D3LM: A DISCRETE DNA DIFFUSION LANGUAGE MODEL FOR BIDIRECTIONAL DNA UNDERSTANDING AND GENERATION

Zhao Yang*
Gaoling School of Artificial Intelligence
Renmin University of China
yangyz1230@ruc.edu.cn

Hengchang Liu*
Gaoling School of Artificial Intelligence
Renmin University of China
liuhengchang@ruc.edu.cn

Chuan Cao
Zhongguancun Academy
chuancao@bza.edu.cn

Bing Su
Gaoling School of Artificial Intelligence
Renmin University of China
bingsu@ruc.edu.cn

ABSTRACT

Early DNA foundation models adopted BERT-style training, achieving good performance on DNA understanding tasks but lacking generative capabilities. Recent autoregressive models enable DNA generation, but employ left-to-right causal modeling that is suboptimal for DNA where regulatory relationships are inherently bidirectional. We present D3LM (**D**iscrete **D**NA **D**iffusion **L**anguage **M**odel), which unifies bidirectional representation learning and DNA generation through masked diffusion. D3LM directly adopts the Nucleotide Transformer (NT) v2 architecture but reformulates the training objective as masked diffusion in discrete DNA space, enabling both bidirectional understanding and generation capabilities within a single model. Compared to NT v2 of the same size, D3LM achieves improved performance on understanding tasks. Notably, on regulatory element generation, D3LM achieves an SFID of 10.92, closely approaching real DNA sequences (7.85) and substantially outperforming the previous best result of 29.16 from autoregressive models. Our work suggests diffusion language models as a promising paradigm for unified DNA foundation models. We further present the first systematic study of masked diffusion models in the DNA domain, investigating practical design choices such as tokenization schemes and sampling strategies, thereby providing empirical insights and a solid foundation for future research. D3LM has been released at https://huggingface.co/collections/Hengchang-Liu/d3lm.

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.

*Equal contribution.

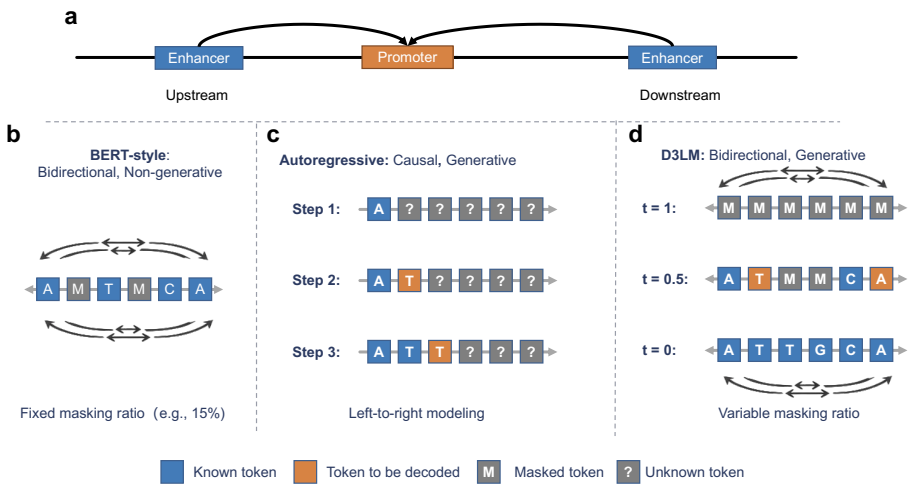


Figure 1: **Comparison of DNA modeling approaches.** **(a)** Enhancers regulate promoters from either upstream or downstream positions, demonstrating bidirectional regulatory relationships in DNA. **(b)** BERT-style models use bidirectional attention but employ fixed masking ratios and lack generative capabilities. **(c)** Autoregressive models generate sequences left-to-right but cannot adjust earlier positions once generated, making it difficult to satisfy global constraints. **(d)** D3LM combines bidirectional modeling with generation through masked diffusion with variable masking ratios, enabling iterative refinement of all positions simultaneously.

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., 2025b), 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 generation 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., 2025b). 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.

To isolate the impact of the training objective, we adopt the NT v2 architecture (Dalla-Torre et al., 2024), resulting in two model variants: D3LM-R, which is randomly initialized with the NT v2 architecture, and D3LM, which is initialized from pre-trained NT v2 weights. Remarkably, D3LM demonstrates improved performance on understanding tasks compared to NT v2 while achieving strong generative capabilities that outperform state-of-the-art DNA generation models, including autoregressive models (Nguyen et al., 2024b;a; Wu et al., 2025) and continuous-space latent diffusion models (Li et al., 2024a). We attribute this to the synergy between unified bidirectional understand-

ing and generation, where both capabilities mutually reinforce each other. These results suggest that discrete diffusion language models represent a promising paradigm for DNA foundation models.

Our contributions can be summarized as follows:

- We propose D3LM, a unified DNA foundation model that combines bidirectional modeling with generative capabilities through masked diffusion in discrete DNA space.
- D3LM demonstrates stronger representation learning than NT v2 of comparable size on downstream understanding tasks, indicating that the masked diffusion generative objective does not degrade representational quality and may even improve it.
- On regulatory element generation, D3LM achieves an SFID of 10.92, closely approaching real biological sequences (7.85) and substantially outperforming autoregressive models (29.16) and continuous-space latent diffusion approaches (62.74), demonstrating the advantages of bidirectional generative modeling directly in the discrete sequence space.
- We present the first systematic analysis of masked diffusion models applied to DNA, focusing on practical design aspects and delivering empirical insights that facilitate future work.

2 METHOD

In this section, we first introduce the bidirectional probabilistic formulation based on discrete masked diffusion models in Section 2.1. We then present the details of our generative training approach in Section 2.2, including model architecture and tokenization strategy. In Section 2.3, we describe the sampling algorithm for sequence generation with D3LM. Finally, in Section 2.4, we explain how D3LM can serve as a representation extractor for fine-tuning on diverse genomic downstream tasks.

2.1 BIDIRECTIONAL PROBABILISTIC FORMULATION

We consider a DNA sequence $\mathbf{x} = (x^1, x^2, \ldots, x^L)$, where each $x^i \in \{A, C, G, T\}$ and L is the sequence length. The sequence is drawn from the true distribution $\mathbf{x} \sim q(\mathbf{x})$, where q represents the real DNA sequence distribution. Our goal is to learn this true distribution through generative modeling using a parametric 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., 2025b) 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. Recent work (Liu et al., 2025; Chen et al., 2025) suggests that unified understanding and generation can mutually benefit model performance, motivating the development of models that combine both capabilities.

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., 2025b; Ye et al., 2025). Unlike the autoregressive formulation in Eq. (1), D3LM defines a

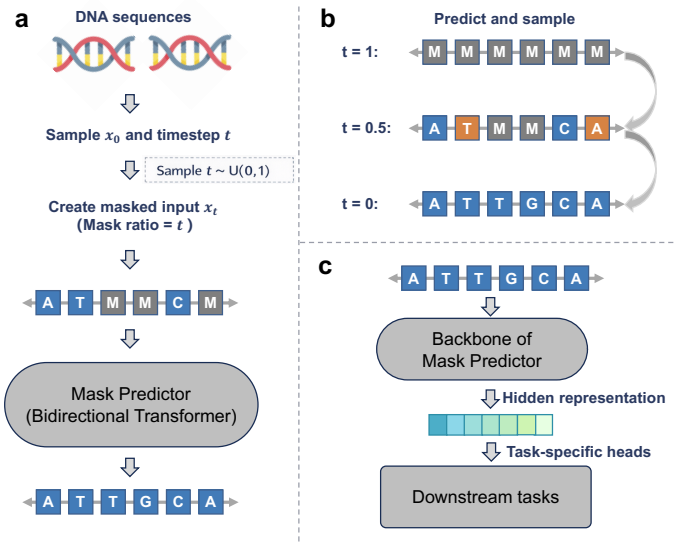


Figure 2: **D3LM framework overview.** **(a)** Training with variable masking ratios sampled from a uniform distribution. The model learns to predict masked tokens using bidirectional attention. **(b)** Iterative generation starting from fully masked sequences, progressively unmasking tokens through repeated sampling. **(c)** Fine-tuning on downstream genomic tasks using frozen or trainable encoder with task-specific heads for promoter classification, histone modification, and splice site prediction.

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 following Nie et al. (2025b).

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:

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

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):

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

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.

2.2 Details of Generative Training

Previous BERT-style and autoregressive DNA models have extensively studied various design choices, including model architecture (Dalla-Torre et al., 2024; Schiff et al., 2024; Nguyen et al.,

2024b), model scale (Dalla-Torre et al., 2024; Nguyen et al., 2024a), and tokenization strategies (Zhou et al., 2024; Wu et al., 2025). Recognizing the importance of these choices, we also conduct systematic studies on key components, with detailed ablation results presented in Section 3.4. Here we present our final design choices.

Model Architecture. D3LM employs the NTv2 (Dalla-Torre et al., 2024) as the backbone for the mask predictor, a Transformer-based (Vaswani et al., 2017) architecture that has proven successful for DNA representation learning. Compared to the original NTv1 (Dalla-Torre et al., 2024), NTv2 introduces key architectural improvements including Rotary Position Embeddings (RoPE) (Su et al., 2024) and SwiGLU activations (Shazeer, 2020). We utilize bidirectional attention to enable bidirectional modeling, which fundamentally differs from GPT-style autoregressive DNA generative models (Nguyen et al., 2024b;a; Wu et al., 2025). We experiment with four model sizes ranging from 50M to 500M parameters, following the scaling configurations of NT v2 (Dalla-Torre et al., 2024).

Tokenization. Tokenization strategy is critical for genomic language models (Wu et al., 2025). We experiment with four different tokenization schemes: 1-mer (single nucleotide), 3-mer, 6-mer, and 9-mer. Following Dalla-Torre et al. (2024) and Wu et al. (2025), we adopt a non-overlapping 6-mer tokenization strategy, which achieves the best balance between vocabulary size and sequence representation. This results in a vocabulary size of 4105, comprising 4096 possible 6-mers ($4^6 = 4096$) plus 9 special tokens including the mask token [M], padding token [PAD], and other control tokens.

2.3 Sampling

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). Surprisingly, we find that **random sampling**—where positions $\mathcal{U}_t$ are selected uniformly at random from $\mathcal{M}_t$—achieves the best generation quality (Table 3). This aligns with the training objective where masks are placed uniformly, and may reflect the non-local nature of DNA regulatory dependencies where confidence scores do not reliably indicate optimal generation order. 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 C.

Temperature scaling. To control generation diversity, we adjust the sampling distribution via temperature scaling:

$$p_\theta^{\text{temp}}(x_0^i|\mathbf{x}_t) \propto p_\theta(x_0^i|\mathbf{x}_t)^{1/\tau}, \tag{5}$$

where $\tau > 0$ is the temperature parameter. Lower temperatures ($\tau < 1$) produce deterministic sequences closely following the training distribution, while higher temperatures ($\tau > 1$) increase diversity at the cost of biological fidelity. We use $\tau = 1.1$ and $T = 50$ steps by default (ablations in Section 3.4).

2.4 Representation learning

D3LM is trained to predict masked tokens at all masking levels, including the original unmasked data (i.e., masking ratio at 0%). As a result, D3LM can simultaneously serve as a DNA sequence

representation learner over massive genomic data, providing useful sequence embeddings for various downstream predictive tasks, such as sequence-level or position-level classification and regression. The sequence embedding can be obtained by simply letting D3LM take as input the given DNA sequence: $\mathbf{h}(\mathbf{x}) = \text{D3LM}_\theta(\mathbf{x}, t = 0) \in \mathbb{R}^{L \times d}$, where d is the embedding dimension and L is the sequence length.

Task-specific adaptation. For a downstream task with labeled data $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where y_n is the task-specific label, we add a lightweight task-specific head on top of the pre-trained encoder. For sequence-level tasks (Dalla-Torre et al., 2024) (e.g., promoter classification), we aggregate token representations using mean pooling: $\bar{\mathbf{h}} = \frac{1}{L}\sum_{i=1}^L \mathbf{h}_i(\mathbf{x})$. For position-level tasks, we directly use the token-level representations $\mathbf{h}_i(\mathbf{x})$. The model is then fine-tuned using task-specific objectives:

$$\mathcal{L}_{\text{task}}(\theta, \phi) = -\sum_{n=1}^{N} \log p_\phi(y_n | \mathbf{h}(\mathbf{x}_n)), \tag{6}$$

where ϕ denotes the parameters of the task-specific head.

3 Experiment

We evaluate D3LM on extensive understanding and generative tasks, starting with the generative training details in Section 3.1. We assess DNA generation capabilities in Section 3.2 and performance on predictive downstream tasks in Section 3.3. Finally, we conduct ablation studies on tokenization strategies, model scale and sampling methods in Section 3.4.

3.1 Generative Training Details

Dataset and Tokenization. We trained D3LM on the **EPD-GenDNA** dataset (Li et al., 2024a), a multi-species genomic corpus comprising 160k DNA sequences across 15 species, each with a fixed length of 2048 base pairs (bp). To focus on mammalian genomics, we curated a subset restricted to mammalian sequences (Li et al., 2024a), which constitute approximately 50% of the total corpus ($\approx$ 80,000 sequences). Adopting the tokenization strategy of the NT (Dalla-Torre et al., 2024), we process DNA sequences using non-overlapping **6-mers**. This yields a vocabulary size of 4,105, comprising all possible 6-mers ($4^6 = 4,096$) and 9 special tokens.

Optimization and Initialization. The model was optimized by minimizing the discrete diffusion objective defined in Eq. (2). We employed the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The learning rate followed a cosine decay schedule, peaking at 8×10^{-5} after a linear warmup phase of the first 5% of training steps. To ensure training stability, we applied a weight decay of 0.01 and a gradient clipping threshold of 1.0. The effective batch size is set to 32 sequences, corresponding to approximately 65,536 **base pairs** per step. We consider two model variants: **D3LM-R**, initialized randomly with the NT v2 architecture, and **D3LM**, initialized from pre-trained NT v2 weights.

Computational Infrastructure. Generative training is performed on a cluster of NVIDIA A800 GPUs with Mixed Precision (BF16) training for computational efficiency. The resulting models are evaluated on both generative and discriminative tasks, with detailed performance metrics provided in Section 3.2 and Section 3.3.

3.2 Performance of Unconditional Generation

Evaluation Setup and Baselines. We evaluated the generative performance on the EPD-GenDNA dataset (Li et al., 2024a;b) by conducting *de novo* generation of 2048bp DNA sequences. The dataset was partitioned into training, validation, and test splits with an 8:1:1 ratio, using the latter solely for subsequent evaluation. To benchmark D3LM against distinct modeling paradigms, we compare it with three classes of baselines: (1) **Autoregressive (AR) models**, including HyenaDNA (Nguyen et al., 2024b)and Evo (Nguyen et al., 2024a), which rely on causal left-to-right modeling; (2) **Latent Diffusion**, represented by DiscDiff (Li et al., 2024a;b), which projects DNA into continuous space; and (3) **Discrete Diffusion**. We adapt the protein-specific DPLM (150M) (Wang et al., 2024) to genomics. Key differences from our approach include its use of single-nucleotide (1-mer) tokenization and a linear $(1 - t)$ loss weighting, rather than our $1/t$ inverse weighting. We also include real validation sequences (Truth) and uniform random sampling (Random) as reference bounds.

Quality is assessed by comparing 1,000 generated sequences to a held-out reference set using four complementary metrics. We report Diversity (Div) and Novelty (Nov) to verify that the model generates varied sequences without memorizing the training corpus. Div is the average pairwise Levenshtein distance among generated sequences, and Nov is the average Levenshtein distance from each generated sequence to its nearest neighbor in the training set. For compositional validity, we use the G/C ratio (G count divided by C count), which reflects strand symmetry under Chargaff's parity rule (typically ≈ 1.0 in natural DNA). Finally, we measure functional fidelity with Sei-based Fréchet Inception Distance (SFID) (Li et al., 2024a): sequences are embedded by Sei (Chen et al., 2022) into a regulatory feature space, and SFID computes the distributional distance between generated and real sequences in that space (lower is better). Metric details are provided in Appendix B.

Results. As detailed in Table 1, D3LM-R achieves superior generation fidelity. While most models exhibit comparable Diversity and Novelty—indicating no trivial memorization—significant performance distinctions emerge in biological plausibility measures. D3LM-R attains an SFID of 10.92, closely approximating the real data baseline (Truth: 7.85) and substantially outperforming both continuous latent diffusion (DiscDiff: 62.74) and autoregressive baselines (HyenaDNA: 29.16). Notably, Evo suffers from severe distributional mismatch (SFID > 500) and distorted GC ratio (0.86). We attribute this to the challenge of adapting large-scale pretrained models to our specific dataset: these models were trained on substantially larger and more diverse genomic corpora, and may require careful fine-tuning or domain adaptation to generalize to the regulatory element generation task. In contrast, D3LM-R preserves a GC ratio (1.07) nearly identical to natural sequences (1.06), suggesting that our discrete diffusion framework effectively captures global genomic constraints that other paradigms miss.

Table 1: Evaluation results of unconditional generation on different sequence lengths.

	Sequence Length = 256				Sequence Length = 2048			
Model	Div (↑)	Nov (↑)	GC	SFID (↓)	Div (↑)	Nov (↑)	GC	SFID (↓)
Truth	189.0	170.0	1.10	31.94	1529	1470	1.06	7.85
Random	192.0	176.0	1.02	148.20	1537	1491	1.00	122.97
HyenaDNA	**192.0**	**176.0**	1.05	122.79	**1533**	1479	1.05	29.16
Evo	191.0	170.0	1.51	123.06	**1533**	1471	0.86	1359.98
DPLM	191.0	175.0	0.91	124.96	1531	**1485**	0.85	95.34
DiscDiff	191.0	**176.0**	0.96	147.90	1527	1482	1.03	62.74
D3LM	191.0	172.0	1.75	85.41	1531	1474	1.10	25.21
D3LM-R	**192.0**	173.0	1.25	**74.14**	1528	1471	1.07	**10.92**

3.3 Downstream Classification Tasks

Datasets. We evaluated representation quality on the NT downstream benchmark suite (Dalla-Torre et al., 2024), including histone modification prediction, enhancer and promoter classification, and splice site prediction. All tasks are based on the revised NT downstream dataset released on HuggingFace. We follow the same dataset splits and evaluation pipeline as NT.

Baselines. We compare D3LM against DNABERT-2 (Ji et al., 2021), Enformer (Avsec et al., 2021), HyenaDNA (Nguyen et al., 2024b), and NT-MSv2 (50M) (Dalla-Torre et al., 2024) under the same dataset and evaluation protocol. Our methods include D3LM (50M), initialized from NT-MSv2 weights and further pre-trained with our discrete diffusion objective, and D3LM-R (50M), which uses the same architecture but starts from random initialization before discrete pre-training.

Fine-tuning details. We finetuned all models using the NT training pipeline implemented with HuggingFace `Trainer` and LoRA (Hu et al., 2022). LoRA is applied to the attention projection modules (`query` and `value`) with rank r=1, α=32, and dropout 0.1. We trained for up to 1,000 update steps with a cosine learning-rate schedule and early stopping (patience 4), selecting the checkpoint that maximizes validation MCC. The final test performance is also reported using MCC, with detailed metric definitions provided in Appendix B.

Results and analysis. Table 2 reports MCC results. D3LM consistently matches or improves upon NT-MSv2, with the most pronounced gains on splice site prediction: D3LM achieves 0.947/0.945/0.959 on splice acceptor/site/donor, exceeding NT-MSv2 (0.922/0.928/0.915) and sub-

Table 2: Performance on downstream NT understanding tasks (MCC ± Std).

TASK	DNABERT-2	ENFORMER	HYENADNA	NT-MSV2 (50M)	D3LM (50M)	D3LM-R (50M)
H2AFZ	0.49 (±0.013)	**0.522 (±0.019)**	0.467 (±0.012)	0.501 (±0.012)	0.501 (±0.020)	0.451 (±0.017)
H3K27AC	0.491 (±0.01)	**0.52 (±0.015)**	0.421 (±0.01)	0.483 (±0.012)	0.497 (±0.019)	0.381 (±0.017)
H3K27ME3	**0.599 (±0.01)**	0.552 (±0.007)	0.55 (±0.009)	0.595 (±0.004)	0.590 (±0.029)	0.547 (±0.014)
H3K36ME3	**0.637 (±0.007)**	0.567 (±0.017)	0.553 (±0.011)	0.607 (±0.006)	0.602 (±0.019)	0.524 (±0.009)
H3K4ME1	0.49 (±0.008)	**0.504 (±0.021)**	0.423 (±0.016)	0.478 (±0.012)	0.479 (±0.016)	0.402 (±0.020)
H3K4ME2	0.558 (±0.013)	**0.626 (±0.015)**	0.515 (±0.018)	0.549 (±0.014)	0.537 (±0.024)	0.488 (±0.014)
H3K4ME3	**0.646 (±0.008)**	0.635 (±0.019)	0.603 (±0.02)	0.627 (±0.013)	0.634 (±0.028)	0.597 (±0.025)
H3K9AC	0.564 (±0.013)	**0.593 (±0.02)**	0.487 (±0.025)	0.538 (±0.011)	0.508 (±0.026)	0.447 (±0.026)
H3K9ME3	0.443 (±0.025)	0.453 (±0.016)	0.419 (±0.03)	0.43 (±0.018)	**0.464 (±0.039)**	0.315 (±0.033)
H4K20ME1	**0.655 (±0.011)**	0.606 (±0.016)	0.59 (±0.007)	0.637 (±0.008)	0.641 (±0.020)	0.574 (±0.020)
ENHANCER	0.517 (±0.011)	**0.614 (±0.01)**	0.476 (±0.021)	0.511 (±0.006)	0.556 (±0.023)	0.431 (±0.046)
ENHANCER TYPES	0.476 (±0.009)	**0.573 (±0.013)**	0.445 (±0.009)	0.459 (±0.006)	0.517 (±0.011)	0.403 (±0.018)
PROMOTER ALL	0.754 (±0.009)	0.745 (±0.012)	0.698 (±0.011)	0.731 (±0.01)	**0.764 (±0.019)**	0.722 (±0.015)
PROMOTER N-TATA	**0.769 (±0.009)**	0.763 (±0.012)	0.729 (±0.009)	0.756 (±0.01)	0.761 (±0.018)	0.745 (±0.014)
PROMOTER TATA	0.784 (±0.036)	0.793 (±0.026)	0.666 (±0.041)	0.747 (±0.043)	**0.867 (±0.081)**	0.765 (±0.042)
SPLICE ACCEPTOR	0.837 (±0.006)	0.749 (±0.007)	0.808 (±0.009)	0.922 (±0.006)	**0.947 (±0.012)**	0.711 (±0.090)
SPLICE SITE ALL	0.855 (±0.005)	0.739 (±0.011)	0.907 (±0.018)	0.928 (±0.005)	**0.945 (±0.013)**	0.609 (±0.057)
SPLICE DONOR	0.861 (±0.004)	0.78 (±0.007)	0.915 (±0.047)	0.915 (±0.006)	**0.959 (±0.006)**	0.798 (±0.021)

stantially outperforming DNABERT-2 and Enformer. This improvement is expected, as NT's masked language modeling objective (fixed 15% masking) can be viewed as a special case of our variable-ratio masked diffusion training. We hypothesize that training on NT's original large-scale genomic corpus with our diffusion objective would yield even stronger performance. In contrast, D3LM-R shows substantially lower performance (e.g., splice site: 0.609 vs. 0.945), which we attribute to the limited scale of our training data (80K sequences) being insufficient to learn robust genomic representations from scratch without leveraging pretrained knowledge.

3.4 ABLATION

In this section, we present ablation studies on tokenization strategies, model scaling, sampling algorithms, denoising steps, and temperature scaling for unconditional generation. Unless otherwise stated, we use the D3LM-R model to generate 2048bp sequences and evaluate them on the EPD-GenDNA dataset, maintaining training and evaluation protocols identical to those in Section 3.2.

Tokenization. As shown in Figure 3b, 6-mer tokenization yields the best overall fidelity (SFID 10.92), outperforming 1-mer (15.77), 3-mer (29.31), and 9-mer (32.01). This suggests that 6-mers provide a favorable inductive bias: they are expressive enough to capture local genomic motifs and short-range dependencies, yet avoid the severe sparsity and rare-transition issues introduced by larger k-mers. This finding is consistent with observations in prior work such as DNABERT (Ji et al., 2021) and Generator (Wu et al., 2025), which also identified 6-mer tokenization as optimal for genomic modeling.

Model scaling. We study scaling behavior by training NT-v2 backbones at 50M, 100M, 250M, and 500M parameters under the same unconditional generation protocol. Figure 3a shows that performance is relatively stable from 50M to 250M (SFID 10.92–11.69), indicating that generation quality is not strongly bottlenecked by parameter count in this range. We hypothesize this is due to information saturation relative to the dataset size, where the larger model overfits to high-frequency noise.

Sampling strategy. We further compare decoding/sampling strategies at inference time, maintaining consistency with previous settings. As shown in Table 3, MaskGit-style sampling, entropy-based sampling, and Topk-M yield consistently high fidelity (SFID 10.93–11.05), while simple random masking achieves the best SFID (10.92) among the tested strategies. In contrast, P2 produces extremely high diversity but suffers from a catastrophic collapse in fidelity (SFID 3847.58) and exhibits severely distorted GC statistics (12.7 vs. $\approx$1.0 for Truth), indicating a failure to preserve basic compositional constraints. Overall, these results suggest that our model is robust to most standard sampling choices, whereas overly aggressive strategies can severely distort the generated distribution. Detailed formulations of these sampling strategies are provided in Appendix C.

Denoising Steps and Temperature. We also investigated the impact of the number of decoding steps and the sampling temperature during the denoising process on the quality of the generated sequences. Please refer to Appendix F and Appendix G for details.

4 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.

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.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

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*, 2025a.

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*, 2025b. URL `https://openreview.net/forum?id=DjJmre5IkP`.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

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.

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, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. In *ICLR*, 2025a.

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*, 2025b. 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.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

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.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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, and Ji-Rong Wen. Synthesizing long-term human motions with diffusion models via coherent sampling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3954–3964, 2023.

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.

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 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. Diffusion models have emerged as powerful generative models in recent years. By defining a forward data corruption process, they learn to reverse the denoising procedure from noisy data (Song et al., 2020; Ho et al., 2020). Initially developed for continuous spaces, these models have achieved promising results in continuous data space, such as image generation (Rombach et al., 2022), audio synthesis (Kong et al., 2020), human motion generation (Yang et al., 2023), and 3D molecular generation (Hoogeboom et al., 2022). Li et al. (2024a) explored encoding discrete DNA sequences into a continuous latent space, followed by performing diffusion in the continuous domain.

Concurrently, researchers began to explore the application of diffusion models in discrete data. D3PM (Austin et al., 2021) pioneered discrete diffusion models, defining discrete state transitions as the corruption process for diffusion models and exploring preliminary results in discrete spaces. Masked diffusion models (Nie et al., 2025a;b; Sahoo et al., 2024) represent a special class of discrete diffusion models with only two states: preserving the original token or transitioning to a masked state. MDMs can be viewed as BERT (Devlin et al., 2019) with variable mask ratios, but are fundamentally modeled from a generative perspective, thus possessing generative capabilities.

A series of theoretical (Lou et al., 2023; Kim et al., 2025a) and empirical (Sahoo et al., 2024; Nie et al., 2025a;b; Ye et al., 2025) works have gradually demonstrated the substantial potential of masked diffusion models for discrete space modeling. Recent efforts in scaling masked diffusion models to billion-scale parameters (Nie et al., 2025b; Ye et al., 2025) have achieved performance comparable to autoregressive models in text generation.

Notably, MDLM (Sahoo et al., 2024) introduced a simplified training objective and demonstrated strong performance on language modeling. MDLM also explored applying this framework to DNA sequences, achieving competitive perplexity on the HG38 genome. However, these experiments were conducted on small-scale models (467K parameters) and evaluated only using perplexity-based metrics, which do not capture biological validity. Our concurrent work NT-v3 (Boshar et al., 2025) fine-tuned their pre-trained model following the MDLM approach for regulatory element generation, but provided only case studies without systematic evaluation. In contrast, our work presents the first comprehensive study of masked diffusion for DNA foundation models. We scale the framework to 500M parameters and conduct systematic evaluation across both understanding tasks and regulatory element generation. Importantly, we introduce biologically meaningful generation metrics including SFID and compositional constraints, and perform extensive ablation studies to understand the role of key components (Section 3.4) of masked diffusion models.

B Evaluation Metrics

We evaluate D3LM on both generative and discriminative tasks using complementary metrics designed to assess different aspects of model performance.

B.1 Metrics for Understanding Tasks

For downstream classification tasks, we report the Matthews Correlation Coefficient (MCC), a balanced metric suitable for datasets with class imbalance. For a binary classification task with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), MCC is computed as:

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

MCC ranges from -1 (perfect disagreement) to $+1$ (perfect agreement), with 0 indicating random prediction. Unlike accuracy, MCC accounts for all four confusion matrix categories and remains informative even when classes are imbalanced.

B.2 Metrics for Generative Tasks

We assess the quality of generated DNA sequences using four complementary metrics that capture different aspects of generation fidelity.

Diversity (Div). Diversity measures the internal variance within the set of generated sequences, quantifying whether the model produces varied outputs rather than repetitive patterns. Given a set of N generated sequences $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, diversity is computed as the average pairwise Levenshtein distance:

$$\text{Div} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{\text{edit}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), \tag{8}$$

where $d_{\text{edit}}(\cdot, \cdot)$ denotes the Levenshtein (edit) distance between two sequences. Higher diversity indicates greater variation in the generated distribution.

Novelty (Nov). Novelty quantifies how different generated sequences are from the training corpus, ensuring the model generates novel sequences rather than memorizing training data. For each generated sequence $\mathbf{x}^{(i)}$, we compute its minimum distance to the training set $\mathcal{D}_{\text{train}}$:

$$\text{Nov} = \frac{1}{N} \sum_{i=1}^{N} \min_{\mathbf{x} \in \mathcal{D}_{\text{train}}} d_{\text{edit}}(\mathbf{x}^{(i)}, \mathbf{x}). \tag{9}$$

Higher novelty indicates that generated sequences are distinct from training examples, demonstrating generalization rather than overfitting.

GC Ratio. The GC ratio measures compositional validity by computing the ratio of Guanine (G) to Cytosine (C) nucleotides. Natural DNA typically maintains approximate strand symmetry (Chargaff's parity rules), with a GC ratio close to 1.0. For a sequence $\mathbf{x}$:

$$\text{GC Ratio} = \frac{\#\{\text{G in } \mathbf{x}\}}{\#\{\text{C in } \mathbf{x}\}}, \tag{10}$$

where $\#\{\cdot\}$ denotes the count of occurrences. Deviations from 1.0 indicate violations of basic compositional constraints in DNA.

Sei-based Fréchet Inception Distance (SFID). SFID measures the distributional similarity between generated and real sequences in a biologically meaningful feature space. We use Sei (Chen et al., 2022), a deep learning model trained to predict over 20,000 chromatin profiles, to map sequences into a high-dimensional regulatory feature space. Given the feature representations of real sequences $\{\mathbf{f}_r^{(i)}\}$ and generated sequences $\{\mathbf{f}_g^{(i)}\}$, SFID is computed as:

$$\text{SFID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\right), \tag{11}$$

where $\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r$ and $\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$ are the mean vectors and covariance matrices of real and generated feature distributions, respectively, and $\text{Tr}(\cdot)$ denotes the matrix trace. Lower SFID indicates that generated sequences possess regulatory properties closer to natural DNA, capturing functional similarity beyond sequence-level statistics.

C Sampling Strategies

In the discrete diffusion generative process, the model iteratively recovers masked tokens over T timesteps. At each step t, the model predicts the probability distribution $p_\theta(x_0|x_t)$ for all currently masked positions. Let $\mathcal{M}_t$ denote the set of indices that are masked at step t. The core difference between sampling strategies lies in two aspects: (1) how to select the candidate token $\hat{x}_i$ for each masked position $i \in \mathcal{M}_t$, and (2) how to compute a confidence score c_i to decide which tokens to unmask (keep) for the next step $t-1$.

We denote the predicted logits at position i as $z_i \in \mathbb{R}^V$, where V is the vocabulary size. The probability distribution is $p_i = \text{Softmax}(z_i/\tau)$, where τ is the temperature.

1. MaskGit (Confidence-based Sampling) Following Chang et al. (2022), this strategy selects the most probable token and uses its probability as the confidence measure.

$$\begin{aligned}\hat{x}_i &= \arg\max p_i \\ c_i &= \max(p_i)\end{aligned}$$

In our implementation, we allow sampling $\hat{x}_i$ from the distribution p_i rather than strictly taking the argmax if $\tau > 0$, but the confidence remains the probability of the selected token. At each step, we unmask a subset of tokens with the highest confidence scores according to a linear schedule.

2. Entropy-based Sampling This strategy measures the uncertainty of the model's prediction using Shannon entropy. Lower entropy indicates higher confidence.

$$\begin{aligned}\hat{x}_i &\sim p_i \\ c_i &= -H(p_i) = \sum_{v=1}^{V} p_i^{(v)} \log p_i^{(v)}\end{aligned}$$

Tokens with lower entropy (higher negative entropy) are prioritized for unmasking. This approach favors positions where the model is certain about the distribution, even if the top-1 probability is not extremely high.

3. Top-k Margin (Topk-M) This metric defines confidence as the margin between the most likely token and the second most likely token.

$$\begin{aligned}\hat{x}_i &\sim p_i \\ c_i &= p_i^{(\text{top1})} - p_i^{(\text{top2})}\end{aligned}$$

where $p_i^{(\text{top1})}$ and $p_i^{(\text{top2})}$ are the largest and second-largest probabilities in p_i. A large margin indicates that the model clearly distinguishes the best candidate from alternatives, reducing the risk of sampling ambiguity.

4. Random Sampling In this baseline strategy, we disregard confidence scores entirely for the selection of unmasking positions.

$$\hat{x}_i \sim p_i$$
$$c_i \sim \text{Uniform}(0, 1)$$

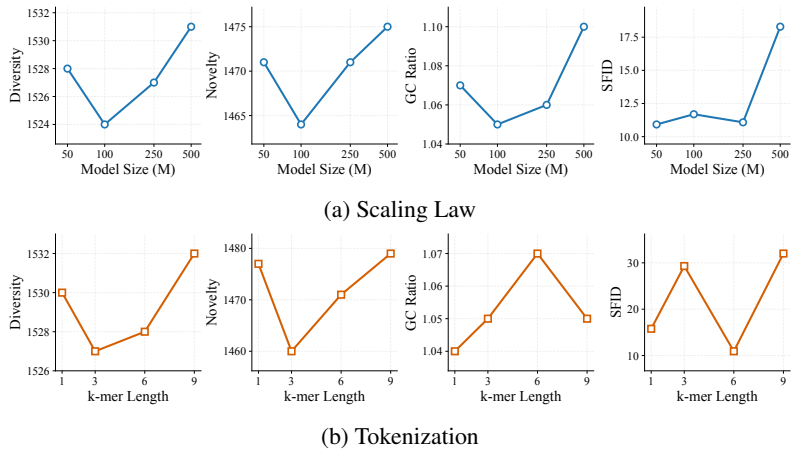
Tokens to be unmasked are selected uniformly at random from the set of currently masked tokens $\mathcal{M}_t$. This serves as an ablation to verify the necessity of confidence-based scheduling.

5. P2 (Proposal-based Probability) The P2 strategy (Peng et al., 2025) employs a "re-masking" mechanism distinct from the standard parallel decoding. Instead of monotonically reducing the mask set, P2 allows previously unmasked tokens to be re-masked if their confidence drops. At step t, for *all* positions (both currently masked and unmasked), we compute confidence based on the current prediction (typically using the max probability). We then select the top k_t tokens with the highest confidence to keep unmasked, where k_t follows the diffusion schedule.

$$\mathcal{U}_{t-1} = \text{TopK}_{i \in \text{All Indices}}(c_i, k_t)$$

This allows the model to "correct" earlier mistakes by re-masking them, but as shown in our experiments (Table 3), this instability can lead to catastrophic failure in maintaining long-range genomic constraints (e.g., GC ratio) when applied to DNA generation.

D Ablation on Tokenization



(a) Scaling Law

(b) Tokenization

Figure 3: Analysis of Model Properties. (a) Illustration of the scaling law behavior. (b) Visualization of the tokenization process.

E Ablation on Sampling Strategy

Table 3: Abaltion study of sampling strategy.

STRATEGY	DIV (↑)	NOV (↑)	GC	SFID (↓)
TRUTH	1529	1470	1.06	7.85
RANDOM	1537	1491	1.00	122.97
P2	**1662.0**	1451.0	12.7	3847.58
MASKGIT	1531.0	1469.0	1.06	10.93
ENTROPY	1530.0	**1472.0**	1.07	10.99
TOPK-M	1529.0	1468.0	1.05	11.05
RANDOM	1528	1471	1.07	**10.92**

F Ablation on Denoising Steps

We investigate the impact of the number of denoising steps T on generation quality by varying T from 1 to 400. As shown in Table 4, increasing the number of steps initially improves biological fidelity significantly, with SFID dropping sharply from 62.40 at $T = 1$ to an optimal 10.92 at $T = 50$. Interestingly, further increasing the steps to 100 and beyond leads to a slight degradation in fidelity (SFID rises to $\approx$ 12–13), suggesting a potential over-refinement issue where the model may drift from the natural manifold or over-smooth the distribution. Consequently, we select $T = 50$ as the default setting, striking the best balance between generation quality and computational efficiency.

Table 4: Ablation study of denoising steps.

Steps (T)	Div (↑)	Nov (↑)	GC	SFID (↓)
Truth	1529	1470	1.06	7.85
Random	1537	1491	1.00	122.97
1	**1536**	**1488**	1.04	62.40
5	1534	1482	1.04	28.96
10	1532	1478	1.04	17.90
50	1528	1471	1.07	**10.92**
100	1528	1467	1.06	12.14
200	1526	1468	1.07	13.39
400	1527	1467	1.06	12.29

G Ablation on Sampling Temperature

We further analyze the effect of sampling temperature τ on the trade-off between diversity and fidelity, testing temperatures from 0.9 to 1.2. Table 5 demonstrates that $\tau = 1.1$ yields the optimal performance, achieving the lowest SFID (10.92) while maintaining diversity (1528) and GC ratio (1.07) remarkably close to the ground truth (1529 and 1.06, respectively). Lower temperatures (e.g., 0.9, 1.0) reduce diversity and degrade fidelity (SFID 38.50 at $\tau = 0.9$), likely due to the model focusing on repetitive, high-probability modes. Conversely, increasing τ to 1.2 boosts diversity further (1532) but leads to a slight drop in fidelity (SFID 14.74). We therefore adopt $\tau = 1.1$ as the default temperature for our main experiments.

Table 5: Ablation study of sampling temperature.

Temp (τ)	Div (↑)	Nov (↑)	GC	SFID (↓)
Truth	1529	1470	1.06	7.85
Random	1537	1491	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	1471	1.07	**10.92**
1.2	**1532.0**	**1475.0**	1.05	14.74