From Circles to Signals: Representation Learning on Ultra-Long Extrachromosomal Circular DNA

Jien Li[1,*], Zhenke Liu[2,*], Ziqi Zhang[2,*], Jiaqi Zhang[2] and Ritambhara Singh,[2,3,†]

[1]Department of Molecular Biology, Cell Biology and Biochemistry , Brown University, Providence, RI 02906
[2]Department of Computer Science, Brown University, Providence, RI 02912
[3]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

Abstract

Extrachromosomal circular DNA (eccDNA) is a covalently closed circular DNA molecule that plays an important role in cancer biology. Genomic foundation models have recently emerged as a powerful direction for DNA sequence modeling, enabling the direct prediction of biologically relevant properties from DNA sequences. Although recent genomic foundation models have shown strong performance on general DNA sequence modeling, their application to eccDNA remains limited: existing approaches either rely on computationally expensive attention mechanisms or truncate ultra-long sequences into kilobase fragments, thereby disrupting long-range continuity and ignoring the molecule's circular topology. To overcome these problems, we introduce eccDNAMamba, a bidirectional state space model (SSM) built upon the Mamba-2 framework, which scales linearly with input sequence length and enables scalable modeling of ultra-long eccDNA sequences. eccDNAMamba further incorporates a circular augmentation strategy to preserve the intrinsic circular topology of eccDNA. Comprehensive evaluations against state-of-the-art genomic foundation models demonstrate that eccDNAMamba achieves superior performance on ultra-long sequences across multiple task settings, such as cancer versus healthy eccDNA discrimination and eccDNA copy-number level prediction, and real versus pseudo-eccDNA discrimination. Moreover, the Integrated Gradient (IG) based model explanation indicates that eccDNAMamba focuses on biologically meaningful regulatory elements and can uncover key sequence patterns in cancer-derived eccDNAs. Overall, these results demonstrate that eccDNAMamba effectively models ultra-long eccDNA sequences by leveraging their unique circular topology and regulatory architecture, bridging a critical gap in sequence analysis. Our codes and datasets are available at `https://github.com/zzq1zh/eccDNAMamba`.

*These authors contributed equally to this work. Alphabetical order.
†Corresponding author. Email: ritambhara@brown.edu

1 Introduction

Extrachromosomal circular DNA (eccDNA) is widespread in eukaryotic cells and often derives from the genome [1]. Specifically, eccDNAs possess a covalently closed, head-to-tail circular structure, introducing *wrap-around dependencies* that play important roles in DNA rearrangement. EccDNA is especially prominent in cancer, with sequence lengths that often exceed 500 kilobases and can even reach megabase scale [2]. These ultra-long molecules frequently harbor oncogenes and distal regulatory sequences that influence tumor evolution, therapeutic resistance, and intratumor heterogeneity [3, 4].

Recently, foundation models trained directly on large-scale DNA sequences have emerged as powerful tools for learning the underlying functional structures of the genome. Despite their success in DNA sequence modeling, existing genomic foundation models still face significant challenges when modeling the ultra-long and circular structure of eccDNA molecules. First, in cancer cells, the eccDNA sequences often span tens of kilobases to several megabases [5] [6], containing entire genes and regulatory regions. Modeling such ultra-long sequences requires a framework that can efficiently capture long-range contexts across extensive genomic sequences. However, the state-of-the-art attention-based foundation models, such as DNABERT-2 [7] and Nucleotide Transformer [8], while powerful for short sequences, scale poorly with input length due to the quadratic complexity of the attention mechanism. This computational bottleneck makes their direct application to ultra-long eccDNA infeasible. In practice, these models typically rely on truncating input sequences into kilobase fragments, disrupting the continuity of the original molecule and eliminating critical long-range contextual information, potentially leading to incomplete or misleading downstream analyses.

To meet the long-range context modeling requirement, other architectures replace attention with more computationally efficient operators. But these approaches still break the circular continuity at the head–tail junction of eccDNA. For example, HyenaDNA and its variant [9, 10] combine information across far-apart sequence positions without computing every intermediate interaction to achieve efficient scaling on very long sequences, yet its unidirectional computation limits the ability to model circular topology and loses dependencies spanning the head-to-tail connection. Caduceus [11] adopts bidirectional state-space models (SSMs) that read sequences step-by-step while maintaining a running summary to capture long-range patterns efficiently with base-level (per-nucleotide) tokenization, yielding fine-grained, symmetric context at single-base resolution. While SSMs provide linear-time sequence processing, making them well-suited for long inputs, the per-base tokenization expands the sequence length drastically. As a result, even with linear computation complexity, memory constraints often necessitate sequence truncation. Consequently, no existing method simultaneously achieves efficient modeling of ultra-long eccDNA sequences while preserving their intrinsic circular topology.

In this paper, we aim to resolve these modeling gaps by introducing eccDNAMamba [12], the first bidirectional state-space models tailored to eccDNA. eccDNAMamba first applies Byte-Pair Encoding (BPE) [13, 14] to compress repetitive nucleotide patterns into compact tokens. It then fuses forward and reverse Mamba-2 [15] passes to obtain full-context representations in linear time. Finally, through a circular augmentation implemented by appending a short snippet of the sequence's beginning to its end, eccDNAMamba explicitly preserves the circular junction and wrap-around dependencies. This augmentation ensures topology-awareness while enabling eccDNAMamba to learn the representation of eccDNA sequences spanning megabases without truncation efficiently. We also systematically define and evaluate the core predictive tasks that make eccDNA modeling both biologically meaningful and computationally challenging.

To evaluate the effectiveness of our model, we preprocess data from CircleBase, eccDNAdb, and eccDNA Atlas [5, 6, 16] and establish the EccDNA Multi-Task Benchmark. This benchmark spans three biologically relevant classification tasks. The cancer-versus-healthy eccDNA discrimination task assesses whether an eccDNA sequence originates from cancer tissues. The

copy-number level prediction task evaluates the degree of oncogene amplification [2, 17]. The real-versus-pseudo eccDNA discrimination task examines whether a model can distinguish genuine eccDNA molecules from length-matched linear genomic fragments based solely on sequence-intrinsic signals. This final task also validates the predictability of short eccDNA formation and extends the evaluation to both short and ultra-long eccDNA molecules [18]. Our results show that eccDNAMamba outperforms existing state-of-the-art foundation models for all these tasks. For cancer-versus-healthy eccDNA discrimination, eccDNAMamba surpasses the best-performing baseline by 1.2% on short sequences and 2.4% on ultra-long sequences. For eccDNA copy-number level prediction, eccDNAMamba exceeds the strongest baseline by 6.4% on average. For real–versus–pseudo eccDNA discrimination, eccDNAMamba outperforms the top baseline by 1.8%. Beyond accuracy, eccDNAMamba offers linear-time scaling and a low, stable memory footprint, enabling efficient eccDNA modeling.

We further demonstrate that eccDNAMamba can provide biological insights on ultra-long eccDNA sequences, unlike existing eccDNA tools that either overlook sequence-level modeling [19] or are restricted to short inputs [18]. Using Integrated Gradients (IG), our eccDNAMamba model highlights which regions of the ultra-long eccDNA sequences the model focuses on, therefore enabling direct biological insight into how model predictions relate to circular topology and regulatory features. These analyses reveal that eccDNAMamba focuses on regulatory and transposable elements that amplify oncogenic programs. Motif discovery on high-attribution regions recovers known oncogenic transcription factor families and uncovers novel sequence patterns of cancer eccDNA. The model effectively leverages the unique topology and regulatory architecture of eccDNA, providing a representation-learning framework for predicting and analyzing properties of eccDNA from sequence for the biology community.

2 Method

Overview. We frame eccDNA analysis as a sequence modeling and prediction task: given an ultra-long eccDNA sequence, the goal is to learn representations that capture both its internal architecture and its dependencies, and to use these representations to support downstream predictions. (Fig. 1). Each raw DNA sequence, composed of the four nucleotide bases (A, T, C, and G), is first tokenized using Byte-Pair Encoding (BPE), which merges frequent nucleotide patterns into compact tokens to reduce sequence length while preserving biological meaning (Sec. 2.1). The model then applies circular augmentation by appending the first 64 tokens of each sequence to its end(Sec. 2.2), explicitly exposing the head–tail junction and enabling the model to learn wrap-around dependencies (Sec. 2.3). The model employs Mamba-2 encoders to process the augmented sequence from both directions—one pass moving forward and the other moving backward. The information gathered from the two directions is then integrated into a single, unified representation of the entire eccDNA sequence(Sec. 2.3).

During pretraining, eccDNAMamba is optimized with a span-masked language modeling objective, where random contiguous spans of tokens are masked, and the model learns to reconstruct the missing segments from the surrounding context (Sec. 2.4). This pretraining scheme encourages the model to capture long-range dependencies and circular topology in an unsupervised manner, providing general-purpose representations that can later be fine-tuned for downstream tasks such as cancer classification, copy-number prediction, or real versus pseudo-eccDNA discrimination. We describe all these steps below.

2.1 Efficient Tokenization

To process genomic sequences with modern deep learning models, we need to convert raw nucleotide sequences into discrete tokens—a process known as tokenization. Effective tokenization defines the basic units of representation, directly influencing representation learning. We adopt

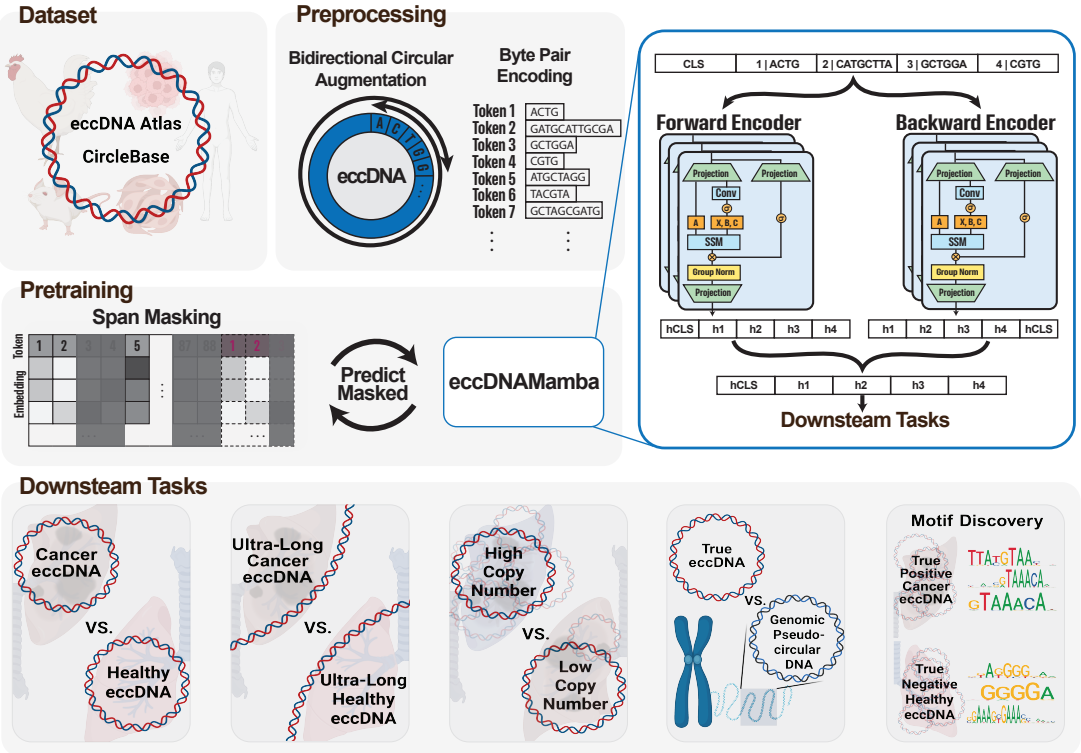


Figure 1: eccDNAMamba processes eccDNA through a topology-aware pipeline combining BPE tokenization, circular augmentation, and bidirectional Mamba-2 encoding to capture wrap-around dependencies. The forward and reverse representations are fused via a shared MLP to generate unified sequence embeddings. During pretraining, a span-masked language modeling objective enables the model to learn long-range and circular dependencies for downstream genomic tasks.

Byte Pair Encoding (BPE) [13, 14] to construct vocabulary:

$$(a^*, b^*) = \arg\max_{(a,b)} \text{freq}_{C_t}(a, b), \qquad \text{where } C_0 = \{\,[A, T, C, G, G, T, A, \ldots]\}; \tag{1}$$

$$\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{a^* b^*\}, \qquad \text{where } \mathcal{V}_0 = \{A, T, C, G\} \tag{2}$$

Here, C_t denotes the eccDNA corpus and $\mathcal{V}_t$ the vocabulary consisting of the nucleotide patterns at iteration t. The symbols a and b represent two adjacent nucleotide tokens in the corpus, and $\text{freq}_{C_t}(a, b)$ denotes their co-occurrence frequency. The BPE algorithm iteratively merges the most frequent adjacent pair of symbols (a^*, b^*), where the superscript * indicates that the pair is selected as the most frequent in the current iteration. The merged token a^*b^* is then added to the vocabulary. Fixed-length k-mer tokenization, as adopted in models such as DNABERT [20] and the Nucleotide Transformer [8], treats all substrings of length k equally and leads to exponential vocabulary growth as k increases. In contrast, modern genomic models such as DNABERT-2 employ BPE to alleviate this limitation [7]. Following this direction, we use BPE to learn variable-length nucleotide patterns based on frequency, allowing the tokenizer to capture recurrent nucleotide patterns while keeping sequence length compact and enabling efficient modeling of ultra-long eccDNA.

2.2 Circular Data Augmentation

eccDNAs are circular molecules, yet standard linear representations ignore potential head–tail dependencies. Prior studies suggest the ends of eccDNA molecules often contain repetitive ele-

ments involved in circularization [21]. To preserve this circular topology, we introduce a data augmentation strategy that appends the first 64 tokens of each sequence to its end. Formally, for a sequence with L tokens, we construct the augmented input as $\tilde{x} = (x_1, x_2, \ldots, x_L, x_1, x_2, \ldots, x_{64})$. This lightweight operation preserves the continuity between the head and tail of the eccDNA sequence, enabling the model to capture long-range dependencies across the circular junction.

We choose 64 tokens because they provide sufficient contextual coverage for most circular junctions while also enabling efficient large-scale pretraining (see Appendix A.8 for more details). To achieve successful pretraining, we empirically chose to append approximately 25% of each sequence to its end. Most eccDNA sequences used for pretraining contain fewer than 256 tokens, so appending 25% corresponds to 64 tokens, aligning with our design choice. We also empirically validate the benefits of this augmentation in Section 3.2. Across all downstream tasks, integrating this circular augmentation leads to consistent performance improvements.

2.3 Bidirectional Mamba-2 Encoding

We aim to build a sequence encoder that can process eccDNA in both directions and remain aware of its circular structure while staying efficient enough to handle ultra-long inputs. To achieve this, we adopt Mamba-2 [15], a state-space model that processes sequences one token at a time instead of comparing all pairs of positions as in attention. This formulation enables linear-time computation and makes it feasible to scan megabase-scale DNA.

Formally, at each position t, the model maps the input token x_t to an embedding

$$s_t = \text{Embedding}(x_t), \tag{3}$$

$$u_{t+1} = A_t u_t + B_t s_t, h_t = C_t u_t + D_t s_t, \tag{4}$$

where u_t summarizes the information accumulated up to step t, and h_t is the contextual output for that position. These updates rely on fixed-cost operations, allowing the encoder to move through the sequence in a single pass, much like scanning along a chromosome. Meanwhile, the dynamically generated matrices A_t, B_t, C_t, D_t introduce nonlinear mixing that enables the model to capture complex sequence patterns while preserving the efficiency of state-space recurrence.

eccDNAMamba uses two independent Mamba-2 encoders that scan the sequence in opposite directions, allowing the model to capture both long-range interactions and circular dependencies in ultra-long eccDNA while maintaining linear-time computation. The detailed mathematical formulation of Mamba-2 is provided in Appendix A.2.

As shown in Fig. 1, each eccDNA sequence is first tokenized into nucleotide pattern embeddings using the BPE vocabulary described in Section 2.1, which reduces redundancy and shortens the effective sequence length. The model then applies the circular augmentation described in Section 2.2, where the first 64 tokens are appended to the end of the sequence to explicitly expose the head–tail junction. The resulting augmented sequence $\tilde{\mathbf{x}} = [x_1, x_2, \ldots, x_L, x_1, \ldots, x_{64}]$ is fed into the forward and reverse Mamba-2 encoders.

The forward encoder scans the augmented sequence in its natural order, while the reverse encoder reads the reversed sequence:

$$\overrightarrow{\mathbf{h}} = \overrightarrow{\text{Mamba}}(\overrightarrow{\tilde{\mathbf{x}}}), \ \overleftarrow{\mathbf{h}} = \overleftarrow{\text{Mamba}}(\overleftarrow{\tilde{\mathbf{x}}}). \tag{5}$$

Each h_i represents the contextual embedding of token x_i, capturing both local patterns and long-range dependencies. The two directional streams are then merged through a shared projection layer to form unified sequence representations:

$$\mathbf{h} = \text{MLP}([\overrightarrow{\mathbf{h}}, \text{FLIP}(\overleftarrow{\mathbf{h}})]). \tag{6}$$

Here, the FLIP operator reverses the temporal order of the reverse encoder output, ensuring that forward and reverse positions are aligned, particularly at the head–tail junction. This

alignment preserves symmetric contextual information and maintains awareness of the circular topology. After the large-scale pretraining phase, we further finetune the model on a specific biological task. This finetuning helps the model adapt its general understanding of eccDNA to the particular problem.

2.4 Pretraining

During pretraining, we further enhance representation learning by masking contiguous spans of tokens rather than isolated ones [22]. Each span consists of roughly three consecutive tokens and covers about 15% of the sequence in total. 80% of masked tokens are replaced with `[MASK]`, 10% are replaced with random tokens, and 10% remain unchanged. Formally, given a augmented sequence $\tilde{\mathbf{x}} = [x_1, x_2, \ldots, x_L, x_1, \ldots, x_{64}]$ and a set of masked spans M, the model predicts the original tokens by minimizing $\mathcal{L}$, the span-masked cross-entropy loss:

$$\mathcal{L} = -\sum_{i \in \mathcal{M}} \log p_\theta(x_i \mid x_{\setminus \mathcal{M}}) \tag{7}$$

Compared to standard pretraining loss, our method jointly reconstructs tokens within a span, capturing intra-span dependencies and contextual coherence, properties crucial for long-range circular dependencies.

3 Experimental Setup

3.1 EccDNA Multi-Task Benchmark

To comprehensively assess model performance, we introduce the EccDNA Multi-Task Evaluation benchmark and assemble a unified eccDNA corpus by standardizing entries from CircleBase, eccDNA Atlas, and eccDNAdb [5, 16, 6], with additional details provided in Appendix A.1.

Cancer vs. healthy eccDNA discrimination. These tasks evaluate whether the sequence of an eccDNA molecule carries predictive signals associated with oncogenic activity. We consider two length regimes: short eccDNAs (<10 kbp) and ultra-long eccDNAs (10–200 kbp). This design separates baseline performance from the ability to model ultra-long eccDNAs.

EccDNA copy-number level prediction. In this task, the model is asked to infer whether the copy number status is low or high. Copy-number alterations (CNAs) represent a major mechanism of oncogene amplification, and eccDNAs are known to frequently harbor high-copy oncogenic elements. We adopt two biologically motivated thresholds (copy number>4.5) and (copy number>10) to define high-level copy number states [2, 17].

Real vs. pseudo-eccDNA discrimination. This task evaluates whether a model can distinguish genuine eccDNA elements from pseudo-eccDNA sequences that are randomly extracted from linear genomes and matched to the same length distribution, thereby testing whether the model can detect sequence-intrinsic signals unique to eccDNA rather than relying on length cues, and further validating that the formation of eccDNA molecules is predictable [18].

Task	Sequence Length	Num. Train	Num. Eval	**Dataset**
Cancer vs. healthy eccDNA discrimination (short)	<10k	16000	4000	CircleBase [5]
Cancer vs. healthy eccDNA discrimination (ultra-long)	10k–200k	5011	811	CircleBase [5]
EccDNA copy-number level prediction	<1M	760	192	eccDNAdb [6]
Real vs. pseudo-eccDNA discrimination (*Homo sapiens*)	<1M	3200	800	eccDNA Atlas [16]
Real vs. pseudo-eccDNA discrimination (*Gallus gallus*)	<1M	3200	800	eccDNA Atlas [16]
Real vs. pseudo-eccDNA discrimination (*Arabidopsis thaliana*)	<1M	873	219	eccDNA Atlas [16]

Table 1: Overview of the three main tasks and cross-species extension. Each task is summarized by sequence length, training/evaluation splits, and dataset resource.

The dataset is consistently split with a train-to-test ratio of 80%:20%, ensuring comparability across tasks and models. By standardizing sampling, task definition, and evaluation metrics, the EccDNA Multi-Task Evaluation benchmark provides a rigorous and reproducible framework for testing sequence modeling approaches on eccDNA.

3.2 Evaluation Metrics

We employed the standard cross-entropy loss to train all classification heads. Given an input augmented sequence $\tilde{x}$, we first obtain unified sequence representations $\mathbf{h}$ using bidirectional Mamba-2 encoder (Sec. 2.3) and compute a mean-pooled representation h_{pool}. The task-specific classification head f_θ then produces a predicted probability distribution $\hat{p} = \text{softmax}(f_\theta(h_{\text{pool}}))$ for the ground-truth label $y \in \{1, \ldots, C\}$. The loss is computed as:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} \mathbf{1}_{[y=c]} \log \hat{p}_c. \tag{8}$$

We use the Matthews Correlation Coefficient (MCC) as the evaluation metric. MCC takes into account true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a measure of overall performance through

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{9}$$

We also use F1 score and Area Under the Receiver Operating Characteristic curve (AUROC) to evaluate model predictions (Appendix A.5).

3.3 Baselines

We compare eccDNAMamba with four representative genomic sequence models covering both attention-based and state-space paradigms: DNABERT2 [20], a Transformer pretrained on genomic k-mers; HyenaDNA [9], a long-range implicit-convolution model that scales efficiently; Caduceus [11], a bidirectional state-space model with base-level tokenization that captures symmetric context ; and DeepCircle [18], a CNN-based classifier specifically designed for the real-versus-pseudo eccDNA discrimination task, and therefore included as a baseline only for this specific task (Table. 4). In addition, we include eccDNAMamba–1M (w/o CA) as training eccDNAMamba without circular augmentation during both the pretraining and fine-tuning phases. This is an ablation model to verify the effect of circular augmentation.

Each model name is suffixed with its maximum supported input length, and any sequence exceeding the maximum length supported by a given baseline model is truncated during tokenization. All baselines are fine-tuned on the same eccDNA Multi-Task benchmark using their recommended learning rates, while eccDNAMamba and its ablation variant are fine-tuned with a learning rate of 3e-4. Each experiment is repeated with three random seeds, and we report the mean and standard deviation of all metrics (MCC, F1, and AUROC) for fair comparison.

4 Results

eccDNAMamba identifies cancer-associated eccDNAs across short and ultra-long length regimes. We first evaluate the ability of models to identify cancer-associated eccDNAs across short and ultra-long length regimes. As shown in Table 2, eccDNAMamba achieves the best overall performance across both regimes, reaching 59.0% MCC and 79.3% F1 on short sequences, and 57.9% MCC and 82.1% F1 on ultra-long ones. Its ablated variant without circular augmentation shows slight degradation, confirming the effectiveness and necessity of topology-aware modeling.

In contrast, DNABERT-2 and Caduceus only slightly outperform HyenaDNA on short sequences with MCC values of 57.7% and 57.8% compared to 54.0%, but their performance drops sharply on long sequences. HyenaDNA remains relatively robust with 55.5% MCC, whereas DNABERT-2 collapses to 10.9%. DNABERT-2 is inapplicable to sequences longer than 10k and requires truncation for ultra-long sequences. This highlights the importance of modeling long-range dependencies for maintaining the topological integrity of eccDNA.

Task	Model	MCC (%)	F1 (%)	AUROC (%)
Cancer vs. healthy eccDNA discrimination (short)	eccDNAMamba–1M	**59.0±0.6**	**79.3±0.2**	**86.7±0.3**
	eccDNAMamba–1M (w/o CA)	57.5±1.6	78.6±0.7	86.3±0.1
	Caduceus–131k	57.8±1.2	78.6±0.5	86.9±0.5
	DNABERT-2–10k	57.7±1.3	78.6±0.6	86.6±0.5
	HyenaDNA–450k	54.0±0.5	76.9±0.2	84.6±0.2
Cancer vs. healthy eccDNA discrimination (ultra-long)	eccDNAMamba–1M	**57.9±4.5**	**82.1±1.1**	**86.9±0.8**
	eccDNAMamba–1M (w/o CA)	56.8±4.3	81.6±1.0	86.3±1.2
	Caduceus–131k	48.4±3.5	73.1±2.6	82.2±3.3
	DNABERT-2–10k	10.9±3.8	49.3±6.7	56.5±2.0
	HyenaDNA–450k	55.5±2.1	80.8±2.0	85.8±1.7

Table 2: Performance of different models on **Cancer vs. healthy eccDNA prediction** across sequence length ranges. EccDNAMamba outperforms all other baseline models in EccDNA sequences with lengths below 10k and between 10k and 200k.

eccDNAMamba infers eccDNA copy-number levels. eccDNA is routinely replicated in cancer cells, selectively amplifying sequences that can exacerbate the cells' oncogenic potential [2]. We evaluate whether eccDNA copy-number levels can be inferred directly from sequences. As shown in Table 3, under a small-sample setting of only 760 training sequences (Table 1), Caduceus and DNABERT-2 show substantial degradation, with MCC values of 20.4% and 20.5% at the higher threshold and 2.5% and 20.3% at the lower threshold. HyenaDNA performs more stably with 29.4% MCC and 64.6% F1. In contrast, eccDNAMamba achieves the best and most stable results across both threshold settings, reaching 36.0% MCC and 67.5% F1 at the higher threshold and 28.7% MCC and 63.9% F1 at the lower threshold, effectively capturing copy-number–related signals despite limited data.

Task	Model	MCC (%)	F1 (%)	AUROC (%)
eccDNA copy-number level prediction (copy number > 10)	eccDNAMamba–1M	**36.0±10.6**	**67.5±5.5**	**70.0±4.6**
	eccDNAMamba–1M (w/o CA)	35.6±8.0	67.1±4.4	69.8±5.4
	Caduceus–131k	20.4±22.8	59.3±10.4	61.9±6.8
	DNABERT-2–10k	20.5±11.3	60.2±5.5	63.5±7.8
	HyenaDNA–450k	29.4±8.9	64.6±4.4	66.9±5.8
eccDNA copy-number level prediction (copy number > 4.5)	eccDNAMamba–1M	**28.7±4.5**	**63.9±1.9**	**67.2±2.3**
	eccDNAMamba–1M (w/o CA)	24.6±6.2	62.1±2.9	65.0±0.6
	Caduceus–131k	2.5±12.5	51.5±5.8	59.7±4.5
	DNABERT-2–10k	20.3±2.0	59.1±0.9	62.7±2.9
	HyenaDNA–450k	22.5±17.6	61.2±8.7	62.5±9.4

Table 3: Performance of different models on **eccDNA copy-number level prediction**. eccDNAMamba demonstrated consistently robust performance under both settings of the copy number classification task, while Caduceus likely collapses because its base pair-level tokenization produces extremely long inputs that overfit and destabilize under small data.

eccDNAMamba distinguishes real eccDNA sequences from pseudo-eccDNA sequences We finally validate whether models can distinguish true eccDNA sequences from pseudo-eccDNA sequences that are randomly extracted from linear genomic regions of *Homo sapiens*. The pseudo-eccDNA sequences are generated to match the length distribution of true eccDNAs, ensuring that models cannot rely on length cues for classification. This task evaluates how well the

model captures sequence features or compositions that distinguish true eccDNAs from random sequences.

As in Table 4, eccDNAMamba achieves the highest performance among all models with 42.7% MCC, 71.0% F1, and 75.8% AUROC, demonstrating superior ability to capture structural and contextual signals embedded in ultra-long eccDNAs. In contrast, DNABERT-2 and DeepCircle perform substantially worse with MCC values of 37.3% and 29.9%. Caduceus and HyenaDNA remain relatively robust yet still fall short of eccDNAMamba. Even though the mean MCC difference is 1.8 points (42.7 vs. 40.9), eccDNAMamba has tighter variance($\pm$ 1.8 vs. $\pm$ 3.9), indicating that its performance is more reliable, more reproducible, and less sensitive to random initialization. These results suggest that eccDNAMamba captures the fundamental biological features of real eccDNA, rather than relying on non-biological artifacts during classification. We also assess the generalization ability of our model by extending this task to *Gallus gallus* and *Arabidopsis thaliana* and observed similar trends, with results reported in Appendix A.6.

Task	**Model**	**MCC (%)**	**F1 (%)**	**AUROC (%)**
Real vs. pseudo-eccDNA discrimination (*Homo sapiens*)	eccDNAMamba–1M	**42.7±1.8**	**71.0±1.2**	**75.8±0.7**
	eccDNAMamba–1M (w/o CA)	41.5±0.3	70.5±0.2	75.3±1.9
	Caduceus–131k	40.9±3.9	70.4±1.9	75.4±0.8
	Deepcircle–0.5k	29.9±1.7	64.6±0.8	69.8±0.9
	DNABERT-2–10k	37.3±0.8	68.5±0.5	74.2±0.7
	HyenaDNA–450k	35.8±0.7	67.8±0.3	72.4±1.1

Table 4: Performance of different models on **Real vs. pseudo-eccDNA discrimination (*Homo sapiens*)**. eccDNAMamba achieves the best performance among all models, with a slight performance drop in its non-augmented version, yet still outperforms the baselines.

5 Biological Interpretations

To understand what drives eccDNAMamba predictions, we use Integrated Gradients (IG) for analysis, because it satisfies completeness and measures each token's true contribution by integrating gradients along a path, avoiding the noise of single-step saliency, while remaining naturally compatible with Mamba's attention-free, ultra-long-sequence architecture. IG assigns token-level attributions along a path from a pad baseline x_{pad}, which is fully filled with PAD tokens, to the input sequence x. For each token x_i, IG integrates the gradient of the model's output with respect to x_i along the straight-line path $x_\alpha = x_{pad} + \alpha(x - x_{pad}), \alpha \in [0, 1]$, using the logit difference between the target and other classes as the decision signal:

$$\mathrm{IG}_i(x) = (x_i - x_{pad}) \int_0^1 \frac{\partial\big(\mathrm{logit}_{\mathrm{target}}(x_\alpha) - \mathrm{logit}_{\mathrm{other}}(x_\alpha)\big)}{\partial x_i}\, d\alpha \tag{10}$$

Therefore, IG yields a principled, faithful importance score for each nucleotide, capturing sensitivity across the entire interpolation rather than a single gradient snapshot, and manages to reveal sequence features that drive eccDNAMamba's decisions.

IG reveals distinct landscapes across prediction groups. After stratifying sequences by prediction outcome, markedly different IG profiles emerged between eccDNA sequences predicted as cancerous and those predicted as healthy. In the ultra-long model (10-200k), cancerous sequences (true positives) exhibited consistent positive signal throughout the entire sequences, with slightly stronger values in the first half. In contrast, healthy sequences (true negatives) present a prominent bipartite IG profile with negative signals at the first half of the sequence and positive signals at the second half (Figure 2). These differences highlight the global differences our model is able to clearly distinguish between healthy and cancerous eccDNA sequences.

Regulatory element groups and transposable element families receive differential attribution. Using ENCODE annotations (ENCODE cCRE v4) [23], we analyzed the aggre-

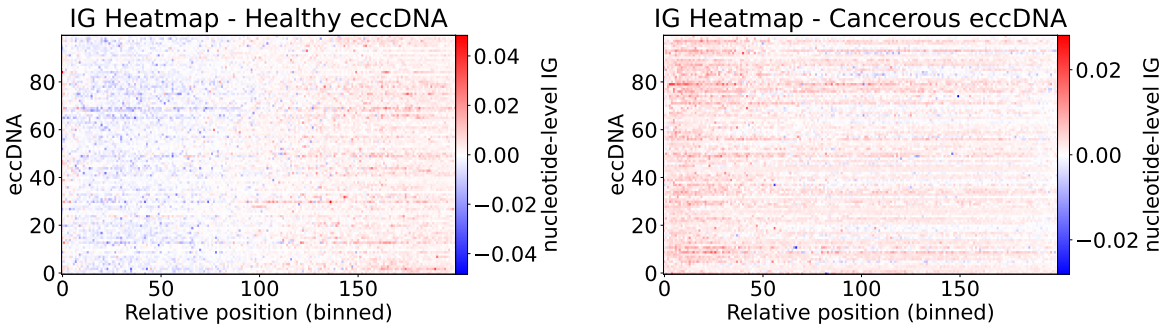


Figure 2: IG values across eccDNAs of correctly predicted healthy (left) and cancerous (right) sequences by the ultra-long model.

gated IG contributions of each class of regulatory elements. Across both short (<10 kbp) and ultra-long (10–200 kbp) settings, cCRE categories such as PLS (promoters), pELS (proximal enhancers), and dELS (distal enhancers) contribute significantly higher per-base |IG| values than length- and context-matched background sequences randomly selected from other regions of the same eccDNAs (Figure 3A,B). Based on RepeatMasker annotations [24], we also observed retrotransposon (RTE) families-specific |IG| patterns: Alu elements showed significantly lower, while LINE-1 and ERV families exhibited higher per-base |IG| than matched background (Figure 3C). These results indicate that eccDNAMamba systematically prioritizes certain regulatory DNA and selectively attends to specific TE classes, consistent with eccDNA's ability to harbor RTE and promoter/enhancer-like sequences that can amplify oncogenic programs. [1, 25]

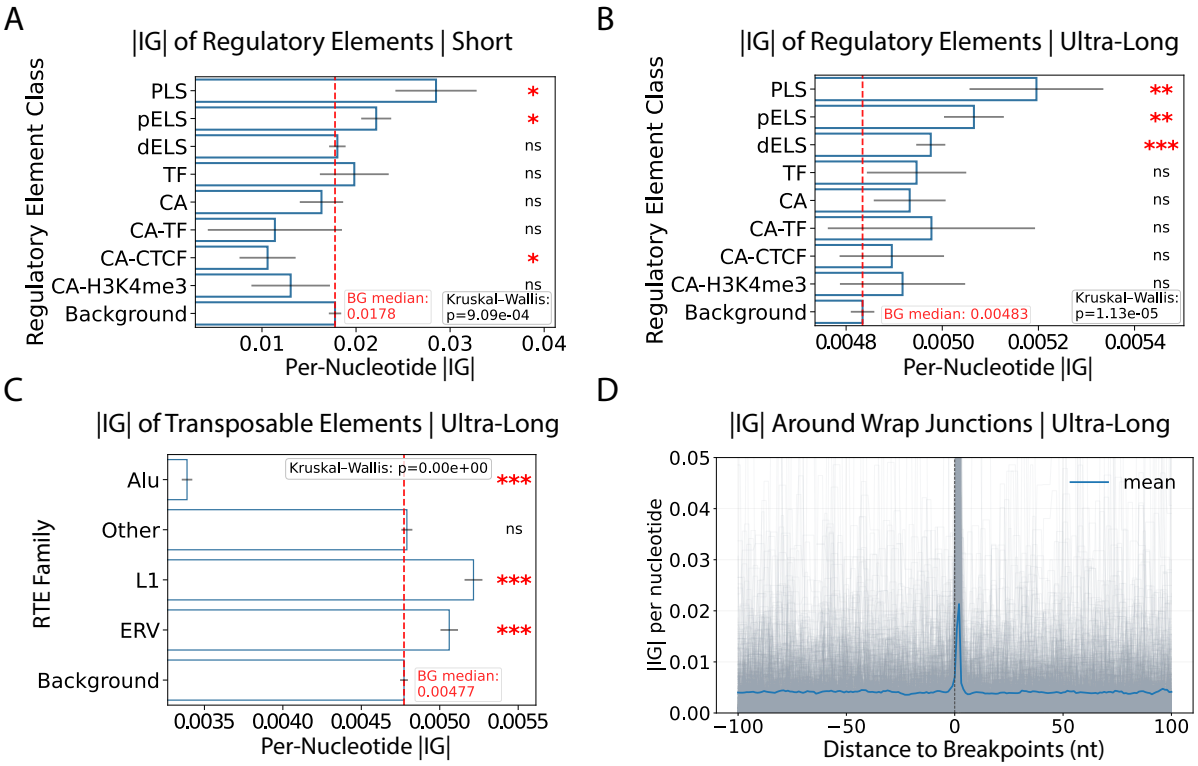


Figure 3: (A,B) Per-nucleotide |IG| values across gene regulatory elements from the short (below10k) model (A) and ultra-long (10-200k) model (B). (C) Per-nucleotide |IG| values across transposable elements from the ultra-long (10-200k) model. (D) |IG| peaks at breakpoints of eccDNA by the ultra-long model. Kruskal Wallis test determines significant differences between groups. Stars indicate significance from two-tailed Mann–Whitney U test, as compared to the "Background." ("*" = $p < 0.05$, "**" = $p < 0.01$, "***" = $p < 0.001$)

Breakpoint-centered attribution highlights circular topology. Aggregating IG scores around the sequence-end to appended-token breakpoints revealed a pronounced enrichment at the head–tail junction of sequences (Figure 3D). This enrichment validates that the circular augmentation strategy is recognized and important to our model, capturing head–tail dependencies. It provides direct interpretability evidence on the role of augmentation in the model's fidelity to eccDNA topology.

eccDNAMamba discovers known and novel motifs relevant for cancer eccDNA. Interpreting high–|IG| subsequences from eccDNAMamba using XSTREME [26] uncovered 69 motifs, 40 of which (58%) matched known matrices by TomTom(CIS-BP_2.00 | *Homo sapiens*) [27, 28]. Twenty-three motifs were significantly enriched in cancer eccDNA ($\log_2 \mathrm{OR} > 0$; $q < 0.05$), led by A/T-rich seeds such as **TAATTAATAT** ($\log_2 \mathrm{OR} = 15.65$; $q = 9.0 \times 10^{-12}$). Among mapped examples, **TTTCCTGAGAAAT** aligns to a STAT-like matrix (M08173_2.00), **GTAAACAATAAACA** aligns to a FOX-like matrix (M03016_2.00), and **TAATTAATAT** aligns to an ARID-family matrix (M01659_2.00), collectively pointing to transcription-factor programs frequently implicated in oncogenesis (cytokine/STAT signaling, FOX pioneer-factor networks, and A/T-binding ARID factors) [29, 30, 31]. Strikingly, only 8/23 motifs aligned to known motifs, all of which match transcription factor families relevant to cancer biology. This indicates that the other 15/23 cancer-enriched motifs, which lack an unambiguous database match, could serve as a resource to reveal previously unannotated sequence patterns associated with cancer eccDNA. These results could provide mechanistic interpretability and testable hypotheses about regulatory grammar on cancer-derived eccDNA.(Figure 4 and Appendix A.9)

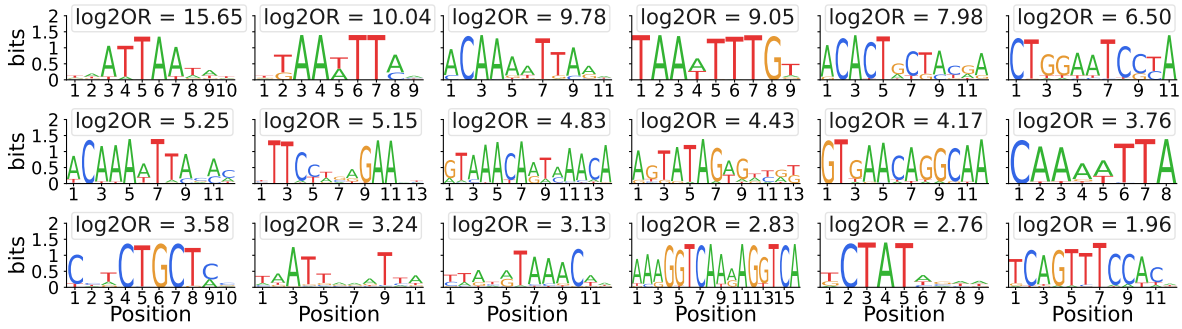


Figure 4: Representative cancer-overrepresented motifs discovered by IG analysis of eccDNAMamba.

Taken together, these analyses provide three convergent insights. eccDNAMamba learns to prioritize biologically important elements and motifs across length regimes. The model distinguishes correct from incorrect predictions through clear differences in IG landscapes. Our circular augmentation is not only beneficial for state-of-the-art performance but also interpretable: it induces consistent, localized attribution signals at the head–tail junction, directly reflecting the circular topology of eccDNA.

6 Discussion

eccDNAMamba presents a topology-aware, bidirectional state-space framework that efficiently models the circular and ultra-long characteristics of eccDNA. By integrating bidirectional Mamba-2 encoding, motif-level BPE tokenization, and lightweight head–tail augmentation, the model captures wrap-around dependencies and long-range interactions in linear time. Integrated Gradients analyses further confirm the biological interpretability of the model.

Despite its strong performance, current training data are primarily human-centric, limiting cross-species generalization. Future work will expand pretraining to broader multi-species eccDNA corpora and apply the framework to more diverse downstream tasks.

References

[1] Dong, Y. *et al.* Extrachromosomal dna (ecdna) in cancer: mechanisms, functions, and clinical implications. *Frontiers in oncology* **13**, 1194405 (2023).

[2] Bailey, C., Pich, O., Thol, K. *et al.* Origins and impact of extrachromosomal dna. *Nature* **635**, 193–200 (2024).

[3] Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).

[4] Koche, R. P. *et al.* Extrachromosomal circular dna drives oncogenic genome remodeling in neuroblastoma. *Nature genetics* **52**, 29–34 (2020).

[5] Zhao, X. *et al.* Circlebase: an integrated resource and analysis platform for human eccdnas. *Nucleic acids research* **50**, D72–D82 (2022).

[6] Peng, L., Zhou, N., Zhang, C.-Y., Li, G.-C. & Yuan, X.-Q. eccdnadb: a database of extrachromosomal circular dna profiles in human cancers. *Oncogene* **41**, 2696–2705 (2022).

[7] Zhou, Z. *et al.* Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006* (2023).

[8] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025). URL `https://doi.org/10.1038/s41592-024-02523-z`.

[9] Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution (2023). URL `https://arxiv.org/abs/2306.15794. 2306.15794`.

[10] Li, F., Lu, W. & Bai, Y. Hyenacircle: a hyenadna-based pretrained large language model for long eccdna prediction. *Frontiers in Genetics* **16**, 1641162 (2025).

[11] Schiff, Y. *et al.* Caduceus: Bi-directional equivariant long-range dna sequence modeling. In *International Conference on Machine Learning*, 43632–43648 (PMLR, 2024).

[12] Liu, Z., Li, J. & Zhang, Z. eccDNAMamba: A pre-trained model for ultra-long eccDNA sequence analysis. In *ICML 2025 Generative AI and Biology (GenBio) Workshop* (2025). URL `https://openreview.net/forum?id=56xKN7KJjy`.

[13] Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[14] Gage, P. A new algorithm for data compression. *The C Users Journal* **12**, 23–38 (1994).

[15] Dao, T. & Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060* (2024).

[16] Zhong, T. *et al.* eccdna atlas: a comprehensive resource of eccdna catalog. *Briefings in Bioinformatics* **24**, bbad037 (2023).

[17] Osama, M., Ali, S., Binte Nawaz, A., Ullah, U. *et al.* "extrachromosomal circular dna (eccdna): A key driver of tumorigenesis, drug resistance, and prognosis across gastrointestinal malignancies". *Health Sciences Review* **15**, 100228 (2025). URL `https://www.sciencedirect.com/science/article/pii/S2772632025000200`.

[18] Chang, K.-L. *et al.* Short human eccdnas are predictable from sequences. *Briefings in Bioinformatics* **24**, bbad147 (2023). URL https://doi.org/10.1093/bib/bbad147. https://academic.oup.com/bib/article-pdf/24/3/bbad147/50410894/bbad147.pdf.

[19] Fang, M. *et al.* eccdna-pipe: an integrated pipeline for identification, analysis and visualization of extrachromosomal circular dna from high-throughput sequencing data. *Briefings in Bioinformatics* **25**, bbae034 (2024). URL https://doi.org/10.1093/bib/bbae034. https://academic.oup.com/bib/article-pdf/25/2/bbae034/56663657/bbae034.pdf.

[20] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[21] Zhao, Y., Yu, L., Zhang, S., Su, X. & Zhou, X. Extrachromosomal circular dna: Current status and future prospects. *Elife* **11**, e81412 (2022).

[22] Joshi, M. *et al.* Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics* **8**, 64–77 (2020).

[23] The ENCODE Project Consortium, Moore, J. E., Purcaro, M. J. *et al.* Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

[24] Smit, A. F. A., Hubley, R. & Green, P. Repeatmasker open-4.0. http://www.repeatmasker.org (2013–2015).

[25] Kraft, K. *et al.* Enhancer activation from transposable elements in extrachromosomal dna. *Nature Cell Biology* (2025).

[26] Grant, C. E. & Bailey, T. L. Xstreme: Comprehensive motif analysis of biological sequence datasets. *bioRxiv* (2021).

[27] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biology* **8**, R24 (2007).

[28] Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

[29] Hu, X., Li, J., Fu, M., Zhao, X. & Wang, W. The jak/stat signaling pathway: from bench to clinic. *Signal Transduction and Targeted Therapy* **6**, 402 (2021).

[30] Seachrist, D. D., Anstine, L. J. & Keri, R. A. Foxa1: A pioneer of nuclear receptor action in breast cancer. *Cancers* **13**, 5205 (2021).

[31] Mullen, J., Kato, S., Sicklick, J. K. & Kurzrock, R. Targeting arid1a mutations in cancer. *Cancer Treatment Reviews* **100**, 102287 (2021).

A Supplementary Information

A.1 Dataset

As summarized in Table 5, we assemble a unified eccDNA corpus by standardizing several public resources: CircleBase [5], which provides 601,036 human eccDNAs including both healthy and cancer-associated sequences; eccDNAdb [6], which contributes 1,270 eccDNAs identified from 480 samples spanning 42 cancer types, derived from a total of 3,395 tumor samples covering 57 cancers across patient tissues and cancer cell lines; and eccDNA Atlas [16], which aggregates 629,987 eccDNAs across multiple species, expanding beyond human to include *Mus musculus*, *Saccharomyces cerevisiae*, *Gallus gallus*, and *Arabidopsis thaliana*.

Dataset	Species	Number of eccDNAs	Additional Information
CircleBase [5]	*Homo sapiens*	601,036	Healthy and cancer-associated eccDNAs
eccDNA Atlas [16]	Multi-species	629,987	Multi-species eccDNAs
eccDNAdb [6]	*Homo sapiens*	1,270	eccDNAs from Cancer samples and cell lines

Table 5: List of eccDNA datasets used in this study.

A.2 Mamba-2 Architecture

The Mamba-2 [15] is a linear state-space model (SSM) that parameterizes sequence dynamics as a continuous-time linear system:

$$x_{t+1} = A_t x_t + B_t u_t, \quad y_t = C_t x_t + D_t u_t, \tag{11}$$

where $u_t \in \mathbb{R}^d$ is the input token embedding, x_t is the latent state, and y_t is the output at time t. This formulation enables constant-time updates per token and thus overall $\mathcal{O}(L)$ complexity for a sequence of length L, in contrast to $\mathcal{O}(L^2)$ attention. The parameters (A_t, B_t, C_t, D_t) are dynamically generated via depthwise convolutions and gated activations, allowing nonlinear feature mixing while maintaining linear-time recurrence.

A.3 Pretraining

We adopt a hybrid method combining contextual embedding-level control and masking as reported in the previous section during pretraining.

We pre-train the model using a span-masked language modeling (SpanMLM) objective adapted for eccDNA modeling. Specifically, we use a bidirectional variant of the Mamba architecture, denoted as **BiMambaForMaskedLM**, initialized from the `state-spaces/mamba2-130m` configuration. The model is trained from scratch on a custom eccDNA corpus, tokenized using a byte-pair encoding vocabulary. The tokenizer includes the special tokens `[PAD]`, `[UNK]`, `[MASK]`, and `[CLS]`.

Let the input sequence be denoted as

$$x = (x_1, x_2, \ldots, x_L)$$

and let $\mathcal{M} \subset \{1, \ldots, L\}$ be a set of span-masked positions. The model is trained to reconstruct the original tokens at these positions by minimizing the span-masked cross-entropy loss:

$$\mathcal{L}_{\text{SpanMLM}} = -\sum_{i \in \mathcal{M}} \log p_\theta(x_i \mid x_{\setminus \mathcal{M}})$$

where p_θ denotes the model's conditional output distribution and $x_{\setminus \mathcal{M}}$ denotes the input with masked spans.

Training is conducted using the HuggingFace `Trainer` API. We set the effective batch size to 144 by using 6 samples per GPU and accumulating gradients over 8 steps. The model is trained for 3 epochs using a learning rate of 5×10^{-4}, with a linear warmup over 6% of total steps. We use the AdamW optimizer with a weight decay of 0.01. All training runs use BF16 mixed precision and are executed on a single node with 3 NVIDIA L40S GPUs.

As shown in Table 6,we filter out sequences longer than 10kbp, resulting in a final corpus of $1,087,886$ eccDNA sequences totaling 524 million base pairs on the dataset we introduced in Appendix A.1. After applying BPE tokenization, this corpus comprises approximately 101.5 million tokens, yielding an average token length of 5.16 base pairs per token after BPE tokenization.

Number of Sequences	**Total Base Pairs**	**Tokens (BPE)**	**Average Token Length (bp)**
1,087,886	524M	101.5M	5.16

Table 6: Statistics of the eccDNA corpus after filtering (sequences longer than 10kbp removed) and BPE tokenization.

A.4 Efficiency

Let $s_{1:L} \in \mathbb{R}^d$ be a single eccDNA sequence tokenized with byte-pair encoding (BPE), where L is the number of *tokens* and d the channel dimension. An eccDNAMamba block processes $u_{1:L}$ as follows: (1) append a fixed k-token prefix to expose head–tail context, yielding length $L' = L + k$; (2) run one forward scan and one reverse scan over the augmented sequence; (3) fuse the two output streams with a constant-cost operation (e.g., sum or concat+linear). We treat k as a small constant (default k=64), so $L' = \Theta(L)$.

Each scan applies the per-step update

$$x_{t+1} = A_t u_t + B_t s_t, \qquad y_t = C_t u_t + D_t s_t,$$

where computing (A_t, B_t, C_t, D_t) and performing the update at each position both cost $O(1)$ work independent of L. Thus, a left-to-right pass costs $O(L')$, the reverse pass costs $O(L')$, and fusion is $O(L')$ in total.

Overall, the block runs in $O(L') + O(L') + O(L') = O(L)$ time; the bidirectional constant factor does not affect the asymptotics. Activation storage is $O(L)$, and with standard chunking or rematerialization, the peak memory can be held to $O(B)$ for a chosen block size B while preserving $O(L)$ total work. Compared to self-attention's $O(L^2 d)$ cost, eccDNAMamba scales linearly in L and accommodates ultra-long eccDNA.

This is supported by the memory profile in Fig. 8: eccDNAMamba reaches a flat plateau at 46% and stays constant over time, whereas HyenaDNA plateaus around 58%, Caduceus continues climbing to 82%, and DNABERT-2 peaks above 90%. The near-constant footprint for eccDNAMamba aligns with our analysis, in contrast to attention-based baselines whose usage grows with sequence length and runtime.

A.5 Evaluation Metrics

We further report the F1 score, which balances precision and recall by taking their harmonic mean:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

where precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

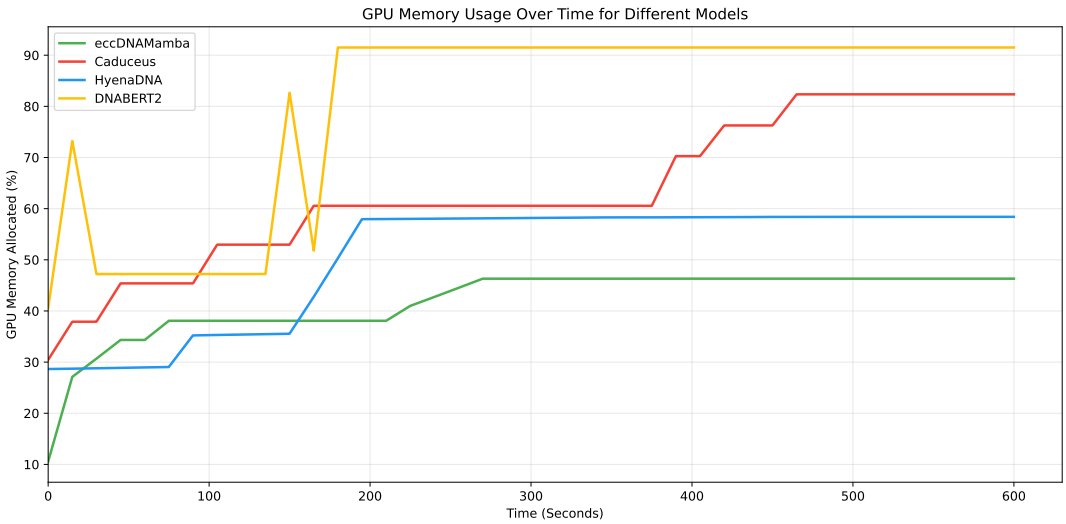


Figure 5: GPU memory usage over time for eccDNAMamba, Caduceus, HyenaDNA, and DNABERT-2, showing eccDNAMamba's lower and stable memory footprint

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

We also report the Receiver Operating Characteristic Area Under the Curve (AUROC), which measures the model's ability to distinguish between positive and negative classes. It is computed by integrating the true positive rate (TPR) over the false positive rate (FPR) across all classification thresholds:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{16}$$

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{17}$$

A higher AUROC indicates better discrimination ability of the model across varying thresholds.

A.6 Generalization on different species

To evaluate the cross-species generalization ability of different genomic models, we further extended the real vs. pseudo-eccDNA discrimination task to two non-human species, *Arabidopsis thaliana* and *Gallus gallus*. This task examines whether a model trained primarily on human eccDNA can reliably identify genuine eccDNA structures in evolutionarily distant genomes. High performance on these datasets indicates robust representation learning and strong transferability beyond the training species.

As shown in Table 7, the performance of various genomic models on the *Arabidopsis thaliana* dataset. Across models, eccDNAMamba shows comparable performance with DNABERT-2 and Caduceus. In contrast, Deepcircle and HyenaDNA exhibit substantial drops in MCC, suggesting that models designed for linear or short DNA contexts struggle to generalize to the *Arabidopsis thaliana* dataset.

Task	Model	MCC (%)	F1 (%)	AUROC (%)
Real vs. pseudo-eccDNA discrimination (*Arabidopsis Thaliana*)	eccDNAMamba–1M	46.8±7.0	73.1±3.2	80.3±4.1
	eccDNAMamba–1M (w/o CA)	45.9±6.4	72.9±3.2	79.1±2.9
	Caduceus–131k	47.9±2.8	73.4±1.1	80.3±1.8
	Deepcircle–0.5k	34.9±6.4	67.1±2.9	71.9±2.6
	DNABERT-2–10k	49.6±4.7	74.4±2.1	79.3±0.5
	HyenaDNA–450k	31.0±7.9	65.4±4.0	72.0±2.7

Table 7: Results on *Arabidopsis thaliana*

As shown in Table 8, eccDNAMamba achieves competitive performance with an MCC of 51.6%, closely matching Caduceus and its own ablated variant. This consistency highlights the robustness of eccDNAMamba in the *Gallus gallus* dataset. In contrast, transformer-based models such as DNABERT-2 and convolutional models like Deepcircle perform notably worse, reflecting limitations in generalizing to the *Gallus gallus* dataset.

Task	Model	MCC (%)	F1 (%)	AUROC (%)
Real vs. pseudo-eccDNA discrimination (*Gallus Gallus*)	eccDNAMamba–1M	51.6±4.2	75.1±2.5	81.6±2.4
	eccDNAMamba–1M (w/o CA)	50.4±2.1	75.0±1.2	81.5±1.6
	Caduceus–131k	52.1±0.8	76.0±0.4	82.8±0.1
	Deepcircle–0.5k	29.9±1.7	64.6±0.8	69.8±0.9
	DNABERT-2–10k	46.1±2.8	72.8±1.4	79.1±0.9
	HyenaDNA–450k	35.4±2.6	67.5±1.3	73.0±2.0

Table 8: Results on *Gallus gallus*

A.7 Performance of Different Models on Cancer vs. Healthy eccDNA Discrimination Task (200k–1M)

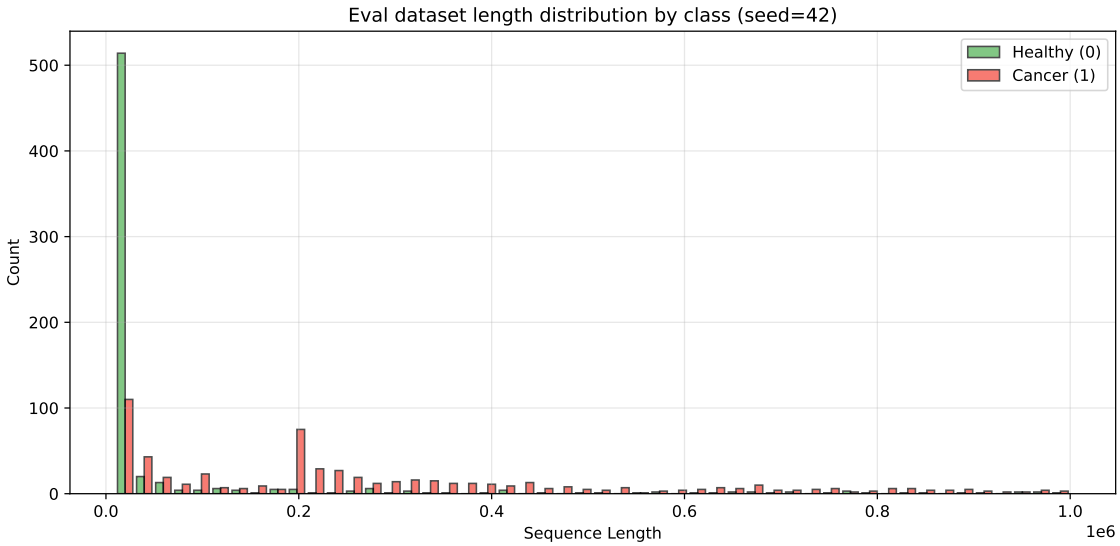


Figure 6: Length distribution of the evaluation dataset by class on **cancer vs. healthy eccDNA discrimination (200k–1M)**.

In the ultra-long eccDNA regime (>200 kb), we observed a pronounced class imbalance where cancer-associated sequences overwhelmingly dominate, while healthy counterparts are sparsely represented (Figure 6). This skewed distribution creates a trade-off for the model: on one hand, sequence length itself becomes a strong implicit signal, encouraging the model to associate longer sequences with cancer; on the other hand, mitigating this bias risks diminishing sensitivity to

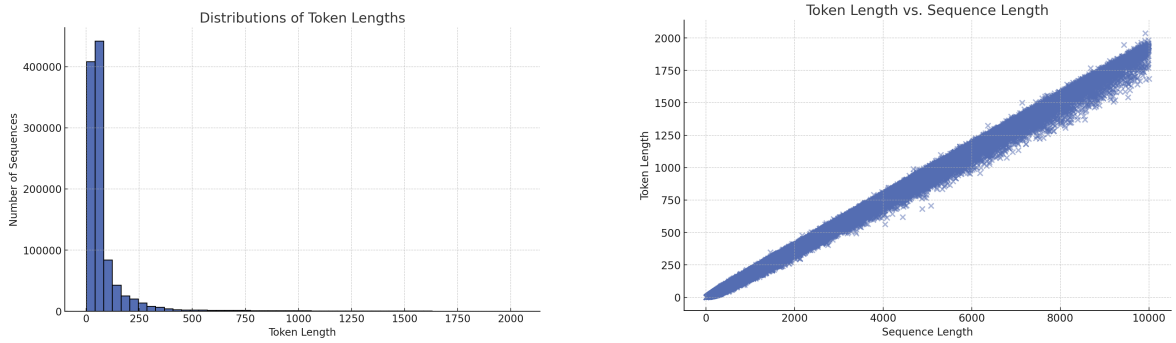
genuine biological correlations between length and oncogenic features. Consequently, as shown in Table 9 eccDNAMamba maintains a high F1 score but exhibits a near-zero MCC, indicating a tendency to classify most ultra-long sequences as cancer-associated. This outcome highlights the need for future work to disentangle biological length effects from dataset-induced bias.

Task	**Model**	**MCC (%)**	**F1 (%)**	**AUROC (%)**
Cancer vs. healthy eccDNA discrimination (200k - 1M)	eccDNAMamba–1M	0.0±0.0	88.1±0.7	55.8±0.5
	eccDNAMamba–1M (w/o CA)	0.0±0.0	88.1±0.7	53.3±1.8
	Caduceus–131k	0.0±0.0	88.1±0.7	61.6±2.0
	DNABERT-2–10k	10.7±1.8	65.6±9.5	61.2±1.5
	HyenaDNA–450k	5.1±4.5	86.4±2.9	68.8±7.7

Table 9: Performance of different models on **Cancer vs. healthy eccDNA discrimination (200k–1M)**.

A.8 Tokenization Statistics and Motivation for Circular Augmentation

We empirically chose to append approximately 25% of the sequence to its end to ensure successful pretraining. As shown in Figure 7a, most eccDNA sequences used for pretraining contain fewer than 256 tokens, so appending 25% corresponds to 64 tokens.



(a) Token length distribution after BPE tokenization.

(b) Nucleotide vs. tokenized length.

Figure 7: (a) Distribution of token lengths after BPE tokenization. Most sequences are tokenized into fewer than 256 tokens, with a long-tailed distribution extending to over 1000 tokens. (b) Relationship between nucleotide sequence length (in base pairs) and tokenized sequence length.

A.9 Motif Identification

A full catalog of motifs enriched in cancer-derived or healthy eccDNA sequences, identified through our IG-based motif discovery pipeline, is shown in Figure 8. Of these, 8 of 23 cancer-enriched motifs and 1 of 6 healthy-enriched motifs matched known transcription factor motifs in CIS-BP v2. Notably, healthy-enriched motifs exhibit higher CG content, even though the underlying eccDNA sequences from cancer and healthy samples display similar overall CG composition (Table 10).

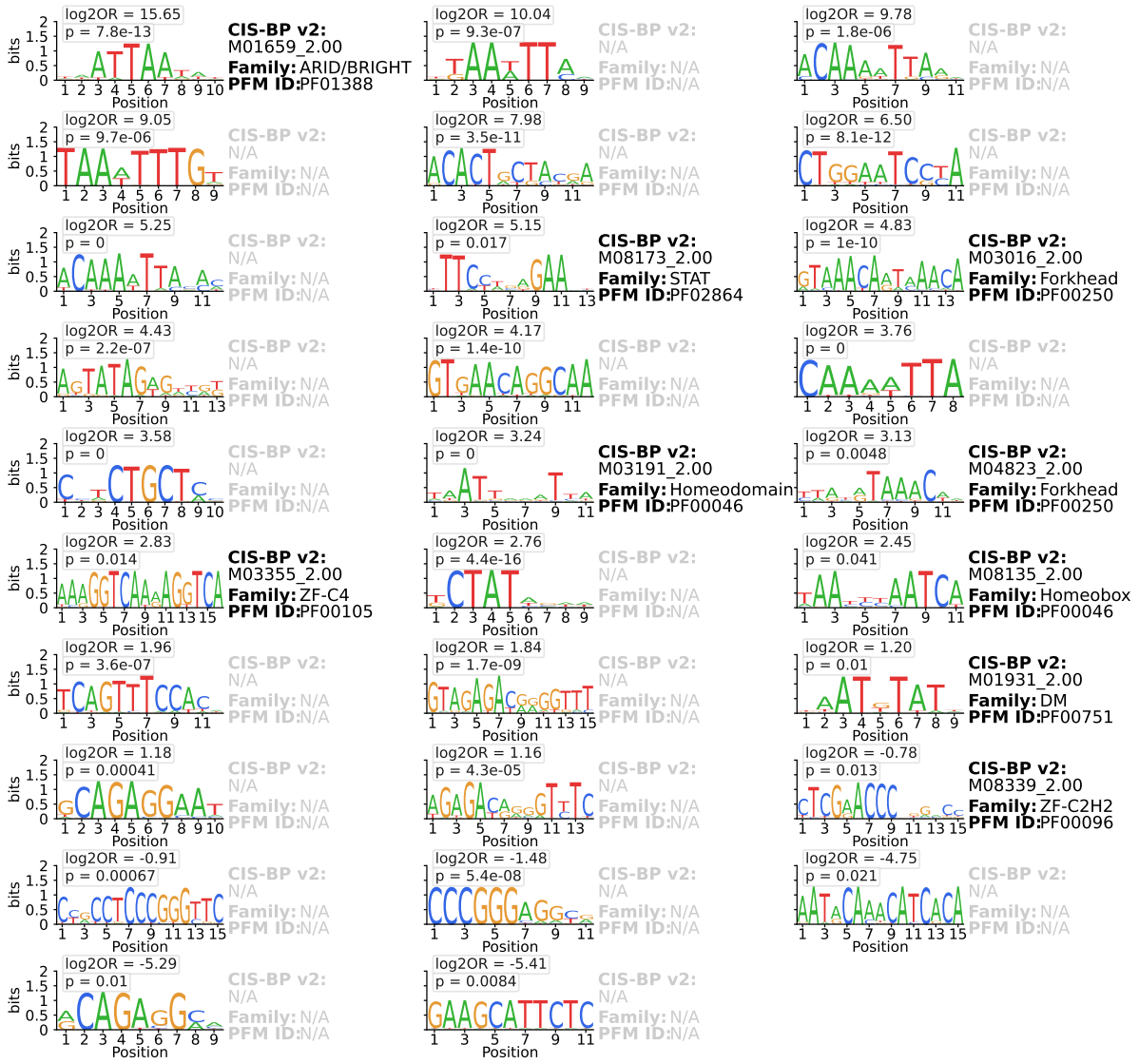


Figure 8: Motifs that are significantly over-represented in either cancer or healthy eccDNA sequences. Motifs with known matching in CIS-BP v2 are labeled with their CIS-BP motif ID, protein family, and Pfam ID.

Group	%A	%C	%G	%T	%GC
Cancer	29.004	21.061	21.051	28.884	42.112
Healthy	28.137	21.848	21.775	28.239	43.623

Table 10: Nucleotide composition across all sequences used for IG analysis and motif discovery. (n=726)