eccDNAMamba: A Pre-Trained Model for Ultra-Long eccDNA Sequence Analysis

Zhenke Liu[1], Ziqi Zhang[1], Jien Li[2]

[1] Department of Computer Science, Brown University, Providence, RI, USA [2] Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, USA;

Background

No existing models support full-length circular eccDNA due to sequence truncation and Transformer inefficiencies.

Introduce eccDNAMamba, the first bidirectional state-space model for circular DNA, enabling full-context modeling.

Combines circular augmentation, span masking, and BPE to achieve strong and robust performance to sequences up to 200 kbp—providing a scalable foundation for eccDNA analysis.

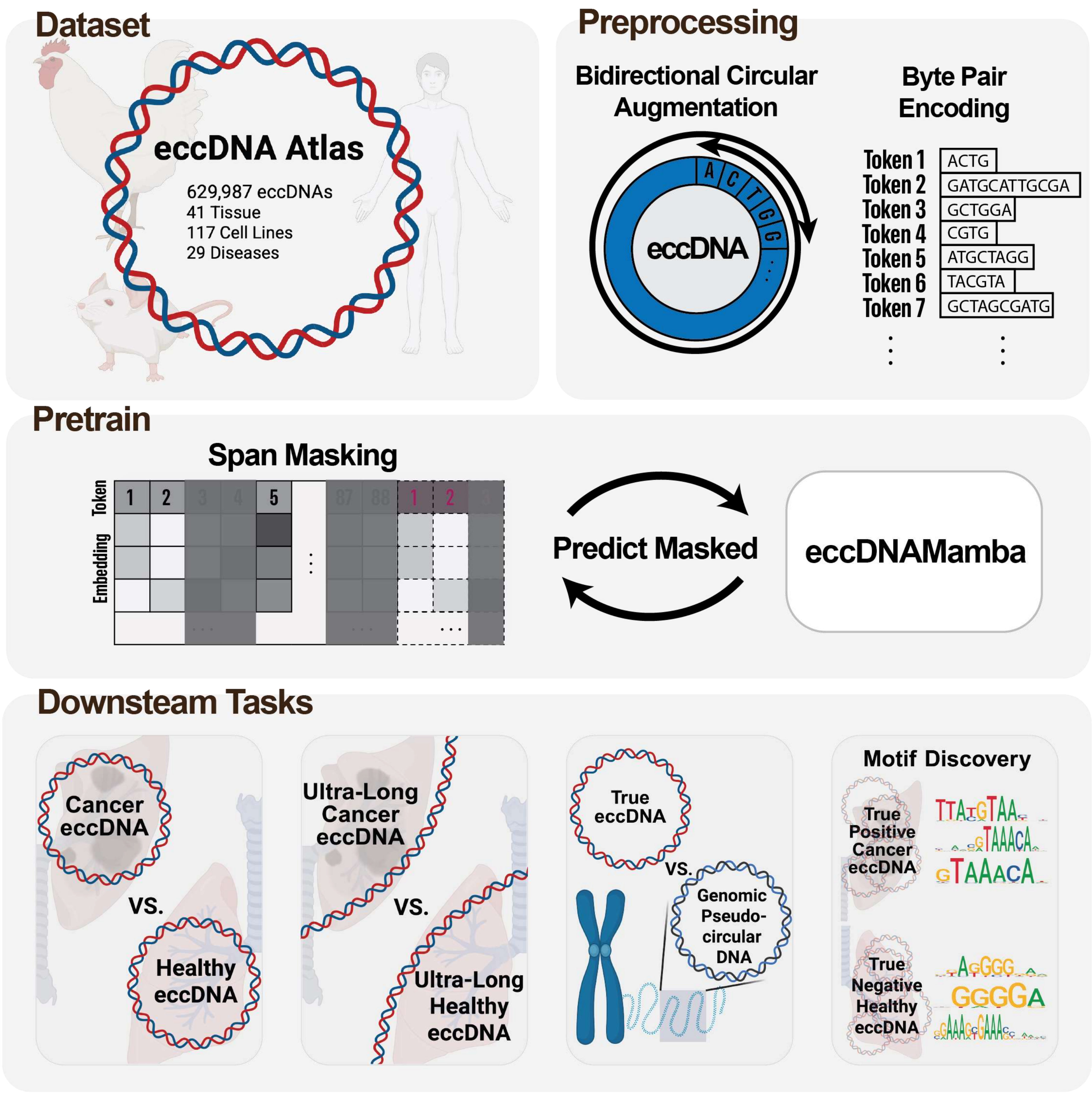
Overview

First bidirectional state-space encoder for circular DNA

Circular-aware input – appends the first 64 tokens to the tail so head-to-tail dependencies are preserved during training and inference

Pre-trained on eccDNAs from diverse species, cell lines, and disease states (≈ 101 M BPE tokens) and fine-tuned for various downstream tasks, positioning it as a versatile foundation for circular-genome analytics

Scales to full-length eccDNAs (200 kbp), enabling ultra-long-range reasoning that classic Transformers cannot handle



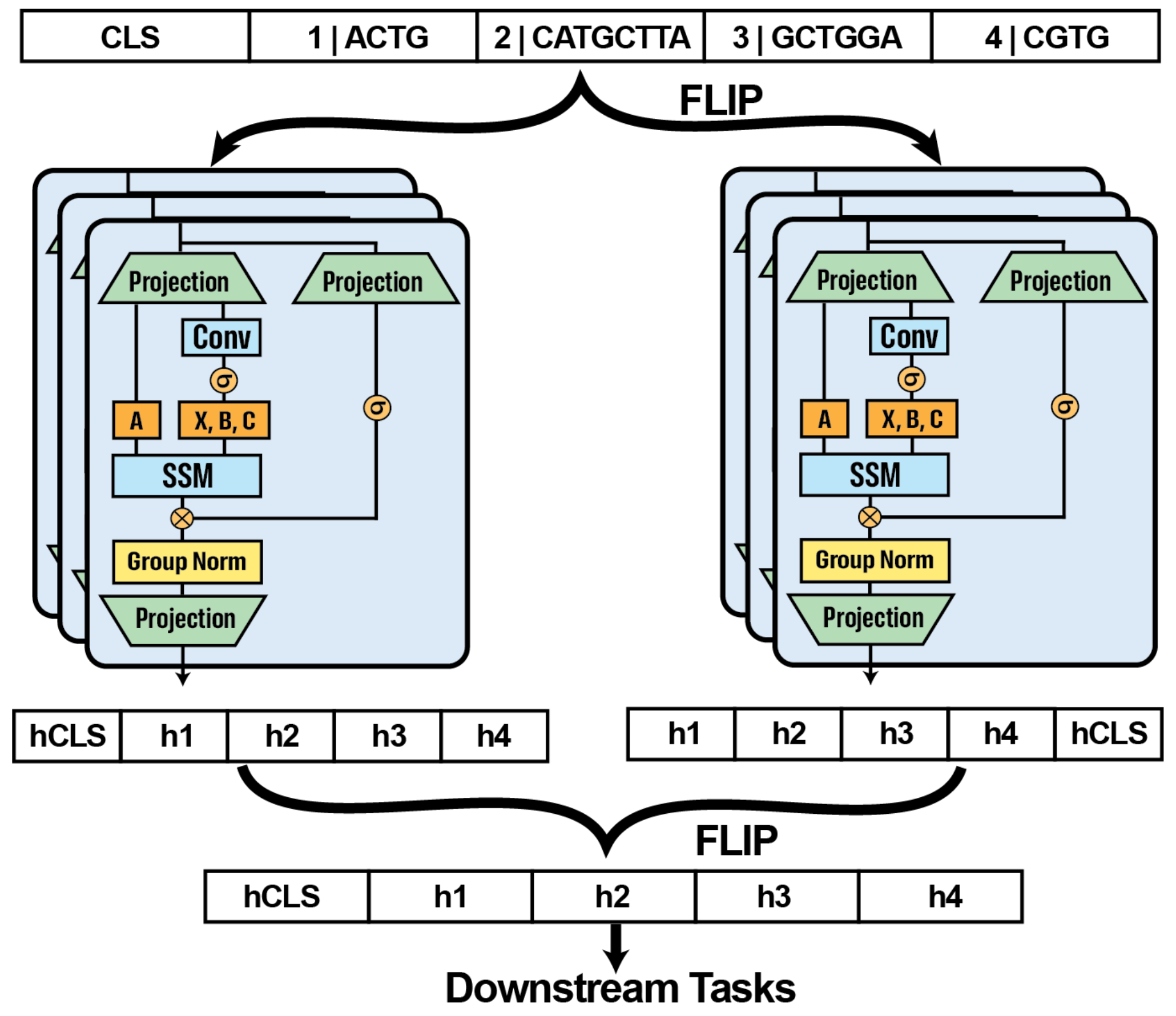
Model

Byte-Pair Encoding (BPE) compresses sequences to a motif-level alphabet (~5 bp per token)

Dual Mamba-2 encoders (forward + reverse) process the sequence bidirectionally; their hidden states are aligned and fused for global context integration

State-space kernels enable linear scaling of compute & memory to sequence length, unlocking routine training/inference on ultra-long sequences

Efficiently provides robust and globally aware embeddings for downstream tasks.



Results & Analysis

Task	Model	Training set (seq)	Test set (seq)	F1	accuracy	precision	recall
Cancer vs. Healthy(<10 kb)	**eccDNAMamba**	20,000 (10,000 cancer + 10,000 healthy)	4,000	**0.8242**	**0.8242**	0.8242	**0.8242**
	DNABERT-2	20,000 (10,000 cancer + 10,000 healthy)	4,000	0.8187	0.8187	0.8187	0.8187
	HyenaDNA	20,000 (10,000 cancer + 10,000 healthy)	4,000	0.8105	0.8104	0.8105	0.8105
	Caduceus	20,000 (10,000 cancer + 10,000 healthy)	4,000	0.8216	0.822	**0.8248**	0.822
Cancer vs. Healthy (10–200 kb)	**eccDNAMamba**	2,000 (1,000 cancer + 1,000 healthy)	400	**0.8147**	**0.8175**	**0.8377**	**0.8174**
	DNABERT-2	2,000 (1,000 cancer + 1,000 healthy)	400	0.5702	0.5725	0.574	0.5725
	HyenaDNA	2,000 (1,000 cancer + 1,000 healthy)	400	0.7261	0.735	0.7699	0.735
	Caduceus	2,000 (1,000 cancer + 1,000 healthy)	400	0.7102	0.7125	0.7192	0.7125
Authentic vs. pseudo	**eccDNAMamba**	20,000 (10,000 authentic + 10,000 pseudo)	4,000	**0.7401**	**0.7407**	**0.7428**	**0.7407**
	DeepCircle (zero-shot)	20,000 (10,000 authentic + 10,000 pseudo)	4,000	0.6363	0.6532	0.6883	0.6532
	DeepCircle (fine-tuned)	20,000 (10,000 authentic + 10,000 pseudo)	4,000	0.6712	0.6742	0.6808	0.6742

Performance

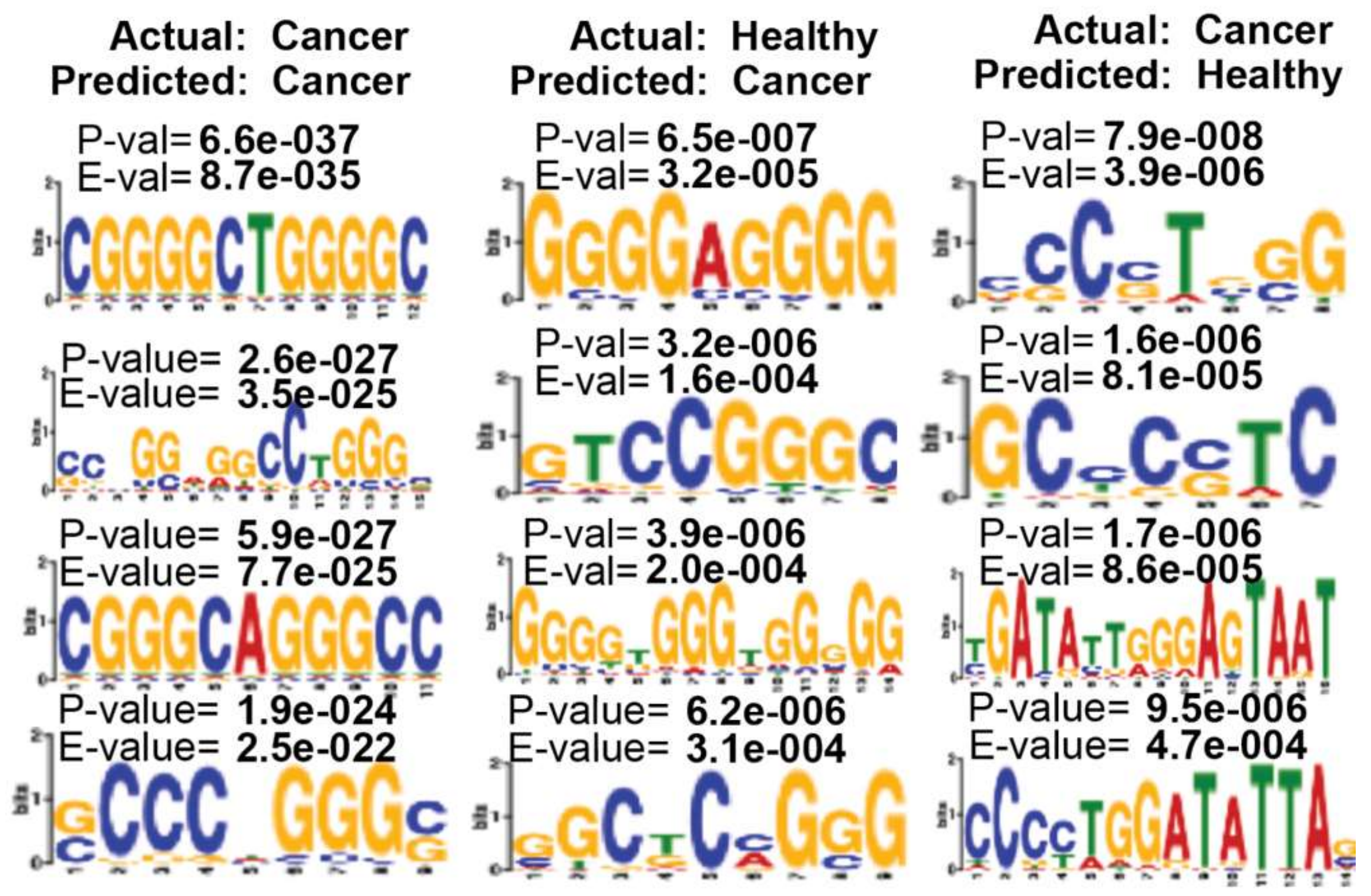
eccDNAMamba achieves improved performance in classifying cancer vs healthy and Authentic vs Pseudo eccDNA compared to other state-of-the-art models, and maintains performance in ultra-long sequences.

Motif Analysis

eccDNAMamba uses motifs with CG-centric cores typical of C2H2 zinc-finger binding sites (ZFs) to classify eccDNA of cancer origin.

ZNF24 & ZNF263 head a list of 218 ZF proteins matching the discovered motifs, linking model predictions to oncogenic regulators.

False-negative cancer eccDNAs are dominated by AT-rich motifs.



Future Work

Future work includes extending interpretability and cross-species experiments
Train model on CG/motif balanced datasets to extract other biologically significant sequence features