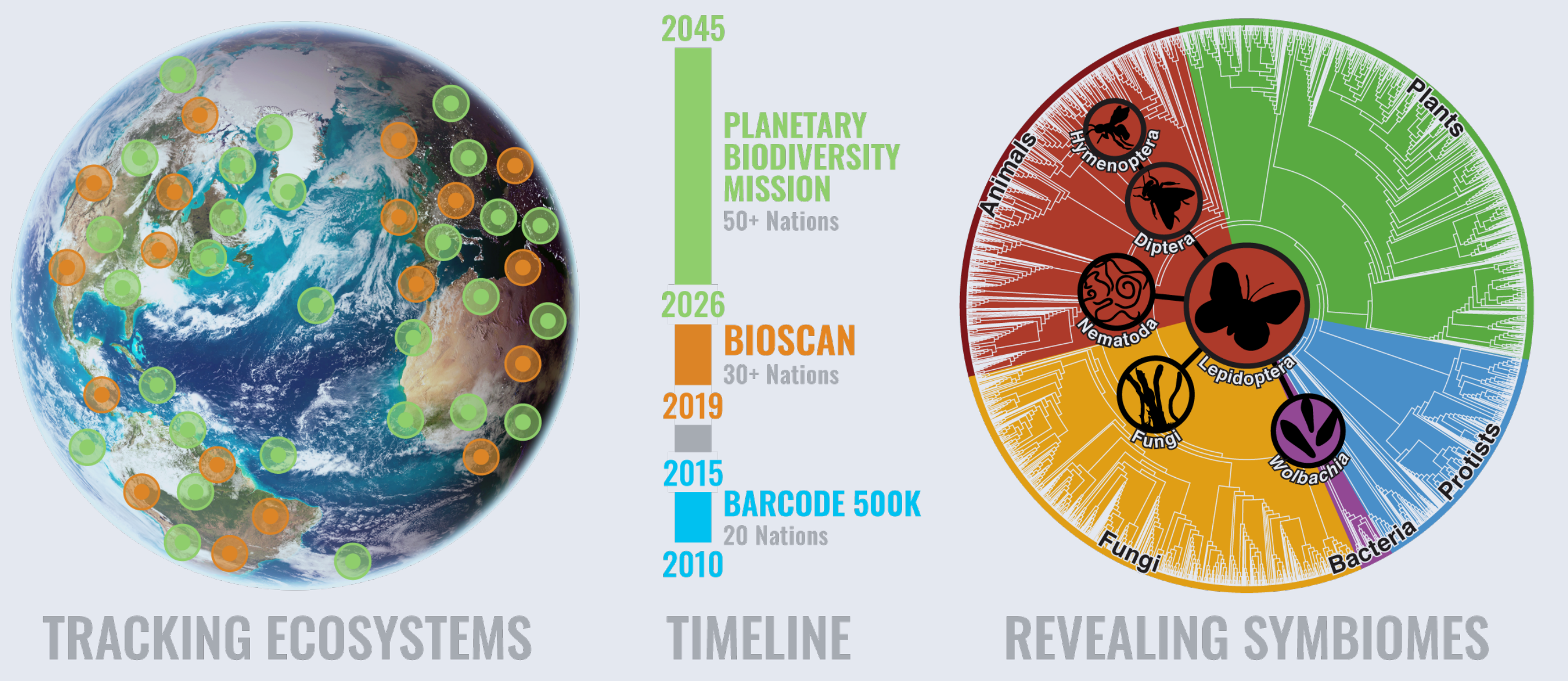
BarcodeMamba: State Space Models for Biodiversity Analysis

Tiancheng Gao[1,2], Graham W. Taylor[1,2*]
[1]University of Guelph, [2]Vector Institute for AI
*Correspondence: Graham W. Taylor (gwtaylor@uoguelph.ca)

1. Introduction



BarcodeMamba was developed to support BIOSCAN's mission of establishing a global biodiversity observation system through DNA barcoding.

DNA barcodes are genetic markers that enable efficient species identification by analyzing short, standardized sections of DNA rather than entire genomes. While transformers and state space models (SSMs) have advanced human genome analysis, **identifying invertebrate species from DNA barcodes** remains challenging due to complex taxonomic relationships and unknown species. BarcodeBERT, a transformer-based foundation model, has been the state-of-the-art solution for these challenges. However, its attention mechanism has quadratic complexity, resulting in substantial computational costs at scale. Building on the success of BarcodeBERT, we present **BarcodeMamba, an efficient and performant foundation model** that leverages the state-of-the-art Mamba-2 architecture to advance biodiversity analysis. Our comprehensive evaluation demonstrates BarcodeMamba's superior performance in both known species classification (99.2% accuracy) and zero-shot identification of unknown species (70.2% genus-level accuracy), **while using only 8.3% of BarcodeBERT's parameters**. Through extensive experiments comparing architectures, analyzing components, and studying scaling behaviours, we show BarcodeMamba's potential for accelerating biodiversity research.

2. Background

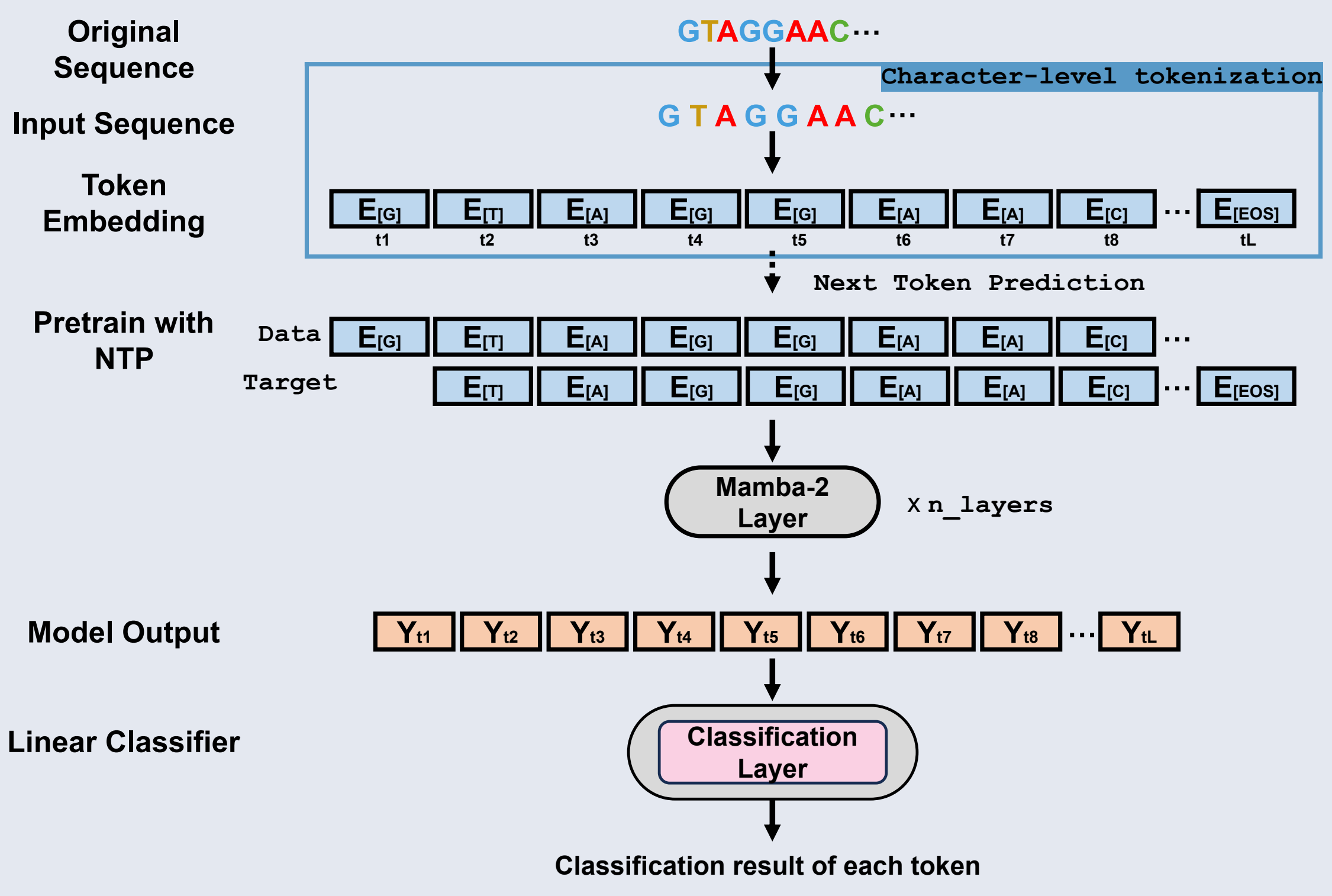
- **Mamba** introduces selective state processing to sequence modeling, allowing it to efficiently handle important information while filtering out noise — a key capability for analyzing complex biological sequences.
- **Mamba-2** enhances this foundation by integrating state space modeling and attention mechanisms, enabling better capture of relationships between different parts of DNA sequences.
- These advances are particularly valuable for DNA barcoding where:
  - Missing or uncertain nucleotides create gaps in sequences
  - Large-scale processing requires efficient hardware-aware computation
  - Complex patterns need to be recognized across different species

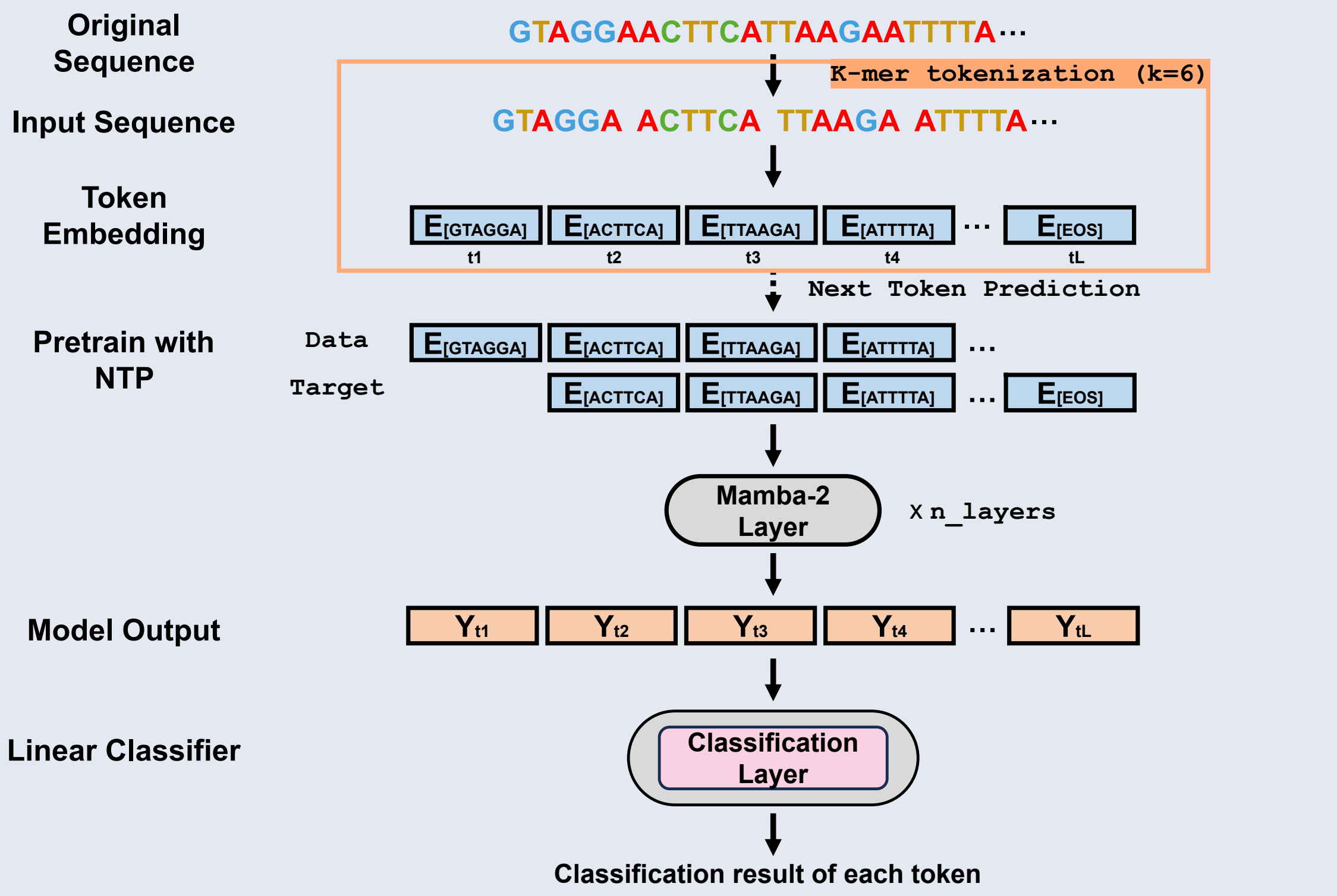
3. Method: Design choices

Data augmentation. Reverse Complement (RC) data augmentation during pretraining

Tokenization. Char-level: learning at single nucleotide resolution; *k*-mer: capturing local patterns

Pretraining objectives. Next token prediction (NTP), preferred by causal models; Masked language modeling (MLM), successfully applied in BarcodeBERT and Caduceus (built upon MambaDNA blocks)


Original Sequence
GTAGGAAC···
Character-level tokenization
Input Sequence
G T A G G A A C···
Token Embedding
E[G] E[T] E[A] E[G] E[G] E[A] E[A] E[C] ··· E[EOS]
t1 t2 t3 t4 t5 t6 t7 t8 tL
Next Token Prediction
Pretrain with NTP
Data
E[G] E[T] E[A] E[G] E[G] E[A] E[A] E[C] ···
Target
E[T] E[A] E[G] E[G] E[A] E[A] E[C] ··· E[EOS]
Mamba-2 Layer
X n_layers
Model Output
Yt1 Yt2 Yt3 Yt4 Yt5 Yt6 Yt7 Yt8 ··· YtL
Linear Classifier
Classification Layer
Classification result of each token



4. Comparison with baselines

	Species-level acc (%) of seen species		Genus-level acc (%) of unseen species	
Model	Fine-tuned	Linear probe	1-NN probe	Params
DNABERT-2	98.3	87.2	40.9	118.9 M
DNABERT	(*k*=6) 97.4	(*k*=4) 47.1	(*k*=6) 48.5	88.1-91.1 M
Caduceus-PS-131k	97.6	5.1	21.1	14.0 M
Caduceus-PH-131k	96.7	2.7	19.3	14.0 M
Caduceus-PS-1k	98.8	16.8	31.4	3.5 M
Caduceus-PH-1k	98.8	6.2	23.1	3.5 M
HyenaDNA-small	98.5	75.2	46.1	3.3 M
HyenaDNA-tiny	**99.1**	93.5	47.0	1.6 M
CNN encoder	98.2	51.8	47.0	1.8 M
BarcodeBERT	(*k*=6) 98.1	(*k*=4) 93.0	(*k*=5) 58.4	86.2-89.2 M
BarcodeMamba-2-large (ours)	(*k*=6) 97.7	(*k*=1) **99.2**	(*k*=6) **70.2**	50.4-56.7 M
BarcodeMamba-2-mini (ours)	(*k*=1) 97.7	(*k*=1) **99.2**	(*k*=6) 63.2	4.3-7.4 M

We evaluate BarcodeMamba's performance through three increasingly challenging tasks:

1. **Traditional Classification** (Fine-tuned): Full model training for known species id.
2. **Representation Quality** (Linear probe): Tests learned representations with a simple classifier.
3. **Unknown Species Detection** (1-NN probe): Identifies new species to the genus level.

In the more challenging test of SSL-trained representations (**Linear Probe** and **1-NN Probe**), improving on BarcodeBERT, our results demonstrate a **substantial improvement** compared to all other models. Our BarcodeMamba model exhibits **superior performance to BarcodeBERT with less than 7.4 M parameters** (vs. 86.2–89.2 M) demonstrating both effectiveness and efficiency.

5. Ablation study

NTP-pretrained

		Species-level acc (%) of seen species				Genus-level acc (%) of unseen species		Representation of unseen barcodes	
		Fine-tuned		Linear probe		1-NN probe		Perplexity	
Tokenizer	*k*	Mamba	Mamba-2	Mamba	Mamba-2	Mamba	Mamba-2	Mamba	Mamba-2
Char	-	**98.7**	**98.1**	**97.0**	**95.9**	41.2	33.0	1.41	1.37
k-mer	4	95.0	97.4	92.9	94.0	43.5	55.3	3.19	3.09
k-mer	5	94.2	95.6	91.5	92.6	**48.5**	57.7	4.16	4.04
k-mer	6	95.9	96.5	91.8	91.9	47.7	**58.7**	5.51	5.31

MLM-pretrained

		Species-level acc (%) of seen species				Genus-level acc (%) of unseen species		Representation of unseen barcodes	
		Fine-tuned		Linear probe		1-NN probe		Perplexity	
Tokenizer	*k*	Mamba	Mamba-2	Mamba	Mamba-2	Mamba	Mamba-2	Mamba	Mamba-2
Char	-	88.4	**98.2**	91.8	91.5	32.1	38.7	1.23	1.22
k-mer	4	**97.3**	96.6	**94.0**	**94.3**	47.4	50.4	1.89	1.86
k-mer	5	97.1	97.5	92.9	93.1	52.2	**51.9**	2.20	2.17
k-mer	6	96.7	95.4	92.7	92.7	**54.5**	51.0	2.46	2.45

Architecture:

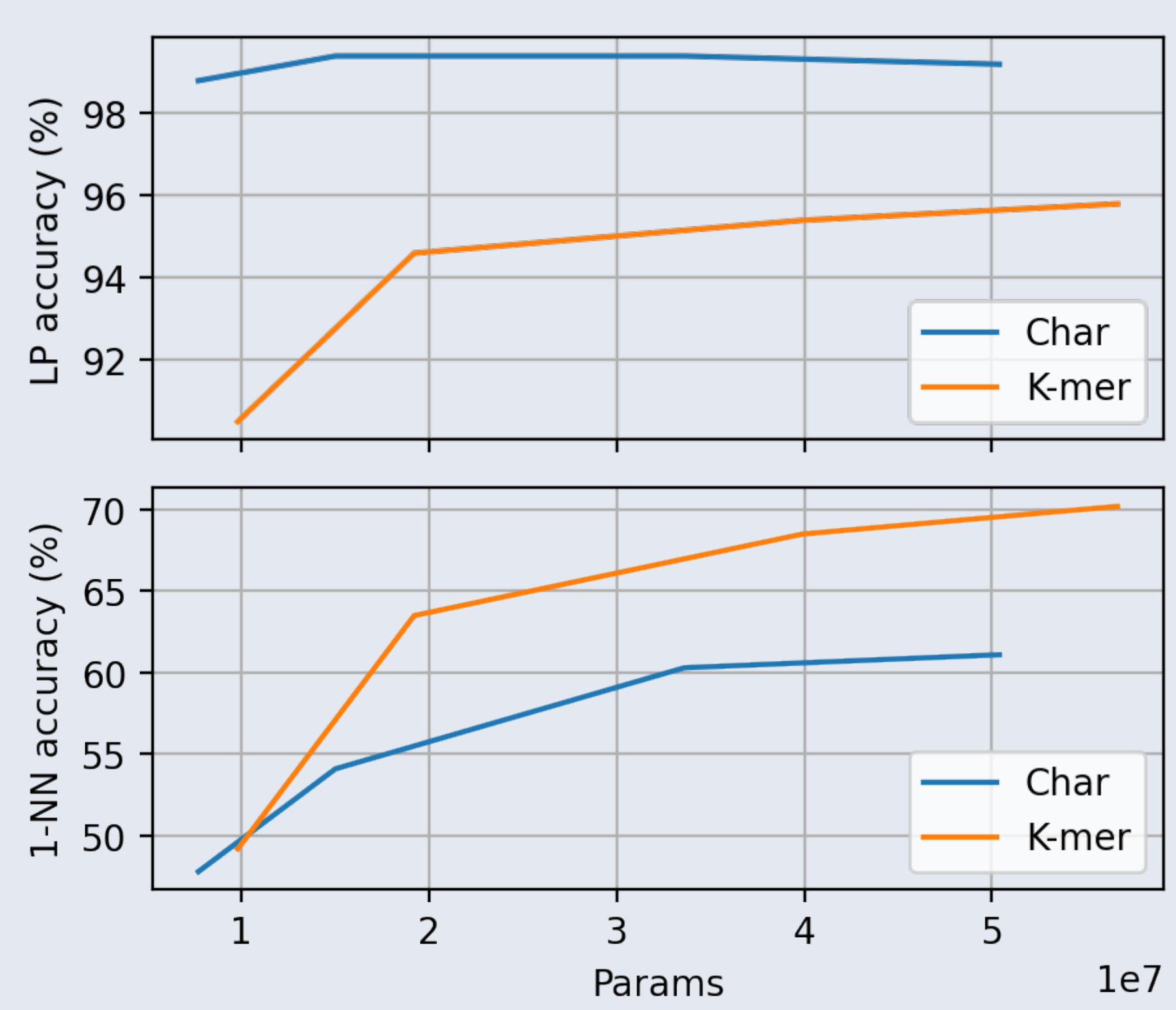
- For both pretraining tasks, Mamba-2 performs better as the mixing layer.

Tokenization:

- The character-level tokenizer enhances the known species classification of BarcodeMamba (Fine-tuned, Linear Probe).
- For 1-NN probing, *k*-mer tokenization enables BarcodeMamba to achieve significantly better results.

Overall, both tokenizers demonstrate that next token prediction (NTP) consistently outperforms masked language modeling (MLM) for BarcodeMamba.

6. Scaling study



Performance Scaling on NTP pretraining:

- Linear probe accuracy reaches 99.4% at 30M parameters
- Unknown species detection (1-NN) improves steadily to 70.2% at 56.7M parameters
- Both tokenization strategies (character-level and *k*-mer, *k*=6) show consistent improvements with scale

Key Findings:

- Larger models particularly benefit unknown species detection
- Performance gains continue even at largest tested size (56.7M param)
- Maintains efficiency advantage over BarcodeBERT across all scales

Future Impact: Scaling behavior suggests potential for improving global species identification. Future work will evaluate performance beyond Canadian invertebrates.

7. Conclusions

Key Achievements:

- Successfully applied state space models to DNA barcode analysis, improving BarcodeBERT's taxonomic classification performance with only 8.3% of its parameters.

Impact for Biodiversity Science:

- Enables more efficient processing of DNA barcodes for taxonomic classification
- Improves accuracy of both known and unknown species identification
- Supports BIOSCAN's mission of global biodiversity monitoring

Future Directions:

- Scale to BIOSCAN-5M dataset (5M arthropod specimens)
- Explore bi-directional architectures for improved accuracy
- Develop robust variants for additional barcode markers (e.g., fungal ITS)

Scan the code to see our repository

8. Acknowledgements

BIOSCAN is supported in part by funding from the Government of Canada's New Frontiers in Research Fund (NFRF).