A BIDIRECTIONAL GENERATIVE DNA MODEL WITH MOTIF-INPAINTING POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Early DNA foundation models adopted BERT-style training, achieving strong performance on sequence understanding tasks but offering limited support for generation. Recent autoregressive models enable DNA generation, yet their left-to-right factorization does not naturally reflect the bidirectional dependencies underlying gene regulation. We present D3LM, which unifies both capabilities through masked diffusion. D3LM retains the Nucleotide Transformer (NT) architecture while reformulating its training objective as discrete masked diffusion. It retains NT's sequence understanding capabilities while generating regulatory sequences that more closely resemble real DNA than those produced by the evaluated autoregressive baselines. Building on D3LM's bidirectional generation capabilities, we introduce Motif-Inpainting Policy Optimization (MIPO) for cell-specific regulatory sequence design. MIPO anchors regulatory motifs within high-scoring sequences and inpaints their local neighborhoods while preserving the outer context, using functional feedback to select improved candidates for reinforcement learning. Our results highlight masked diffusion as a promising framework for unified DNA understanding, generation, and functional design.

1 INTRODUCTION

DNA, the blueprint of biology, encodes the fundamental instructions for life and plays a pivotal role in the central dogma of molecular biology (Crick, 1970). Understanding and designing DNA sequences is crucial for applications ranging from target discovery (Lindsay, 2003) and personalized medicine (Huang et al., 2023) to synthetic biology (Gosai et al., 2024; Yang et al., 2025a; DaSilva et al., 2025).

The recent progress of language models (Devlin et al., 2019; Brown et al., 2020) has inspired the development of genomic foundation models that learn universal representations from large-scale unlabeled DNA sequence data. *BERT-style* models such as DNABERT (Ji et al., 2021; Zhou et al., 2024) and Nucleotide Transformer (NT) (Dalla-Torre et al., 2024; Boshar et al., 2025) employ bidirectional masked language modeling, learning rich representations through predicting randomly masked nucleotides (shown in Figure 1(b)). While these models excel at understanding tasks and effectively capture bidirectional dependencies in DNA sequences, they lack generative capabilities. Recent works (Liu et al., 2025; Chen et al., 2025) suggest that unified understanding and generation can mutually benefit model performance. On the other hand, *autoregressive models* (shown in Figure 1(c)) such as HyenaDNA (Nguyen et al., 2024b), Evo (Nguyen et al., 2024a; Merchant et al., 2026), and Generator (Wu et al., 2025) adopt next-token prediction (Brown et al., 2020) with causal left-to-right modeling, enabling sequence generation. However, this causal modeling paradigm faces fundamental challenges when applied to DNA sequences.

Unlike natural language, where left-to-right autoregressive modeling has proven effective (Kim et al., 2025), DNA regulatory relationships are inherently bidirectional. Multiple lines of biological evidence support this, including that transcription factor binding sites often exhibit palindromic symmetry (Georgakopoulos-Soares et al., 2023) and CpG islands require global GC content constraints (Lal et al., 2024). The most illustrative example is enhancer-promoter interactions. As shown in Figure 1(a), enhancer elements can regulate target genes from variable genomic positions. Some function from upstream locations (Schuur et al., 1996), while others regulate from downstream (Li et al., 2013) of the promoter. Causal autoregressive models, which strictly follow left-to-right genera-

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

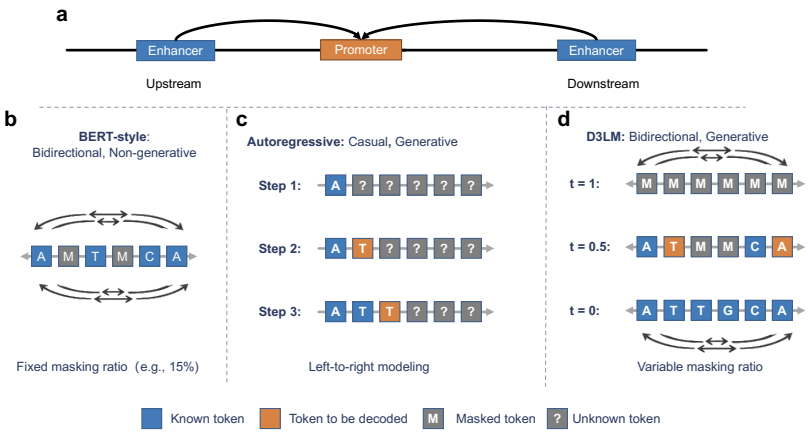


Figure 1: **Comparison of DNA modeling approaches. (a)** Enhancers regulate promoters from either upstream or downstream positions, demonstrating bidirectional regulatory relationships in DNA. **(b)** BERT-style models. **(c)** Autoregressive models. **(d)** D3LM combines bidirectional modeling with generation through masked diffusion with variable masking ratios, enabling iterative prediction and progressive unmasking.

tion order, cannot adequately model downstream enhancers. These biological properties motivate the need for a DNA foundation model that combines bidirectional modeling with unified understanding and generation capabilities.

We address this challenge by introducing D3LM (**D**iscrete **D**NA **D**iffusion **L**anguage **M**odel), a unified framework that bridges understanding and generation through discrete diffusion modeling (Nie et al., 2025). The key distinction between D3LM and NT (Dalla-Torre et al., 2024) lies in the training objective. D3LM learns to predict masked tokens across a continuous range of masking ratios (shown in Figure 1(d)), while NT uses a fixed masking ratio and lacks generative capabilities. This enables D3LM to retain bidirectional modeling capabilities while gaining the ability to generate sequences.

Besides, recent work has applied reinforcement learning (RL) to autoregressive DNA models for cell-type-specific regulatory design (Yang et al., 2025a; Ma et al., 2026; Chen et al., 2026). Group-based policy optimization, such as GRPO (Shao et al., 2024), relies on comparisons among sampled candidates; redundant groups can provide limited reward-driven learning signals. We introduce **M**otif-**I**npainting **P**olicy **O**ptimization (MIPO) to guide this exploration with biological priors. MIPO inserts a sampled cell-associated motif into a high-scoring parent sequence and uses D3LM's bidirectional inpainting to complete its local neighborhood while preserving the motif and outer sequence context. At inference, the trained model generates without motif hints or reward-based filtering.

To investigate bidirectional generative training on an established DNA modeling backbone, we adopt the NT-v2 architecture (Dalla-Torre et al., 2024). We consider two variants: D3LM-R, trained from random initialization, and D3LM, initialized from pretrained NT-v2 weights. Using a shared backbone reduces architectural confounds and allows us to examine bidirectional generative training with and without pretrained initialization.

Our contributions are threefold:

- We develop D3LM, a masked diffusion DNA model that unifies bidirectional understanding and generation.
- We systematically evaluate DNA masked diffusion across generation, downstream understanding, and key modeling and sampling choices.
- We introduce MIPO, a motif-guided sampling strategy that exploits bidirectional inpainting to improve reinforcement learning for functional DNA design.

2 RELATED WORK

Genomic language models. Early genomic models are trained as supervised models, taking DNA sequences as input to predict thousands of genome tracks (Zhou & Troyanskaya, 2015; Zhou et al., 2018; Avsec et al., 2021; Yang et al., 2025b; Avsec et al., 2025). Recent genomic language models have drawn inspiration from the success of unsupervised pre-training in natural language processing (Devlin et al., 2019; Brown et al., 2020). DNABERT (Ji et al., 2021) pioneered this approach by adapting BERT's masked language modeling objective to DNA sequences, demonstrating that pre-trained representations transfer effectively to downstream genomic tasks. Subsequent work has continued along two primary directions. BERT-style models (Zhou et al., 2024; Dalla-Torre et al., 2024; Sanabria et al., 2024) employ bidirectional masked language modeling with fixed masking ratios, achieving strong performance on understanding tasks but lacking the ability to generate novel sequences for applications such as regulatory element design. Conversely, autoregressive models (Nguyen et al., 2024b;a) adopt next-token prediction with causal attention, enabling sequence generation but constraining the model to causal left-to-right processing inherited from natural language modeling.

Masked diffusion models. Discrete diffusion models generate sequences by reversing categorical corruption processes (Austin et al., 2021; Lou et al., 2023). Masked diffusion specializes this framework to token masking, combining bidirectional context with iterative generation. MDLM (Sahoo et al., 2024) develops a simplified training objective, while LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025) demonstrate its scalability to large language models. Applications to DNA include genomic modeling with MDLM (Sahoo et al., 2024) and controllable regulatory sequence generation with NT-v3 (Boshar et al., 2025). Our work studies masked diffusion as a unified DNA modeling framework, evaluating sequence understanding, regulatory generation, and practical modeling choices, and extending its bidirectional completion capability to reinforcement learning for functional design.

Reinforcement learning for language models. Policy optimization enables language models to improve through task-dependent rewards. PPO (Schulman et al., 2017) uses a clipped surrogate objective to control policy updates, while GRPO (Shao et al., 2024) estimates advantages from groups of sampled responses without a separate value model. Extending these methods to diffusion language models requires handling their non-autoregressive generation process. d1 (Zhao et al., 2026a) introduces diffu-GRPO using a mean-field approximation for efficient likelihood estimation. ESPO (Ou et al., 2026) instead formulates sequence-level policy optimization with an ELBO-based likelihood proxy. Beyond the optimization objective, IGPO (Zhao et al., 2026b) exploits diffusion inpainting to guide exploration: partial ground-truth reasoning traces help construct successful responses when ordinary sampling produces all-incorrect groups, inspiring MIPO.

3 METHOD

In this section, we first introduce the bidirectional probabilistic formulation based on discrete masked diffusion models in Section 3.1. We then present the details of our generative training approach in Section 3.2, including model architecture and tokenization strategy. In Section 3.3, we describe the sampling algorithm for sequence generation with D3LM. In Section 3.4, we explain how D3LM can serve as a representation extractor for fine-tuning on diverse genomic downstream tasks. Finally, Section 3.5 introduces MIPO, which uses motif-guided bidirectional inpainting to guide exploration during reinforcement learning for functional DNA design.

3.1 BIDIRECTIONAL PROBABILISTIC FORMULATION

We represent a DNA sequence as a token sequence $\mathbf{x} = (x^1, x^2, \ldots, x^L)$, where $x^i \in \mathcal{V}$ and $\mathcal{V}$ denotes the DNA-token vocabulary associated with the chosen tokenizer. Here, L denotes the number of tokens. Our default tokenizer uses non-overlapping 6-mers, with single-nucleotide tokens for any remaining bases. The sequence is drawn from the data distribution $\mathbf{x} \sim q(\mathbf{x})$. Our goal is to learn this distribution using a parametric generative model $p_\theta(\mathbf{x})$.

Autoregressive DNA models. Mainstream generative DNA language models such as HyenaDNA (Nguyen et al., 2024b) and Evo (Nguyen et al., 2024a; Merchant et al., 2026) employ

autoregressive modeling:

$$\underbrace{p_\theta(\mathbf{x}) = p_\theta(x^1) \prod_{i=2}^{L} p_\theta(x^i | x^1, \ldots, x^{i-1})}_{\text{Autoregressive formulation}}. \tag{1}$$

While this formulation is well-suited for natural language processing (Kim et al., 2025) due to the inherent left-to-right structure of language, DNA sequences do not necessarily follow such sequential dependencies as discussed in Section 1.

Non-generative masked models. Another class of DNA foundation models, such as NT (Dalla-Torre et al., 2024) and DNABERT (Ji et al., 2021; Zhou et al., 2024), adopts a non-generative masked prediction strategy inspired by BERT (Devlin et al., 2019). These models randomly mask a fixed ratio (typically 15%) of positions in the input sequence and train the model to predict the masked tokens using bidirectional attention. While this approach effectively captures bidirectional dependencies in DNA sequences, it lacks generative capabilities because the fixed masking ratio prevents the training objective from being a proper probabilistic generative model.

Bidirectional discrete DNA diffusion language models. To combine the strengths of both paradigms, we adopt an alternative modeling approach following discrete diffusion models (Nie et al., 2025; Ye et al., 2025). Unlike the autoregressive formulation in Eq. (1), D3LM defines a generative model distribution $p_\theta(\mathbf{x}_0)$ through a *forward masking process* and a *reverse denoising process*, following the masked diffusion language model framework Nie et al. (2025).

The forward process gradually masks tokens independently in $\mathbf{x}_0$ until the sequence is fully masked at $t = 1$. For $t \in (0, 1)$, the sequence $\mathbf{x}_t$ is partially masked, with each position being masked with probability t and remaining unmasked with probability $1 - t$. The reverse process recovers the data distribution by iteratively predicting masked tokens as t moves from 1 to 0.

Similar to NT Dalla-Torre et al. (2024), D3LM employs a *mask predictor*, a parametric model $p_\theta(\cdot|\mathbf{x}_t)$ that takes $\mathbf{x}_t$ as input and predicts all masked tokens simultaneously using bidirectional attention. To isolate the impact of the discrete diffusion training objective, D3LM directly adopts the NT v2 architecture (Dalla-Torre et al., 2024), allowing us to attribute performance differences specifically to the training paradigm rather than architectural innovations. The model is trained using a cross-entropy loss computed only on the masked tokens:

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t} \left[\frac{1}{t} \sum_{i=1}^{L} \mathbf{1}[x_t^i = \text{[M]}] \log p_\theta(x_0^i | \mathbf{x}_t) \right], \tag{2}$$

where $\mathbf{x}_0$ denotes a clean sample from the DNA training dataset, t is sampled uniformly from $[0, 1]$, and $\mathbf{x}_t$ is sampled from the forward process defined above. The loss computation is restricted to masked tokens only, as enforced by the indicator function $\mathbf{1}[\cdot]$.

Once trained, D3LM can simulate a reverse process parameterized by the mask predictor and define the model distribution $p_\theta(\mathbf{x}_0)$ as the marginal distribution induced at $t = 0$. The loss in Eq. (2) provides an upper bound on the negative log-likelihood (Sahoo et al., 2024; Ou et al., 2024):

$$-\mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x})} [\log p_\theta(\mathbf{x}_0)] \leq \mathcal{L}(\theta), \tag{3}$$

making it a principled objective for generative modeling. This masked modeling approach offers several advantages for DNA sequences, including the ability to capture bidirectional dependencies and generate sequences in a non-sequential manner.

3.2 Details of Generative Training

The training process is shown in Figure 2(a).

Model architecture. D3LM adopts the NT-v2 backbone (Dalla-Torre et al., 2024), retaining its bidirectional Transformer architecture, rotary position embeddings (Su et al., 2024), and SwiGLU feed-forward layers (Shazeer, 2020). We evaluate model sizes of 50M, 100M, 250M, and 500M parameters.

Tokenization. We use non-overlapping 6-mer tokenization by default, with 4,105 vocabulary entries: 4,096 possible 6-mers and nine additional entries, including single-nucleotide and special tokens.

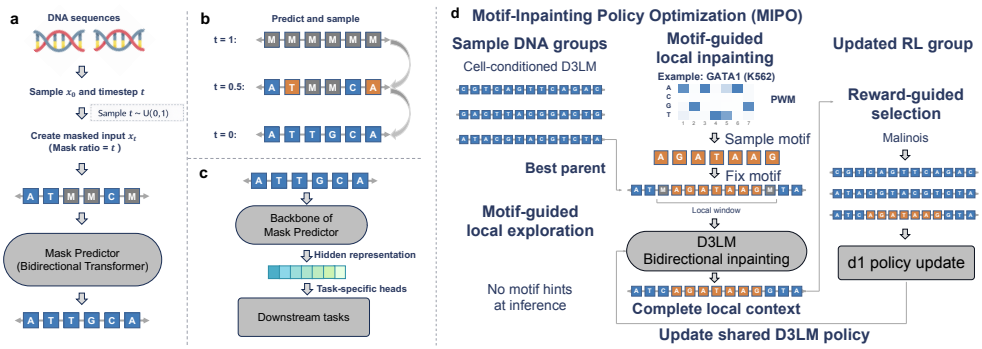


Figure 2: **Overview of D3LM and MIPO. (a)** Masked diffusion training with variable masking ratios and bidirectional attention. **(b)** DNA generation through progressive unmasking. **(c)** Downstream prediction using D3LM representations and task-specific heads. **(d)** MIPO inserts a motif into a high-scoring parent sequence and completes its local surroundings while preserving the outer sequence context. Functional feedback selects improving proposals for d1 policy updates. At inference, the trained policy generates without motif hints.

We also evaluate 1-mer, 3-mer, and 9-mer tokenization. Model-size and tokenization ablations are reported in Section 4.5.

3.3 Sampling

Shown in Figure 2(b), once trained, D3LM generates novel DNA sequences by simulating the reverse denoising process. Starting from a fully masked sequence, we iteratively predict the clean sequence x_0 and selectively unmask tokens until all positions are revealed.

Discrete-time sampling. We discretize the denoising process into T steps. At each step t, the model takes the current partially masked sequence $\mathbf{x}_t$ as input and predicts the probability distribution $p_\theta(x_0^i|\mathbf{x}_t)$ over the original clean tokens for all masked positions. Let $\mathcal{M}_t$ denote the set of currently masked indices. Based on these predictions, we select a subset $\mathcal{U}_t \subset \mathcal{M}_t$ to unmask for the next step. The next state $\mathbf{x}_{t-1}$ is constructed via:

$$\mathbf{x}_{t-1}^i = \begin{cases} \text{sample from } p_\theta(x_0^i|\mathbf{x}_t) & \text{if } i \in \mathcal{U}_t \\ [\text{M}] & \text{if } i \in \mathcal{M}_t \setminus \mathcal{U}_t \\ x_t^i & \text{otherwise} \end{cases} \tag{4}$$

Token selection strategy. We experimented with multiple strategies for selecting which positions to unmask, including confidence-based approaches (MaskGit Chang et al. (2022), entropy-based sampling, top-k margin, and re-masking mechanisms). We use random-order unmasking by default and compare alternative position-selection strategies in Section 4.5. Random order achieves the lowest SFID, while entropy-based sampling achieves the lowest dFID. The number of positions to unmask at each step is determined by a predefined schedule (e.g., linear or cosine). Detailed comparisons of all strategies are provided in Appendix D.

3.4 Representation Learning

D3LM also serves as a bidirectional encoder, extracting contextual representations from unmasked DNA sequences for downstream genomic tasks (Dalla-Torre et al., 2024). We attach lightweight task-specific heads to representations for sequence-level prediction or token-level representations for position-level prediction. The encoder can be frozen or jointly fine-tuned with the task-specific head shown in Figure 2(a).

3.5 MOTIF-INPAINTING POLICY OPTIMIZATION

We extend D3LM to cell-type-specific regulatory sequence design through cell-conditioned supervised fine-tuning followed by reinforcement learning shown in Figure 2(d). Our key idea is to improve the candidates used for policy optimization through motif-guided local exploration. Regulatory motifs provide biologically informed search directions, while D3LM's bidirectional generation allows their surrounding sequences to adapt to both the inserted motif and the retained sequence context. Functional feedback then determines which proposals enter the policy update.

Functional rewards and group-based optimization. Starting from the SFT policy, we sample groups of G sequences conditioned on a target cell type c. The reward model provides a functional score that favors activity in the target cell over the strongest off-target cell, with a scheduled off-target penalty (Ma et al., 2026). We use the same reward-processing procedure for the baseline and MIPO. Our policy optimizer follows d1 (Zhao et al., 2026a), which adapts group-relative policy optimization to masked diffusion models. Its learning signal depends on the sampled alternatives: identical candidates receive identical rewards and yield zero relative advantages. MIPO introduces motif-guided alternatives to expand exploration around promising sequences.

Motif-guided bidirectional inpainting. For each group, we select the sequence with the highest functional score as the parent $\mathbf{b}$. We sample a motif fragment from a cell-associated position-weight matrix, randomly orient it, and place it at a token-aligned internal region $\mathcal{A}$. Let $\mathcal{W} \supseteq \mathcal{A}$ denote a local completion window. We construct the partially masked input

$$\widetilde{x}^i = \begin{cases} a^i, & i \in \mathcal{A}, \\ [\mathrm{M}], & i \in \mathcal{W} \setminus \mathcal{A}, \\ b^i, & i \notin \mathcal{W}, \end{cases} \tag{5}$$

where a^i denotes an inserted motif token. D3LM completes the masked positions using the sampling procedure in Section 3.3, while preserving the motif and the sequence outside the window. This enables local exploration conditioned jointly on the motif and both flanks. Each proposal contains one sampled motif; different proposals can use different placements and motif realizations.

Functional selection within groups. We score the completed proposals using the same functional scorer as the ordinary samples. A proposal is accepted only when its score exceeds the highest score in its original group. Accepted proposals replace the lowest-scoring original members in descending proposal-score order, preserving the group size. Every group receives proposals; their inclusion is determined by functional improvement. Motif information therefore guides candidate construction, while the functional scorer governs selection.

Policy update and inference. After selection, we apply reward preprocessing and recompute the group-relative advantages:

$$A_j = \frac{r_j - \bar{r}}{\mathrm{Std}(r_1, \ldots, r_G) + \epsilon}, \qquad \bar{r} = \frac{1}{G} \sum_{k=1}^{G} r_k, \tag{6}$$

where r_j is the processed reward of the retained sequence. The resulting groups are optimized with d1's clipped diffusion policy-optimization surrogate and its configured reference-policy regularization. We train on complete sequences, including the anchor tokens, treating accepted proposals as guided replay without proposal-density correction. MIPO thus changes the training-group construction while retaining the underlying policy optimizer. All scored proposals, including rejected ones, count toward the training scoring budget. At inference, the trained policy generates sequences conditioned only on the target cell type, without motif hints or reward-based filtering.

4 EXPERIMENTS

We evaluate D3LM on DNA generation and downstream understanding tasks, followed by ablations on tokenization, model size, and sampling. Additional experimental details and complete downstream results are provided in Appendix A.

4.1 Generative Training Details

Dataset and tokenization. We use the mammalian subset of EPD-GenDNA (Li et al., 2024a;b), comprising approximately 80,000 sequences of 2048 base pairs (bp), with an 8:1:1 training, validation, and test split. Following NT (Dalla-Torre et al., 2024), our default tokenizer uses non-overlapping 6-mers. We consider two variants with the NT-v2 architecture: **D3LM**, initialized from pretrained NT weights, and **D3LM-R**, initialized randomly. Unless otherwise stated, both use 50M backbones. We also train a 256 bp variant, D3LM-R@256, to examine generation at matched training lengths. For representation learning, we additionally evaluate human-genome-trained variants, denoted by the suffix **hm**, with both NT and random initialization.

Optimization. We optimize the masked diffusion objective in Eq. (2). The default EPD models use AdamW, a peak learning rate of 8×10^{-5}, cosine decay with 5% warmup, weight decay 0.01, gradient clipping at 1.0, and an effective batch size of 32. Training uses BF16 mixed precision on NVIDIA A800 GPUs. Masking probabilities are sampled independently per sequence, and only masked positions contribute to the denoising loss. The human-genome and 256 bp variants use a learning rate of 10^{-4} and batch size 64. At inference, the default configuration uses 50 denoising steps, temperature 1.1, top-p 0.9, and random selection of positions to unmask.

4.2 Performance of Unconditional Generation

Evaluation setup and baselines. We compare autoregressive models HyenaDNA (Nguyen et al., 2024b), Evo (Nguyen et al., 2024a), and Generator (Wu et al., 2025); latent diffusion via DiscDiff (Li et al., 2024a;b); and a DNA-adapted DPLM (Wang et al., 2024). Truth and uniform random DNA provide reference distributions. Evaluation uses 1,000 sequences per model and length. Evo is fine-tuned at 256 bp, whereas the default comparison models are trained at 2048 bp. Table 1 reports the 2048 bp results and training lengths; model sizes and training budgets differ. Results at 256 bp are provided in Appendix B.

We assess distributional similarity using Sei-based Fréchet distance (SFID) (Chen et al., 2022; Li et al., 2024a) and DNABERT-2-based Fréchet distance (dFID) (Zhou et al., 2024). We also report GC content, G/C ratio, diversity (Div), and nearest-reference novelty (Nov). Div and Nov summarize Hamming distances and serve as descriptive checks, rather than standalone measures of generation quality or memorization. Complementary 3-mer and 6-mer frequency distances are reported in Appendix C.

Table 1: Unconditional generation at 2048 bp. Rows trained at 256 bp evaluate length transfer.

Model	Train bp	Div	Nov	GC content	G/C	SFID ↓	dFID ↓
Truth	–	1530.0 ± 38.3	1470.0 ± 25.9	0.53 ± 0.08	1.06	7.85	1.96
Random	–	1536.0 ± 19.6	1491.5 ± 8.7	0.50 ± 0.01	1.00	122.88	13.80
Generator	2048	1514.0 ± 246.3	1438.0 ± 45.1	0.44 ± 0.16	27.46	599.51	19.27
HyenaDNA	2048	**1534.0 ± 21.9**	1480.0 ± 15.0	0.52 ± 0.06	1.05	29.20	5.82
Evo	256	1534.0 ± 120.3	1468.0 ± 44.4	0.46 ± 0.14	8.56	1360.14	7.18
DPLM	2048	1531.0 ± 19.7	**1484.0 ± 9.8**	0.53 ± 0.01	0.85	95.46	10.51
DiscDiff	2048	1529.0 ± 20.5	1483.0 ± 11.7	0.51 ± 0.03	1.02	64.54	10.07
D3LM	2048	1532.0 ± 28.0	1474.0 ± 20.2	0.50 ± 0.09	1.10	25.32	4.15
D3LM-R	2048	1530.0 ± 28.0	1469.0 ± 23.8	0.54 ± 0.09	1.07	**10.88**	**3.37**
HyenaDNA@256	256	1531.0 ± 77.2	1473.0 ± 35.5	0.52 ± 0.16	0.82	201.77	7.39
HyenaDNA-R@256	256	1513.0 ± 26.5	1474.0 ± 17.1	0.58 ± 0.05	1.13	82.50	10.27
D3LM-R@256	256	1485.0 ± 23.6	1445.0 ± 16.3	0.66 ± 0.03	1.04	125.32	12.82

Div/Nov: median ± distance SD; GC content: mean ± SD; G/C: mean ratio. @256 denotes training at 256 bp; R denotes random initialization.

Results. At 2048 bp, D3LM-R achieves SFID 10.88 and dFID 3.37, compared with 29.20/5.82 for HyenaDNA, 64.54/10.07 for DiscDiff, and 95.46/10.51 for DPLM. Its GC content (0.5352) and G/C ratio (1.0664) are close to the real-sequence reference (0.5265 and 1.0591). NT-initialized D3LM also achieves competitive fidelity, with SFID 25.32 and dFID 4.15, although random initialization performs better on this generation task.

Table 2: **Cell-specific functional design.** Malinois MinGap (↑), reported as mean ± training-seed SD over five seeds for HepG2/K562 and ten for SK-N-SH, using 512 generated sequences per seed. D3LM and SFT are single-checkpoint references. The d1 comparison uses shared hyperparameters with MIPO. Bold indicates the largest mean among the displayed methods.

Method	HepG2	K562	SK-N-SH
D3LM	-0.29	-0.07	-0.27
D3LM + SFT	4.24	2.10	1.05
D3LM + SFT + d1	8.31 ± 0.22	$\mathbf{23.24} \pm 6.15$	5.29 ± 1.60
D3LM + SFT + d1 + MIPO	$\mathbf{24.48} \pm 2.10$	22.78 ± 3.81	$\mathbf{6.75} \pm 1.02$

Table 3: **Motif-prior ablation.** Malinois MinGap (↑), with the same evaluation protocol and training-seed counts as Table 2. All methods start from the same SFT checkpoint. Bold indicates the largest mean among the displayed methods.

Method	HepG2	K562	SK-N-SH
d1	8.31 ± 0.22	$\mathbf{23.24} \pm 6.15$	5.29 ± 1.60
MIPO (shuffled motifs)	8.02 ± 0.08	21.60 ± 4.71	4.48 ± 1.42
MIPO	$\mathbf{24.48} \pm 2.10$	22.78 ± 3.81	$\mathbf{6.75} \pm 1.02$

4.3 Cell-Type-Specific Regulatory Sequence Design

We evaluate whether MIPO improves functional design beyond cell-conditioned SFT and standard diffusion policy optimization. Following the GO-CRE setting (Ma et al., 2026), we generate 200-bp regulatory sequences targeting HepG2, K562, or SK-N-SH. Performance is measured by Malinois MinGap: predicted activity in the target cell minus the highest predicted activity in the other two cells. We compare pretrained D3LM, D3LM with SFT, and SFT followed by d1 (Zhao et al., 2026a), with or without MIPO. The two RL variants share the learning rate, group size, and KL coefficient within each cell, and receive equal training scoring budgets, including rejected MIPO proposals. All final sequences are generated without motif hints or reward-based filtering. Appendix H provides the dataset construction and training details.

Results. As shown in Table 2, SFT increases predicted cell specificity over pretrained D3LM, and RL provides further gains. Under shared hyperparameters, MIPO increases MinGap from 8.31 to 24.48 in HepG2 and from 5.29 to 6.75 in SK-N-SH, corresponding to relative improvements of 194.6% and 27.7%. In K562, MIPO obtains 22.78 compared with 23.24 for d1. These results demonstrate that motif-guided local exploration can improve diffusion policy optimization, with its benefit depending on the target cell.

Ablation of the Motif Prior

The shuffled-motif ablation in Table 3 further examines the contribution of the biological prior. To assess the contribution of motif structure, we retain the MIPO proposal, inpainting, and acceptance procedures but randomly permute the nucleotides in each sampled anchor fragment. This preserves the fragment length and nucleotide composition. The motif-family sampling, placement procedure, proposal count, optimizer settings, and scoring budget are otherwise unchanged.

Shuffling the motif fragments reduces mean MinGap from 24.48 to 8.02 in HepG2 and from 6.75 to 4.48 in SK-N-SH. The K562 difference is smaller. These results support a contribution from the supplied motif structure beyond the fragment's nucleotide composition and the additional proposal-generation procedure.

4.4 Downstream Classification Tasks

Datasets and protocol. We evaluate on the 18 revised NT downstream tasks (Dalla-Torre et al., 2024), covering histone modifications, enhancers, promoters, and splice sites. DNABERT-2 (Zhou et al., 2024), Enformer (Avsec et al., 2021), HyenaDNA, and NT-v2 scores are taken from the published NT benchmark; we additionally report a Caduceus-PH reproduction (Schiff et al., 2024). D3LM variants are adapted using LoRA (Hu et al., 2022) on the query and value projections, with rank 1,

scaling factor 32, and dropout 0.1. We train for up to 1,000 updates with validation-MCC checkpoint selection and early stopping (patience 4). Ten train/validation folds are evaluated on the fixed test set; full results and protocol details appear in Table 5 and Appendix A.

Table 4: Average MCC across the 18 revised NT downstream tasks.

Baseline	Average MCC ↑	D3LM variant	Average MCC ↑
DNABERT-2	0.635	D3LM	0.654
Enformer	0.630	D3LM-R	0.551
HyenaDNA	0.593	D3LM-hm	**0.661**
NT-v2 (50M)	0.634	D3LM-hm-R	0.567
Caduceus-PH	0.516	D3LM-hm-R-BPE	0.539

Unweighted mean across tasks. D3LM variants use 50M backbones; hm denotes human-genome training and R random initialization. Full results are in Table 5.

Results and analysis. Table 4 reports the unweighted mean MCC across the 18 tasks. D3LM achieves an average MCC of 0.654, compared with 0.634 for NT-v2, with gains on splice acceptor/site/donor prediction (0.947/0.945/0.959 versus 0.922/0.928/0.915). Improvements are not uniform across tasks; for example, H3K9ac decreases from 0.538 to 0.508. Human-genome training raises the average to 0.661 for D3LM-hm. In contrast, random initialization gives 0.551 on EPD and 0.567 on the human genome. Thus, pretrained initialization is important for representation quality in these settings, while scratch training yields stronger EPD generation. The gains combine continued diffusion training with additional pretraining data and compute.

4.5 Ablation Studies of Pretraining

Unless otherwise stated, ablations use D3LM-R to generate 2048 bp sequences. Complete results are provided in Appendix E.

Tokenization and model size. Among the tested tokenizers, 6-mer achieves the lowest SFID/dFID (10.88/3.37), compared with 15.69/3.47 for 1-mer, 29.36/4.27 for 3-mer, and 31.99/5.07 for 9-mer (Table 11). Sampling uses 200, 100, 50, and 50 steps, respectively, so these are not equal-cost decoding comparisons. With the NT backbone and random initialization on the human genome, BPE yields average downstream MCC 0.539, below 0.567 for 6-mer. Across 50M, 100M, 250M, and 500M backbones, dFID is 3.37, 3.31, 3.45, and 3.57, respectively (Table 10); increasing model size does not yield a monotonic improvement in this setting.

Sampling strategy. Random-order, MaskGit, entropy, and top-k margin decoding obtain similar SFID values (10.88–11.01). Random order has the lowest SFID, whereas entropy has the lowest dFID (3.12 versus 3.37 for random order). The tested P2 variant performs poorly, with SFID 3830.70 and dFID 40.33 (Table 9). Random order is therefore a competitive default; the preferred strategy depends on the metric.

Denoising steps and temperature. In the step sweep, SFID decreases from 62.40 at one step to 10.92 at 50 steps, with no further improvement at 100–400 steps. Temperature 1.1 gives the lowest SFID among 0.9–1.2. We therefore use 50 steps and temperature 1.1 by default; detailed results are given in Appendices F and G.

5 Conclusion

In this work, we introduced D3LM, which unifies bidirectional representation learning and DNA generation through masked diffusion in discrete token space. By training with variable masking ratios while maintaining bidirectional attention, D3LM addresses fundamental limitations of both BERT-style and autoregressive DNA models. Our work establishes discrete diffusion as a promising paradigm for DNA foundation models. Building on D3LM, MIPO integrates motif-guided local inpainting into reinforcement learning, improving predicted cell-type specificity in HepG2 and SK-N-SH under matched training scoring budgets.

AI USE STATEMENT

Generative AI tools assisted with writing and language polishing, literature search and related-work discovery, and coding for some experiments. The authors take full responsibility for all results reported in this paper.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. *bioRxiv*, pp. 2025–06, 2025.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, pp. 2025–12, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, 54(7):940–949, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Xingyu Chen, Shihao Ma, Runsheng Lin, Jiecong Lin, and Bo Wang. Ctrl-dna: Controllable cell-type-specific regulatory dna design via constrained rl. *Advances in Neural Information Processing Systems*, 38:14037–14065, 2026.

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Lucas Ferreira DaSilva, Simon Senan, Judith F Kribelbauer-Swietek, Zain Munir Patel, Lithin Karmel Louis, Aniketh Janardhan Reddy, Sameer Gabbita, Jonathan D Rosen, Zach Nussbaum, César Miguel Valdez Córdova, et al. Designing synthetic regulatory elements using the generative ai framework dna-diffusion. *Nature Genetics*, pp. 1–15, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Ilias Georgakopoulos-Soares, Chengyu Deng, Vikram Agarwal, Candace SY 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(1):2333, 2023.

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, et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, pp. 1–10, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Connie Huang, Richard W Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail, and Nilah M Ioannidis. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, 2023.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DjJmre5IkP.

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

Long Li, Jingli A Zhang, Marei Dose, Hao Yuan Kueh, Ruzbeh Mosadeghi, Fotini Gounari, and Ellen V Rothenberg. A far downstream enhancer for murine bcl11b controls its t-cell specific expression. *Blood, The Journal of the American Society of Hematology*, 122(6):902–911, 2013.

Zehui Li, Yuhao Ni, William AV Beardall, Guoxuan Xia, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. Discdiff: Latent diffusion model for dna sequence generation. *arXiv preprint arXiv:2402.06079*, 2024a.

Zehui Li, Yuhao Ni, Guoxuan Xia, William Beardall, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. Absorb & escape: Overcoming single model limitations in generating heterogeneous genomic sequences. *Advances in Neural Information Processing Systems*, 37:21949–21978, 2024b.

Mark A Lindsay. Target discovery. *Nature Reviews Drug Discovery*, 2(10):831–838, 2003.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Mingqian Ma, Wanjuan Bu, Guoqing Liu, Yuxuan Liu, Sizhen Liu, Zhen Zhao, Shijie Yao, Qingru Hua, Yujie Zhang, Cuiting Zhong, et al. Reconstructing sequence-grammar trajectories enables interpretable and tunable cis-regulatory element design. *bioRxiv*, pp. 2026–07, 2026.

Aditi T Merchant, Samuel H King, Eric Nguyen, and Brian L Hie. Semantic design of functional de novo genes from a genomic language model. *Nature*, 649(8097):749–758, 2026.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024a.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024b.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KnqiC0znVF.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*, 2024.

Jingyang Ou, Jiaqi Han, Minkai Xu, Shaoxuan Xu, Jianwen Xie, Stefano Ermon, Yi Wu, and Chongxuan Li. Principled rl for diffusion llms emerges from a sequence-level perspective. In *International Conference on Learning Representations*, volume 2026, pp. 4941–4966, 2026.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*, 2025.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

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

Eric R Schuur, Gail A Henderson, Lena A Kmetec, Joseph D Miller, Henry G Lamparski, and Daniel R Henderson. Prostate-specific antigen expression is regulated by an upstream enhancer. *Journal of Biological Chemistry*, 271(12):7043–7051, 1996.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. Generator: a long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025.

Zhao Yang, Bing Su, Chuan Cao, and Ji-Rong Wen. Regulatory DNA sequence design with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhao Yang, Jiwei Zhu, and Bing Su. Space: Your genomic profile predictor is a powerful dna foundation model. In *Forty-second International Conference on Machine Learning*, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:56729–56762, 2026a.

Siyan Zhao, Mengchen Liu, Jing Huang, Miao Liu, Chenyu Wang, Bo Liu, Yuandong Tian, Guan Pang, Sean Bell, Aditya Grover, et al. Inpainting-guided policy optimization for diffusion large language models. In *International Conference on Learning Representations*, volume 2026, pp. 6017–6048, 2026b.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024.

A Additional Experimental Details

Training variants. The default EPD models use 16 training epochs. D3LM-R@256 uses 10 epochs on 256 bp mammalian sequences. Human-genome variants use the hg38 reference-genome dataset with 10 epochs, batch size 64, and learning rate 10^{-4}. The data loader reads 12 kb windows; after tokenization, the collator retains the first chunk of at most 345 tokens per window. The BPE variant uses the DNABERT-2 tokenizer with the same token limit, which does not imply the same number of base pairs per example as 6-mer tokenization. Model inputs are DNA tokens, rather than individual characters except in the 1-mer setting.

Generative baselines. Evo uses `evo-1-8k-base`, adapted on EPD-GenDNA-256 with LoRA rank 16, learning rate 10^{-5}, batch size 8, 3,200 steps over four epochs, and 400 warmup steps. Generator uses `GENERator-eukaryote-1.2b-base`, fully fine-tuned on EPD-GenDNA-2048 with learning rate 10^{-5}, batch size 32, 7,800 steps over five epochs, and 1,000 warmup steps. The default HyenaDNA baseline uses the large 1M-context architecture trained from scratch on EPD-GenDNA-2048. The DPLM baseline uses a 150M architecture, 1-mer tokenization, and scratch training on EPD-GenDNA-2048. Generator samples with temperature 1.5 and top-k 50; the 256 bp Evo and default HyenaDNA samples use temperature 1.0 and top-p 0.9. These comparisons use model-specific adaptation and sampling configurations.

Sequence evaluation. Sequences shorter than the target length are removed, and longer sequences are truncated to the target length before evaluation. We use a fixed shuffle seed of 42 to select at most 1,000 generated sequences. The reference set contains 1,000 real sequences sampled with seed 42; the Truth row uses a second sample with seed 148 from the same reference pool. The two real samples are not explicitly constrained to be disjoint. Fréchet distances are point estimates from the selected sets, rather than averages over independent training runs.

Downstream adaptation. For each task, we partition the official training split into ten shuffled train/validation folds and keep the official test split fixed. D3LM uses a two-layer classification head with a Tanh activation on the first-token representation. We evaluate and save every 100 updates, retaining the checkpoint with highest validation MCC. We report test performance across the ten folds. Caduceus-PH is fully fine-tuned on the same revised benchmark.

B Unconditional Generation at 256 bp

Table 6 reports generation at 256 bp. D3LM-R trained at 2048 bp obtains dFID 3.65, compared with 3.49 for Evo trained at 256 bp. Matching the training length improves D3LM-R@256 to dFID 2.63 and SFID 42.63.

C Evaluation Metrics

Understanding tasks. We report Matthews correlation coefficient (MCC), using its multiclass form for tasks with more than two labels. In the binary case,

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{7}$$

For D3LM variants and Caduceus, tables report mean $\pm$ two standard deviations across the ten fold-specific test scores. Standard deviations use divisor ten; the task average is an unweighted mean. Quoted baseline error terms retain the convention of their source (Dalla-Torre et al., 2024).

Diversity and novelty. Let G denote the evaluated generated sequences, $S \subseteq G$ a subset of up to 128 sequences, and R_{64} a subset of 64 real reference sequences, both selected with seed 42. For Hamming distance d_H, we compute

$$\mathrm{Div} = \mathrm{median}_{x,y \in S,\, x<y}\, d_H(x,y), \tag{8}$$

$$\mathrm{Nov} = \mathrm{median}_{x \in G} \min_{r \in R_{64}} d_H(x,r). \tag{9}$$

Here $x < y$ indexes unordered distinct pairs. Error terms are the standard deviations of the pairwise distances and nearest-reference distances, respectively. Novelty is relative to this reference sample, not an exhaustive search of the training corpus, and cannot establish absence of memorization.

Table 5: Performance on the 18 revised NT downstream tasks (MCC).

(a) Baseline comparison

Task	DNABERT-2	Enformer	HyenaDNA	NT-v2 (50M)	Caduceus-PH	D3LM
H2AFZ	0.490 ± 0.013	0.522 ± 0.019	0.467 ± 0.012	0.501 ± 0.012	0.374 ± 0.040	0.501 ± 0.020
H3K27ac	0.491 ± 0.010	0.520 ± 0.015	0.421 ± 0.010	0.483 ± 0.012	0.398 ± 0.121	0.497 ± 0.019
H3K27me3	0.599 ± 0.010	0.552 ± 0.007	0.550 ± 0.009	0.595 ± 0.004	0.510 ± 0.051	0.590 ± 0.029
H3K36me3	0.637 ± 0.007	0.567 ± 0.017	0.553 ± 0.011	0.607 ± 0.006	0.502 ± 0.095	0.602 ± 0.019
H3K4me1	0.490 ± 0.008	0.504 ± 0.021	0.423 ± 0.016	0.478 ± 0.012	0.365 ± 0.019	0.479 ± 0.016
H3K4me2	0.558 ± 0.013	0.626 ± 0.015	0.515 ± 0.018	0.549 ± 0.014	0.449 ± 0.033	0.537 ± 0.024
H3K4me3	0.646 ± 0.008	0.635 ± 0.019	0.603 ± 0.020	0.627 ± 0.013	0.569 ± 0.032	0.634 ± 0.028
H3K9ac	0.564 ± 0.013	0.593 ± 0.020	0.487 ± 0.025	0.538 ± 0.011	0.434 ± 0.071	0.508 ± 0.026
H3K9me3	0.443 ± 0.025	0.453 ± 0.016	0.419 ± 0.030	0.430 ± 0.018	0.374 ± 0.063	0.464 ± 0.039
H4K20me1	0.655 ± 0.011	0.606 ± 0.016	0.590 ± 0.007	0.637 ± 0.008	0.542 ± 0.051	0.641 ± 0.020
Enhancer	0.517 ± 0.011	0.614 ± 0.010	0.476 ± 0.021	0.511 ± 0.006	0.455 ± 0.042	0.556 ± 0.023
Enhancer types	0.476 ± 0.009	0.573 ± 0.013	0.445 ± 0.009	0.459 ± 0.006	0.439 ± 0.023	0.517 ± 0.011
Promoter all	0.754 ± 0.009	0.745 ± 0.012	0.698 ± 0.011	0.731 ± 0.010	0.705 ± 0.029	0.764 ± 0.019
Promoter n-TATA	0.769 ± 0.009	0.763 ± 0.012	0.729 ± 0.009	0.756 ± 0.010	0.731 ± 0.028	0.761 ± 0.018
Promoter TATA	0.784 ± 0.036	0.793 ± 0.026	0.666 ± 0.041	0.747 ± 0.043	0.803 ± 0.143	0.867 ± 0.081
Splice acceptor	0.837 ± 0.006	0.749 ± 0.007	0.808 ± 0.009	0.922 ± 0.006	0.655 ± 0.141	0.947 ± 0.012
Splice site all	0.855 ± 0.005	0.739 ± 0.011	0.907 ± 0.018	0.928 ± 0.005	0.351 ± 0.034	0.945 ± 0.013
Splice donor	0.861 ± 0.004	0.780 ± 0.007	0.915 ± 0.047	0.915 ± 0.006	0.626 ± 0.103	0.959 ± 0.006
Average	0.635	0.630	0.593	0.634	0.516	0.654

(b) Initialization, training data, and tokenization

Task	D3LM	D3LM-R	D3LM-hm	D3LM-hm-R	hm-R-BPE
H2AFZ	0.501 ± 0.020	0.451 ± 0.017	0.499 ± 0.018	0.461 ± 0.011	0.468 ± 0.030
H3K27ac	0.497 ± 0.019	0.381 ± 0.017	0.494 ± 0.024	0.427 ± 0.031	0.422 ± 0.023
H3K27me3	0.590 ± 0.029	0.547 ± 0.014	0.608 ± 0.014	0.566 ± 0.027	0.569 ± 0.016
H3K36me3	0.602 ± 0.019	0.524 ± 0.009	0.611 ± 0.017	0.538 ± 0.013	0.540 ± 0.014
H3K4me1	0.479 ± 0.016	0.402 ± 0.020	0.487 ± 0.021	0.446 ± 0.018	0.465 ± 0.013
H3K4me2	0.537 ± 0.024	0.488 ± 0.014	0.552 ± 0.024	0.502 ± 0.016	0.511 ± 0.015
H3K4me3	0.634 ± 0.028	0.597 ± 0.025	0.636 ± 0.022	0.600 ± 0.028	0.586 ± 0.032
H3K9ac	0.508 ± 0.026	0.447 ± 0.026	0.526 ± 0.024	0.480 ± 0.031	0.472 ± 0.028
H3K9me3	0.464 ± 0.039	0.315 ± 0.033	0.473 ± 0.048	0.392 ± 0.022	0.398 ± 0.040
H4K20me1	0.641 ± 0.020	0.574 ± 0.020	0.647 ± 0.015	0.596 ± 0.019	0.603 ± 0.021
Enhancer	0.556 ± 0.023	0.431 ± 0.046	0.559 ± 0.033	0.463 ± 0.037	0.472 ± 0.033
Enhancer types	0.517 ± 0.011	0.403 ± 0.018	0.524 ± 0.015	0.442 ± 0.022	0.435 ± 0.024
Promoter all	0.764 ± 0.019	0.722 ± 0.015	0.758 ± 0.019	0.705 ± 0.008	0.717 ± 0.027
Promoter n-TATA	0.761 ± 0.018	0.745 ± 0.014	0.759 ± 0.013	0.724 ± 0.011	0.731 ± 0.014
Promoter TATA	0.867 ± 0.081	0.765 ± 0.042	0.905 ± 0.058	0.662 ± 0.062	0.709 ± 0.105
Splice acceptor	0.947 ± 0.012	0.711 ± 0.090	0.948 ± 0.008	0.719 ± 0.185	0.597 ± 0.037
Splice site all	0.945 ± 0.013	0.609 ± 0.057	0.950 ± 0.008	0.662 ± 0.063	0.502 ± 0.024
Splice donor	0.959 ± 0.006	0.798 ± 0.021	0.961 ± 0.006	0.824 ± 0.047	0.508 ± 0.031
Average	0.654	0.551	0.661	0.567	0.539

D3LM and Caduceus: mean ± 2 SD across ten folds on a fixed test set. Other baselines retain the error terms reported by Dalla-Torre et al. (2024). Average is the task mean. D3LM backbones are 50M; hm denotes human-genome training, R random initialization, and BPE the DNABERT-2 tokenizer.

Table 6: Unconditional generation at 256 bp.

Model	Train bp	Div	Nov	GC content	G/C	SFID ↓	dFID ↓
Truth	–	189.0 ± 8.7	170.0 ± 10.0	0.60 ± 0.12	1.10	31.94	2.29
Random	–	192.0 ± 6.9	176.0 ± 3.3	0.50 ± 0.03	1.01	147.34	5.43
Generator	2048	190.0 ± 27.3	173.0 ± 6.4	0.41 ± 0.17	7.99	421.95	16.65
HyenaDNA	2048	192.0 ± 7.3	176.0 ± 3.9	0.46 ± 0.08	1.05	123.07	4.88
Evo	256	190.0 ± 14.5	170.0 ± 8.9	0.61 ± 0.17	1.52	121.38	3.49
DPLM	2048	192.0 ± 7.0	175.0 ± 3.3	0.53 ± 0.02	0.91	125.16	4.41
DiscDiff	2048	191.0 ± 7.1	176.0 ± 3.7	0.47 ± 0.04	0.96	145.88	6.49
D3LM	2048	192.0 ± 9.9	172.0 ± 6.6	0.54 ± 0.14	1.77	89.35	3.88
D3LM-R	2048	192.0 ± 10.2	173.0 ± 6.9	0.53 ± 0.16	1.26	73.55	3.65
D3LM-R@256	256	189.0 ± 9.7	170.0 ± 6.5	0.62 ± 0.12	1.08	42.63	2.63

Div/Nov: median ± distance SD; GC content: mean ± SD; G/C: mean ratio. All rows use 1,000 sequences. @256 denotes training at 256 bp.

k-mer distances. For $k \in \{3, 6\}$, let $f_k(x)$ be the normalized vector of overlapping k-mer counts. Replacing $d_H(x, y)$ above with $\|f_k(x) - f_k(y)\|_1$ gives k-mer Div and Nov. These quantify sequence-level compositional variation and nearest-reference similarity; they are not distances between pooled dataset-wide frequency distributions. Tables 7 and 8 provide the complete generation comparisons.

GC statistics. For a sequence of length L, we compute

$$\text{GC content} = \frac{n_G + n_C}{L}, \qquad \text{G/C} = \frac{n_G}{n_C}. \tag{10}$$

GC content is reported as mean $\pm$ SD across sequences, and G/C as the mean ratio. The implementation assigns G/C zero when $n_C = 0$. These are compositional statistics, not direct measures of regulatory function.

Feature-space distances. SFID uses Sei features (Chen et al., 2022; Li et al., 2024a); dFID uses the final-layer first-token representation from DNABERT-2 (Zhou et al., 2024). DNABERT-2 inputs are truncated to 512 tokens. For either encoder, the distance between real and generated feature distributions is

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right). \tag{11}$$

Lower values indicate closer distributions in the chosen feature space. Encoder choice, input truncation, and sample size affect these estimates. In particular, Sei's human training domain limits its interpretation on multispecies sequences. We therefore use both feature distances alongside sequence statistics, without interpreting either as experimental validation of biological activity. Truth provides a finite-sample reference, not a universal lower bound.

Table 7: Sequence-level k-mer distances for 256 bp generation.

Model	Train bp	3-mer Div	3-mer Nov	6-mer Div	6-mer Nov
Truth	–	0.73 ± 0.18	0.46 ± 0.06	1.85 ± 0.07	1.70 ± 0.11
Random	–	0.55 ± 0.06	0.53 ± 0.05	1.89 ± 0.04	1.80 ± 0.03
Generator	2048	1.40 ± 0.46	0.96 ± 0.34	1.89 ± 0.28	1.75 ± 0.08
HyenaDNA	2048	0.61 ± 0.12	0.47 ± 0.05	1.86 ± 0.05	1.75 ± 0.04
Evo	256	0.98 ± 0.36	0.49 ± 0.19	1.86 ± 0.16	1.68 ± 0.13
DPLM	2048	0.54 ± 0.06	0.49 ± 0.05	1.88 ± 0.04	1.78 ± 0.03
DiscDiff	2048	0.55 ± 0.07	0.54 ± 0.04	1.85 ± 0.04	1.78 ± 0.03
D3LM	2048	0.91 ± 0.28	0.54 ± 0.15	1.86 ± 0.08	1.71 ± 0.08
D3LM-R	2048	0.86 ± 0.26	0.49 ± 0.08	1.86 ± 0.09	1.71 ± 0.09
D3LM-R@256	256	0.66 ± 0.20	0.44 ± 0.04	1.84 ± 0.09	1.69 ± 0.08

Normalized k-mer frequencies with L_1 distance; median $\pm$ distance SD. Div uses generated pairs; Nov uses nearest references.

Table 8: Sequence-level k-mer distances for 2048 bp generation.

Model	Train bp	3-mer Div	3-mer Nov	6-mer Div	6-mer Nov
Truth	–	0.38 ± 0.15	0.21 ± 0.04	1.34 ± 0.09	1.22 ± 0.06
Random	–	0.20 ± 0.02	0.25 ± 0.02	1.35 ± 0.02	1.34 ± 0.02
Generator	2048	1.38 ± 0.55	1.13 ± 0.42	1.77 ± 0.52	1.62 ± 0.20
HyenaDNA	2048	0.27 ± 0.12	0.20 ± 0.02	1.31 ± 0.05	1.25 ± 0.03
Evo	256	1.10 ± 0.51	0.60 ± 0.53	1.76 ± 0.27	1.44 ± 0.26
DPLM	2048	0.19 ± 0.02	0.21 ± 0.02	1.33 ± 0.02	1.31 ± 0.02
DiscDiff	2048	0.22 ± 0.04	0.25 ± 0.02	1.28 ± 0.03	1.29 ± 0.02
D3LM	2048	0.38 ± 0.18	0.22 ± 0.03	1.28 ± 0.10	1.19 ± 0.04
D3LM-R	2048	0.37 ± 0.15	0.21 ± 0.02	1.29 ± 0.08	1.20 ± 0.05
HyenaDNA@256	256	0.79 ± 0.36	0.32 ± 0.25	1.48 ± 0.22	1.22 ± 0.11
HyenaDNA-R@256	256	0.26 ± 0.09	0.24 ± 0.02	1.29 ± 0.05	1.26 ± 0.04
D3LM-R@256	256	0.21 ± 0.05	0.21 ± 0.02	1.16 ± 0.04	1.15 ± 0.03

Normalized k-mer frequencies with L_1 distance; median $\pm$ distance SD. Div uses generated pairs; Nov uses nearest references.

D SAMPLING STRATEGIES

Starting from masked tokens, the model samples candidate tokens from its temperature-scaled, top-p-filtered distribution p_i at each masked position. A linear time schedule determines how many positions are revealed at each iteration. We compare the following position-selection rules:

$$\text{MaskGit:} \quad c_i = p_i(\hat{x}_i), \qquad \hat{x}_i \sim p_i,$$

$$\text{Entropy:} \quad c_i = \sum_v p_i(v) \log p_i(v),$$

$$\text{Top-}k\text{ margin:} \quad c_i = p_i^{\text{top1}} - p_i^{\text{top2}}.$$

MaskGit-style selection follows the confidence-based decoding principle of Chang et al. (2022). For these three rules, positions are sampled without replacement using Gumbel top-k scores $c_i/0.9 + g_i$, where g_i is standard Gumbel noise. Random-order decoding instead selects masked positions uniformly without replacement. Candidate token sampling is shared across these rules.

The tested P2 variant (Peng et al., 2025) uses a linear target mask count and retains the highest-confidence proposals at currently masked positions. Already generated positions receive infinite confidence and are not rescored. This variant does not reconsider previously generated tokens, so its result applies to this implementation.

Table 9: Sampling strategies for D3LM-R (50M) at 2048 bp, with 50 denoising steps.

Strategy	Steps	Div	Nov	GC content	G/C	SFID ↓	dFID ↓
Truth	–	1530.0 ± 38.3	1470.0 ± 25.9	0.53 ± 0.08	1.06	7.85	1.96
Random DNA	–	1536.0 ± 19.8	1491.0 ± 9.1	0.50 ± 0.01	1.00	122.56	13.72
P2	50	1668.0 ± 520.4	1454.0 ± 37.4	0.52 ± 0.10	12.30	3830.70	40.33
MaskGit	50	1530.0 ± 30.4	1470.0 ± 23.6	0.54 ± 0.09	1.06	10.91	3.40
Entropy	50	1529.0 ± 29.4	1471.0 ± 24.8	0.54 ± 0.09	1.07	10.97	3.12
Top-k margin	50	1530.0 ± 30.8	1469.0 ± 23.6	0.53 ± 0.09	1.06	11.01	3.40
Random order	50	1530.0 ± 28.0	1469.0 ± 23.8	0.54 ± 0.09	1.07	10.88	3.37

Temperature 1.1; top-p 0.9. Random DNA is a reference; Random order is model decoding. Error terms follow Table 1. P2 is defined in Appendix D.

E MODEL AND TOKENIZATION ABLATIONS

Tables 10 and 11 report the complete model-size and tokenization comparisons. The best SFID is attained by the 50M model, while the 100M model has slightly lower dFID (3.31 versus 3.37). The 500M model does not improve either metric. These findings apply to the tested training setting. The tokenizer comparison uses different denoising-step counts.

Table 10: Model-size ablation at 2048 bp with random initialization and 6-mer tokenization.

Parameters (M)	Steps	Div	Nov	GC content	G/C	SFID ↓	dFID ↓
50	50	1530.0 ± 28.0	1469.0 ± 23.8	0.54 ± 0.09	1.07	10.88	3.37
100	50	1525.0 ± 31.5	1464.5 ± 24.1	0.55 ± 0.09	1.04	11.62	3.31
250	50	1525.0 ± 28.0	1468.0 ± 21.6	0.54 ± 0.08	1.06	11.08	3.45
500	50	1529.0 ± 27.5	1473.0 ± 19.6	0.52 ± 0.09	1.10	18.08	3.57

Random-order decoding; temperature 1.1; top-p 0.9. Error terms follow Table 1.

F ABLATION ON DENOISING STEPS

We vary the number of steps from 1 to 400 at temperature 1.1. Table 12 shows a decrease in SFID from 62.40 at one step to 10.92 at 50 steps. Increasing the number of steps to 100, 200, or 400 yields 12.14, 13.39, and 12.29, respectively, without improving on the 50-step setting. We therefore use 50 steps as the default. The step and temperature sweeps report SFID on their evaluation samples; their default value of 10.92 is distinct from the 10.88 obtained in the main generation evaluation.

Table 11: Tokenization ablation at 2048 bp with randomly initialized 50M backbones.

Tokenizer	Steps	Div	Nov	GC content	G/C	SFID ↓	dFID ↓
1-mer	200	1531.0 ± 23.9	1475.0 ± 19.2	0.54 ± 0.06	1.04	15.69	3.47
3-mer	100	1529.0 ± 41.4	1462.0 ± 29.0	0.55 ± 0.12	1.05	29.36	4.27
6-mer	50	1530.0 ± 28.0	1469.0 ± 23.8	0.54 ± 0.09	1.07	10.88	3.37
9-mer	50	1532.0 ± 21.2	1479.0 ± 14.1	0.53 ± 0.05	1.05	31.99	5.07

Random-order decoding; temperature 1.1; top-p 0.9. Step counts differ across tokenizers. Error terms follow Table 1.

Table 12: Denoising-step ablation for D3LM-R at 2048 bp and temperature 1.1.

Steps	Div	Nov	G/C	SFID ↓
Truth	1529.0	1470.0	1.06	7.85
Random	1537.0	1491.0	1.00	122.97
1	1536.0	1488.0	1.04	62.40
5	1534.0	1482.0	1.04	28.96
10	1532.0	1478.0	1.04	17.90
50	1528.0	1471.0	1.07	10.92
100	1528.0	1467.0	1.06	12.14
200	1526.0	1468.0	1.07	13.39
400	1527.0	1467.0	1.06	12.29

Random-order decoding; top-p 0.9. Values are point estimates.

G Ablation on Sampling Temperature

With 50 denoising steps, we vary temperature from 0.9 to 1.2. Table 13 shows the lowest SFID at 1.1 (10.92), compared with 38.50, 14.31, and 14.74 at 0.9, 1.0, and 1.2, respectively. Diversity increases over this range, while the lowest feature distance occurs at an intermediate temperature. We therefore use temperature 1.1 by default.

Table 13: Sampling-temperature ablation for D3LM-R at 2048 bp with 50 denoising steps.

Temperature	Div	Nov	G/C	SFID ↓
Truth	1529.0	1470.0	1.06	7.85
Random	1537.0	1491.0	1.00	122.97
0.9	1513.0	1452.0	1.11	38.50
1.0	1524.0	1464.0	1.08	14.31
1.1	1528.0	1471.0	1.07	10.92
1.2	1532.0	1475.0	1.05	14.74

Random-order decoding; top-p 0.9. Values are point estimates.

H Functional Design: Data and Experimental Details

H.1 Dataset Construction

We construct a cell-conditioned regulatory sequence dataset from public ENCODE MPRA libraries associated with the GO-CRE setting (Ma et al., 2026). The source collection contains 26 distinct design FASTA files linked to 65 experiments across HepG2, K562, and SK-N-SH. We retain valid 200-bp sequences containing only A, C, G, and T. For six libraries containing 230-bp constructs, we remove the two verified 15-bp adapter sequences. Other sequence lengths are excluded. Exact sequence deduplication yields 2,015,540 candidates. This is a reconstruction from public source libraries, rather than the original authors' processed SFT dataset.

We score the candidate pool using the frozen public Malinois predictor (Gosai et al., 2024). For each target cell, we retain sequences whose target-cell prediction is at or above its global 75th percentile

and whose predictions for both other cells are at or below their respective global 25th percentiles. This produces 71,431 HepG2, 26,790 K562, and 16,486 SK-N-SH sequence–cell examples. The resulting cell labels are assigned by predictor-based selection.

We split the selected examples into training and validation sets at an approximately 80:20 ratio using seed 42. A sequence and its reverse complement are assigned to the same split. The resulting sets contain 91,788 training examples and 22,919 validation examples, with per-cell counts in Table 14. The grouping controls exact and reverse-complement overlap; it does not enforce a sequence-homology threshold.

H.2 Supervised Fine-Tuning

All RL experiments start from the same cell-conditioned D3LM SFT checkpoint. We prepend a cell-specific token to each sequence and use the native non-overlapping 6-mer tokenizer, with two single-nucleotide tokens for the remaining bases. Each 200-bp sequence therefore contains 35 DNA tokens, preceded by the classification and cell tokens.

SFT updates all model parameters for three epochs using AdamW, a learning rate of 5×10^{-5}, batch size 32, weight decay 0.01, a linear learning-rate schedule, and a 3% warmup. Training uses FP32 precision. For each example, we uniformly sample the number of masked DNA tokens from 1 to 35 and minimize the mean cross-entropy over masked positions; the prefix tokens remain unmasked. Validation uses a fixed corruption for each sequence. We select the checkpoint with the lowest validation masked-token loss, without using generated-sequence rewards for checkpoint selection.

H.3 Functional Scoring and Evaluation

Malinois predicts activity in K562, HepG2, and SK-N-SH. Following the GO-CRE input convention, we append fixed 200-bp MPRA flanks on both sides of each designed sequence, giving a 600-bp predictor input. Predictions are averaged over this complete input and its reverse complement. The predictor remains frozen throughout data construction, RL training, and evaluation.

The reported endpoint is the unprocessed specificity gap

$$\mathrm{MinGap}(\mathbf{x}, c) = f_c(\mathbf{x}) - \max_{c' \neq c} f_{c'}(\mathbf{x}), \tag{12}$$

where f_c denotes the strand-averaged Malinois prediction for cell c. Higher values indicate greater predicted target-cell specificity. The same predictor supplies training feedback and evaluation scores, so this is an in-domain functional-prediction assessment.

For each trained policy, we generate 512 sequences using 16 diffusion steps, temperature 1.0, top-p 1.0, and random-order unmasking. Generation is conditioned only on the target cell: no motif hints, reward-based filtering, or best-of-N selection are applied. We evaluate the final checkpoint at the fixed training budget. We report the mean and sample standard deviation of per-seed mean MinGap over five training seeds for HepG2/K562 and ten for SK-N-SH. The seed sets are paired across methods, and generation seeds are shared within each comparison cohort.

H.4 Baselines and Controlled Comparison

D3LM. The pretrained generator before cell-conditioned SFT. We evaluate one frozen cohort of 512 unconditional sequences against all three target-cell scoring functions.

D3LM + SFT. The shared cell-conditioned checkpoint described above, without RL. We generate one 512-sequence reference cohort per target cell. D3LM and SFT are single-checkpoint references and therefore do not have a training-seed standard deviation in the table.

D3LM + SFT + d1. We apply d1's diffusion policy-optimization objective (Zhao et al., 2026a) to the SFT checkpoint, using D3LM's native diffusion sampler. Training groups contain ordinary cell-conditioned samples, without motif-guided proposals.

D3LM + SFT + d1 + MIPO. We retain the d1 optimizer and augment training-group construction with the proposals described in Section 3.5. The plain d1 baseline shares MIPO's learning rate, group size, and KL coefficient within each target cell. This comparison isolates the effect of the sampling module under shared hyperparameters.

Table 14: **Data counts and RL settings.** Learning rate, group size, KL coefficient, and scoring budget are shared between plain d1 and MIPO within each cell. Motif and completion-window settings apply to MIPO and its shuffled-motif ablation.

Setting	HepG2	K562	SK-N-SH
SFT training examples	57,168	21,414	13,206
SFT validation examples	14,263	5,376	3,280
RL learning rate	3×10^{-6}	3×10^{-5}	3×10^{-5}
Group size	4	8	4
KL coefficient	0	0	0.03
Motif prior	HNF1B	GATA1, GATA2, RUNX1	NFKB2
Completion window (tokens)	12	12	6
Proposals per group	1	2	1
Training scoring budget	2,560	2,560	1,280
Training seeds	5	5	10

H.5 RL Training Configuration

We optimize a separate policy for each target cell. RL uses full-parameter AdamW updates with BF16 precision, a constant learning rate, zero weight decay, gradient clipping at 1.0, and d1's clipping parameter 0.2. Each rollout batch contains 32 ordinary sequences and supports two policy-update steps.

MIPO additionally scores eight proposals per rollout batch: one per group for HepG2 and SK-N-SH, and two per group for K562. Each proposal inserts one motif and completes a local window around it. The cell-specific settings are listed in Table 14. Motif fragments are sampled from PWM probabilities sharpened with exponent 1.5, padded to token boundaries, and placed at random internal positions with random strand orientation. K562 samples uniformly from its three listed motif families.

The training specificity score uses a scheduled penalty on the strongest off-target prediction. We retain the original 8,000-scored-sequence schedule, with the penalty coefficient increasing from 1 toward 5; early stopping does not compress this schedule. All RL variants use the same GO-CRE reward-processing procedure, including tier-based scaling, running normalization, the configured HepG2 G/C-run penalty, and clipping to $[-2, 10]$. The primary evaluation metric in Eq. (12) uses neither this scheduled coefficient nor these reward transformations.

Hyperparameters and stopping budgets are selected using development runs and fixed before the corresponding confirmation runs. Training budgets count every sequence submitted to the functional scorer, including rejected proposals. Consequently, the 2,560-score budget gives 160 parameter updates for plain d1 and 128 for MIPO; the 1,280-score budget gives 80 and 64 updates, respectively. Evaluation scoring is accounted for separately.