^{-5}, and stage-10 `best_policy` rule as the reported GENERator-v2 reward trials. Both rollout and 59-seed evaluation use temperature 1.0 and top-$p = 1.0$, matching GENERator-v2. The configurations nevertheless differ in pretrained backbone, vocabulary, SFT loss (single-base causal LM versus six-mer nucleotide marginal), action granularity, horizon (200 versus 42), and alignment between sampled actions and scored bases. Matching the data, conditioning, sampling, reward, seeds, and update budget does not isolate a backbone effect; the result is a descriptive system-level comparison, not a causal backbone-only ablation.

Table 8: Enformer embedding similarity before and after PPO, using the figure-aligned Ariadne summaries. Higher similarity indicates greater representation concentration.

Cell	SFT generation seeds	After SFT	Ariadne
HepG2	59	0.401 (0.038)	0.976 (0.006)
K562	17	0.372 (0.034)	0.979 (0.006)
SK-N-SH	17	0.423 (0.051)	0.971 (0.005)

I Reported Multi-Seed Results

Common notation and aggregation. All PPO tables below aggregate each cell line separately; the 15 trials from the three cell lines are never pooled. E denotes Enformer and M denotes Malinois. E Top and M Top are the corresponding normalized top-16 activity scores, whereas E Fitness is the median Enformer score among the top 128 candidates. Each PPO entry is the mean (sample standard deviation) across five PPO training seeds after within-policy aggregation over 59 generation seeds. Existing tables use two decimal places; supplementary normalization and embedding tables use three. The before/after-SFT cross-Enformer table is the stated fixed-checkpoint, 15-generation-seed exception. Because E and M use different predictors and cell-specific normalization intervals, comparisons are meaningful within a column, not as absolute cross-column differences. Diversity and Emb. Sim. are the Enformer-ranked metrics defined in Section F.

I.1 Embedding concentration relative to SFT

Table 8 supplies the SFT reference for the claim that PPO outputs are more concentrated in Enformer representation space. SFT entries summarize generation seeds for one fixed checkpoint; PPO entries summarize five training seeds after within-policy aggregation over 59 generation seeds. Available SFT embedding coverage is 59 seeds for HepG2 and 17 for K562 and SK-N-SH. These dispersions describe different sources of variation and are not a paired training-replicate comparison.

I.2 Aggregation and figure interpretation

The GENERator-based Ariadne/ERT activity, diversity, embedding, and motif summaries use the figure-aligned `figures/uncond_five_seed_s10cut_metrics.json`, with the SFT references from `figures/baseline_metrics.json`. These endpoint summaries are distinct from intermediate rollout statistics in Figure 4. Each endpoint metric is first summarized within a policy and then across five training seeds; no visualization-layer IQR filtering is applied to the endpoint tables.

The archived endpoint summary is identical to the merged summary produced by selecting between available ERT evaluations using Enformer Fitness. This selection reuses the evaluation criterion and can make the reported Enformer endpoint optimistic. Some archived runs also enabled early stopping. Accordingly, these values preserve correspondence with the figures but are descriptive selected summaries, not unbiased estimates from five prespecified fixed-budget trials. They do not establish a causal effect of tail or a strict advantage over protocol-mismatched baselines. Controlled normalization and paired rank comparisons are interpreted separately.

In Figure 3, 1.5×IQR filtering operates on generation-seed ratios within each cell and method: SFT supplies 59 ratios and PPO pools 59 ratios from each of five policies. Markers and bars describe the retained ratios, not five independent training-seed means. In Figure 4, filtering is applied separately to each cell, backbone, stage, and metric across training seeds. It can therefore change which seeds contribute at a plotted point. The curves are descriptive visualizations rather than formal evidence of reduced training variance. The archived HyenaDNA Malinois trace uses a normalized rollout batch mean, whereas the GENERator trace uses a top-16 summary; their absolute levels should therefore not be read as matched Top metrics.

Table 9: Sequence-quality diagnostics over five training seeds; entries are mean (sample SD).

Cell	Uniqueness (%) ↑	Novelty ↑	Violation (%) ↓
HepG2	100.00 (0.00)	131.00 (0.60)	0.21 (0.20)
K562	100.00 (0.00)	131.10 (0.50)	0.01 (0.00)
SK-N-SH	100.00 (0.00)	131.20 (0.20)	0.87 (1.27)

Table 10: Cross-Enformer results before and after shared unconditional SFT. Entries are means over 15 generation seeds for one checkpoint; bold marks the larger mean between the two SFT stages within each cell and column.

Cell	Stage	Target E Top	vs. HepG2	vs. K562	vs. SK-N-SH
HepG2	Before SFT	0.55	—	**−0.13**	**+0.01**
	After SFT	**0.66**	—	−0.14	−0.04
K562	Before SFT	0.68	+0.13	—	**+0.14**
	After SFT	**0.81**	**+0.15**	—	+0.11
SK-N-SH	Before SFT	0.54	−0.01	−0.14	—
	After SFT	**0.71**	**+0.05**	**−0.10**	—

Table 11: Target-minus-off-target Enformer Top gap ΔR; entries are mean (sample SD) over five training seeds.

Target policy	vs. HepG2	vs. K562	vs. SK-N-SH
HepG2	—	+0.07 (0.03)	+0.12 (0.03)
K562	+0.29 (0.01)	—	+0.18 (0.02)
SK-N-SH	+0.20 (0.02)	+0.04 (0.03)	—

I.3 Sequence quality and predicted cell specificity

Table 9 reports three complementary diagnostics for the final Ariadne policies. To preserve two-decimal reporting without hiding small values, uniqueness and mean naturalness violation are expressed as percentages. Novelty remains a basewise Hamming distance.

All three policies generate unique sequences at the reported precision, remain roughly 131 substitutions from the nearest member of the fixed reference subset, and have low compositional-violation rates. These diagnostics do not test enhancer activity or safety.

The new before/after-SFT analysis uses one shared checkpoint at each stage and 15 held-out generation seeds per checkpoint. The same generated distribution is scored with each target-cell predictor. Table 10 reports the generation-seed means available in the author-provided summary.

Target-cell E Top increases after SFT in all three cell lines, whereas the target-minus-off-target gaps are not uniformly positive: both HepG2 gaps are smaller after SFT, and the SK-N-SH score remains below the K562 predictor. Because SFT contains no cell-specific supervision, these changes describe how the predictors score the shared sequence distribution and cannot account for the uniformly positive run-level mean gaps observed only after target-specific Ariadne PPO.

Table 11 applies Equation 26 to the same final evaluation batches. Each run is first averaged over 59 generation seeds and then summarized over five training seeds.

All 30 run–off-target combinations have positive run-level mean ΔR. The gaps are asymmetric: K562-to-HepG2 is larger than HepG2-to-K562, and the SK-N-SH policy separates more strongly from HepG2 than from K562. This supports target-aligned predicted activity, but all scores come from related Enformer predictors with different normalization intervals. The table therefore does not establish causal or experimental cell-type selectivity.

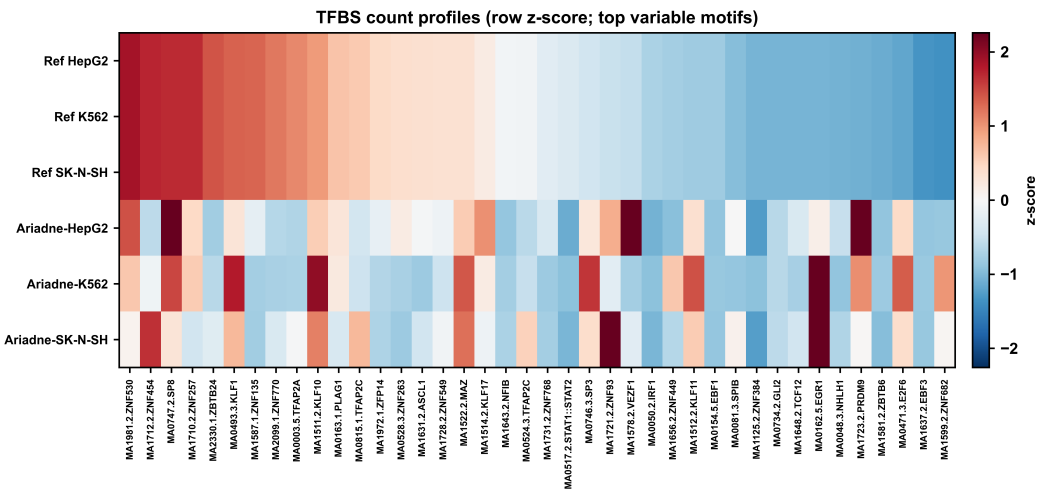


Figure 7: TFBS-count profiles for the reference corpus and final Ariadne generations. Columns are the 40 most variable JASPAR motifs; colors are row-wise z-scores, not absolute motif counts.

I.4 Motif-profile comparison

Figure 7 compares the motif-count profiles of the reference sequences with those of the final Ariadne generations. The display retains the 40 motifs with the largest variation and standardizes each row independently.

The three reference rows show a common broad profile over the selected motif set. In contrast, the final generated sets exhibit target-dependent, motif-specific peaks: examples include SP8, VEZF1, and PRDM9 for HepG2; KLF1, KLF10, and EGR1 for K562; and ZNF454, ZNF93, and EGR1 for SK-N-SH. These differences are consistent with the generated sets occupying distinct motif-composition regimes after cell-specific PPO, but they are descriptive rather than mechanistic evidence. Motifs do not enter the reward, row standardization removes absolute abundance information, and the heatmap neither establishes TF binding nor demonstrates causal regulatory activity. The scalar motif-correlation values in the ablation and backbone tables provide the complementary reference-profile agreement summary.

I.5 Cell-specific reward-component ablation

Controlled reward trials use the shared SFT checkpoint, ten 100-step stages, five matched training seeds, and the same 59-seed evaluation. The seven configurations enable EMA, rank, and tail individually, in pairs, and jointly as EMA+rank+tail (ERT), for $7 \times 3 \times 5 = 105$ trials; the raw-Enformer control adds 15 trials, and fixed-z/batch-z add 30. Tables 12– 14 report the five-training-seed summaries. When enabled, the fixed component weights are $w_{\mathrm{EMA}} = 1.0$, $w_{\mathrm{rank}} = 0.3$, and $w_{\mathrm{tail}} = 0.6$; a disabled component has weight zero. Thus the seven configurations correspond exactly to the seven non-empty subsets of these three weighted components, while sharing EMA decay 0.99 and tail quantile 0.90. ERT entries use the figure-aligned summary described in Appendix I.2; other component entries use their corresponding reward summaries.

For HepG2, ERT has the highest reported Enformer summaries, whereas EMA+rank has the highest Malinois Top. This supports a distinction between improvement under the optimization oracle and transfer to another predictor.

For K562, EMA+rank and ERT have comparable Enformer summaries, while EMA+tail has the highest Malinois Top. The lower diversity of EMA+rank and ERT exposes the accompanying activity–diversity trade-off.

For SK-N-SH, EMA+rank has the highest Enformer summaries, whereas EMA-only has the highest Malinois Top. Across cells, rank provides modest incremental Enformer gains; tail is cell dependent.

Table 12: HepG2 reward ablation across five PPO seeds; lower EmbSim is better.

Reward	E Top	E Fitness	M Top
EMA only	0.90 (0.02)	0.88 (0.02)	0.97 (0.06)
Rank only	0.75 (0.03)	0.71 (0.06)	0.88 (0.12)
Tail only	0.70 (0.03)	0.58 (0.08)	0.73 (0.06)
EMA+rank	0.91 (0.03)	0.89 (0.03)	**0.99 (0.03)**
EMA+tail	0.91 (0.03)	0.89 (0.03)	0.96 (0.03)
Rank+tail	0.77 (0.03)	0.75 (0.04)	0.96 (0.09)
ERT	**0.93 (0.03)**	**0.92 (0.03)**	0.98 (0.06)
Raw Enformer	0.81 (0.02)	0.79 (0.02)	0.93 (0.02)

Reward	Diversity	Emb. Sim.	Motif
EMA only	139.70 (3.00)	0.96 (0.01)	0.43 (0.04)
Rank only	144.70 (2.70)	0.86 (0.12)	0.28 (0.11)
Tail only	**145.70 (4.20)**	**0.55 (0.16)**	0.26 (0.10)
EMA+rank	139.80 (2.10)	0.96 (0.02)	**0.44 (0.03)**
EMA+tail	140.70 (4.00)	0.96 (0.00)	0.42 (0.02)
Rank+tail	141.40 (1.50)	0.91 (0.07)	0.36 (0.04)
ERT	138.34 (3.41)	0.98 (0.01)	**0.44 (0.02)**
Raw Enformer	145.40 (1.20)	0.93 (0.01)	0.32 (0.05)

Table 13: K562 reward ablation across five PPO seeds; lower EmbSim is better.

Reward	E Top	E Fitness	M Top
EMA only	0.98 (0.00)	0.96 (0.00)	0.93 (0.02)
Rank only	0.90 (0.04)	0.85 (0.08)	0.86 (0.09)
Tail only	0.84 (0.06)	0.70 (0.18)	0.73 (0.11)
EMA+rank	**0.99 (0.01)**	**0.97 (0.00)**	0.92 (0.03)
EMA+tail	0.97 (0.01)	0.96 (0.01)	**0.94 (0.03)**
Rank+tail	0.92 (0.03)	0.88 (0.06)	0.89 (0.07)
ERT	**0.99 (0.01)**	**0.97 (0.01)**	0.89 (0.04)
Raw Enformer	0.95 (0.03)	0.93 (0.03)	0.89 (0.03)

Reward	Diversity	Emb. Sim.	Motif
EMA only	135.60 (1.20)	0.97 (0.00)	0.26 (0.04)
Rank only	138.20 (4.40)	0.89 (0.09)	0.26 (0.11)
Tail only	137.60 (8.80)	**0.67 (0.28)**	**0.33 (0.07)**
EMA+rank	132.80 (3.20)	0.98 (0.00)	0.28 (0.03)
EMA+tail	135.60 (1.50)	0.97 (0.01)	0.26 (0.04)
Rank+tail	136.10 (3.60)	0.91 (0.05)	0.29 (0.10)
ERT	132.68 (1.34)	0.98 (0.01)	0.27 (0.02)
Raw Enformer	**140.00 (2.10)**	0.96 (0.02)	0.30 (0.05)

Table 14: SK-N-SH reward ablation across five PPO seeds; lower EmbSim is better.

Reward	E Top	E Fitness	M Top
EMA only	0.92 (0.02)	0.90 (0.01)	**1.12 (0.03)**
Rank only	0.86 (0.04)	0.81 (0.08)	1.00 (0.13)
Tail only	0.82 (0.05)	0.70 (0.14)	0.85 (0.13)
EMA+rank	**0.95 (0.02)**	**0.93 (0.02)**	1.08 (0.05)
EMA+tail	0.92 (0.04)	0.90 (0.05)	1.07 (0.04)
Rank+tail	0.88 (0.03)	0.84 (0.07)	1.03 (0.11)
ERT	0.94 (0.02)	0.92 (0.02)	1.09 (0.05)
Raw Enformer	0.89 (0.01)	0.86 (0.01)	1.08 (0.01)

Reward	Diversity	Emb. Sim.	Motif
EMA only	137.60 (3.10)	0.97 (0.01)	0.19 (0.06)
Rank only	142.70 (3.60)	0.90 (0.09)	0.18 (0.13)
Tail only	140.20 (6.10)	**0.74 (0.20)**	**0.29 (0.10)**
EMA+rank	138.60 (2.10)	0.97 (0.01)	0.24 (0.05)
EMA+tail	139.70 (5.20)	0.96 (0.03)	0.21 (0.11)
Rank+tail	138.50 (3.20)	0.92 (0.08)	0.22 (0.10)
ERT	138.48 (5.47)	0.97 (0.00)	0.23 (0.01)
Raw Enformer	**149.00 (0.00)**	0.97 (0.00)	0.03 (0.03)

Table 15: Normalization controls and the figure-aligned ERT summary. Entries are five-training-seed mean (sample SD); ERT has the selection limitations described in Appendix I.2.

Cell	Reward	E Fitness	M Top
HepG2	Fixed-z	0.768 (0.013)	0.983 (0.033)
HepG2	Batch-z	0.885 (0.015)	0.984 (0.013)
HepG2	EMA only	0.880 (0.019)	0.972 (0.055)
HepG2	ERT	0.916 (0.029)	0.985 (0.059)
K562	Fixed-z	0.928 (0.006)	0.969 (0.024)
K562	Batch-z	0.954 (0.006)	0.937 (0.024)
K562	EMA only	0.963 (0.003)	0.933 (0.024)
K562	ERT	0.973 (0.013)	0.892 (0.044)
SK-N-SH	Fixed-z	0.875 (0.014)	1.079 (0.025)
SK-N-SH	Batch-z	0.907 (0.013)	1.083 (0.027)
SK-N-SH	EMA only	0.901 (0.013)	1.123 (0.030)
SK-N-SH	ERT	0.922 (0.024)	1.090 (0.051)

Raw reward retains greater diversity than ERT. None of these comparisons is a formal significance test, and reward magnitude and realized regularization were not matched.

I.6 Normalization controls and incremental rank effects

Fixed-z uses the mean and standard deviation of Enformer scores under the SFT policy and freezes them during PPO. Batch-z recomputes both statistics from the current rollout. Neither includes rank or tail; EMA-only likewise isolates the normalization term. The fixed-z and batch-z controls comprise $2 \times 3 \times 5 = 30$ trials with ten 100-step stages, no early stopping, and the same SFT initialization, PPO settings, and sampling parameters. Each trained policy is evaluated over 59 generation seeds before averaging across five training seeds. Enformer and Malinois entries use their per-generation-seed summaries (the `seed_mean` fields), rather than mixing Malinois values from a pooled candidate set.

Both batch-z and EMA-only exceed fixed-z in Enformer Fitness in all three cells, supporting a reference that evolves during optimization. EMA-only exceeds batch-z in K562 but falls slightly below it in HepG2 and SK-N-SH. The data therefore do not establish a universal benefit of EMA

Table 16: Paired incremental effect of rank: EMA+rank minus EMA-only. Positive pairs count training seeds with increased Enformer Fitness; generation seeds are not independent training replicates.

Cell	Mean paired Δ E Fitness	Positive pairs
HepG2	$+0.010$	$4/5$
K562	$+0.011$	$5/5$
SK-N-SH	$+0.024$	$5/5$

Table 17: HepG2 backbone comparison across five PPO seeds; lower EmbSim is better.

Metric	GENERATOR-V2 (ERT)	HyenaDNA (ERT)
E Top	$\mathbf{0.93}\,(\mathbf{0.03})$	$0.82\,(0.08)$
E Fitness	$\mathbf{0.92}\,(\mathbf{0.03})$	$0.81\,(0.09)$
M Top	$\mathbf{0.98}\,(\mathbf{0.06})$	$0.89\,(0.07)$
Diversity	$\mathbf{138.34}\,(\mathbf{3.41})$	$122.68\,(11.90)$
Emb. Sim.	$\mathbf{0.98}\,(\mathbf{0.01})$	$\mathbf{0.98}\,(\mathbf{0.01})$
Motif	$\mathbf{0.44}\,(\mathbf{0.02})$	$0.33\,(0.08)$

smoothing over instantaneous normalization. ERT has higher reported Enformer Fitness than either control, but the ordering does not consistently transfer to Malinois; in K562, fixed-z and batch-z have higher Malinois Top than ERT. Fixed-z also falls below raw reward on HepG2 Enformer Fitness, so normalization alone is not uniformly beneficial.

These paired summaries support modest incremental gains from rank; the reported differences use the underlying paired summaries and may differ slightly from subtracting rounded table means. No formal paired significance test is claimed. Tail introduces a distinct upper-tail preference, but the cell-dependent ERT and Malinois results do not establish its necessity.

Coefficient sensitivity. An exploratory HepG2 grid evaluates rank weights $\{0.2, 0.3, 0.4\}$ and tail weights $\{0.45, 0.60, 0.75\}$ while fixing EMA weight 1, decay 0.99, and quantile 0.90. Each setting has one training run; its 59 generation seeds measure within-policy variation. The default $(0.3, 0.60)$ is not uniformly best across Enformer and Malinois. This grid provides local sensitivity evidence, not replicated hyperparameter optimization; decay and quantile robustness remain unestablished.

I.7 CELL-SPECIFIC HYENADNA BACKBONE COMPARISON

The exact archived HyenaDNA protocol and its differences from the GENERATOR-V2 protocol are specified in Section H. All 15 HyenaDNA trials are complete. The means and sample standard deviations below were recomputed from the per-training-seed records in `hyenadna_five_seed_metrics.json`. Figure 4 in the main text summarizes training rollouts, whereas Tables 17–19 report separate held-out generation. Each policy is evaluated over 59 generation seeds before aggregation across five training seeds; the latter remain the independent statistical units. These detailed tables also include embedding similarity and motif agreement.

For HepG2, GENERATOR-V2 achieves higher activity and diversity (Table 17). HyenaDNA has no paired E-Fitness wins and is more variable across training seeds.

For K562, GENERATOR-V2 again achieves higher activity and diversity (Table 18); HyenaDNA wins one of five paired E-Fitness comparisons and is more variable.

For SK-N-SH, HyenaDNA is slightly higher on Enformer activity, with Malinois Top tied after rounding, but retains lower diversity (Table 19). It wins three of five paired E-Fitness comparisons. Overall, GENERATOR-V2 is stronger in HepG2/K562 and more diverse across cells, but not uniformly more active in SK-N-SH. Simultaneous changes in backbone, tokenization, and action horizon prevent causal attribution to the backbone alone.

Table 18: K562 backbone comparison across five PPO seeds; lower EmbSim is better.

Metric	GENERATOR-V2 (ERT)	HyenaDNA (ERT)
E Top	**0.99 (0.01)**	0.93 (0.04)
E Fitness	**0.97 (0.01)**	0.92 (0.05)
M Top	**0.89 (0.04)**	0.84 (0.08)
Diversity	**132.68 (1.34)**	115.82 (14.71)
Emb. Sim.	**0.98 (0.01)**	0.99 (0.00)
Motif	**0.27 (0.02)**	0.21 (0.21)

Table 19: SK-N-SH backbone comparison across five PPO seeds; lower EmbSim is better.

Metric	GENERATOR-V2 (ERT)	HyenaDNA (ERT)
E Top	0.94 (0.02)	**0.95 (0.03)**
E Fitness	0.92 (0.02)	**0.94 (0.03)**
M Top	**1.09 (0.05)**	**1.09 (0.07)**
Diversity	**138.48 (5.47)**	120.95 (9.20)
Emb. Sim.	**0.97 (0.00)**	0.99 (0.00)
Motif	**0.23 (0.01)**	0.21 (0.17)