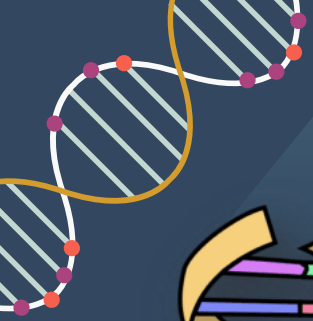
DNA:

A New Language for LLMs to Learn

Presenter: Austin Wang
Date: March 30, 2026
CMPT-413

Credit to Chuanqi Tang for the original slides



DNA

=Adenine

= Thymine

= Cytosine

= Guanine

= Phosphate backbone

DNA

a language with just four letters:

A / T / C / G

Natural Language	DNA
Letters (a, b, c…)	**Bases** (A, T, C, G)
Words	**k-mers** (ATG, GTC,…)
Sentence	**DNA barcode**

Can we teach machines to understand DNA?

If LLMs Could Understand DNA…

What Could We Do? 🤯

Predict DNA expression

Functional annotation of unknown regions

Model species relationships

Phylogenetic patterns from sequences

Accelerate species discovery

Classify new organisms at scale — fast.

Accelerate species discovery

Classify new organisms at scale — fast.

~2.3 million known species — insects alone: **1 million+**

Estimated total: **8–10 million**, maybe **100 million**

That means **>80%** of life remains unknown

1. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How Many Species Are There on Earth and in the Ocean? *PLoS Biology*, 9(8), e1001127. https://doi.org/10.1371/journal.pbio.1001127
2. Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21), 5970–5975. https://doi.org/10.1073/pnas.1521291113
3. Wiens, J. J. (2022). How many species are there on Earth and how many are left to describe? *PLoS Biology*, 20(7), e3001760. https://doi.org/10.1371/journal.pbio.3001760

BarcodeBERT: Transformers for Biodiversity Analyses

Pablo Millan Arias[1,*], Niousha Sadjadi[1,*], Monireh Safari[1,*],
ZeMing Gong[3,†], Austin T. Wang[3,†], Joakim Bruslund Haurum[6], Iuliia Zarubiieva[2,4],
Dirk Steinke[2], Lila Kari[1,♯], Angel X. Chang[3,5], Scott C. Lowe[4,‡], and Graham W. Taylor[2,4,‡,♯]

[1]University of Waterloo
[2]University of Guelph
[3]Simon Fraser University
[4]Vector Institute
[5]Alberta Machine Intelligence Institute (Amii)
[6]Aalborg University and Pioneer Centre for AI

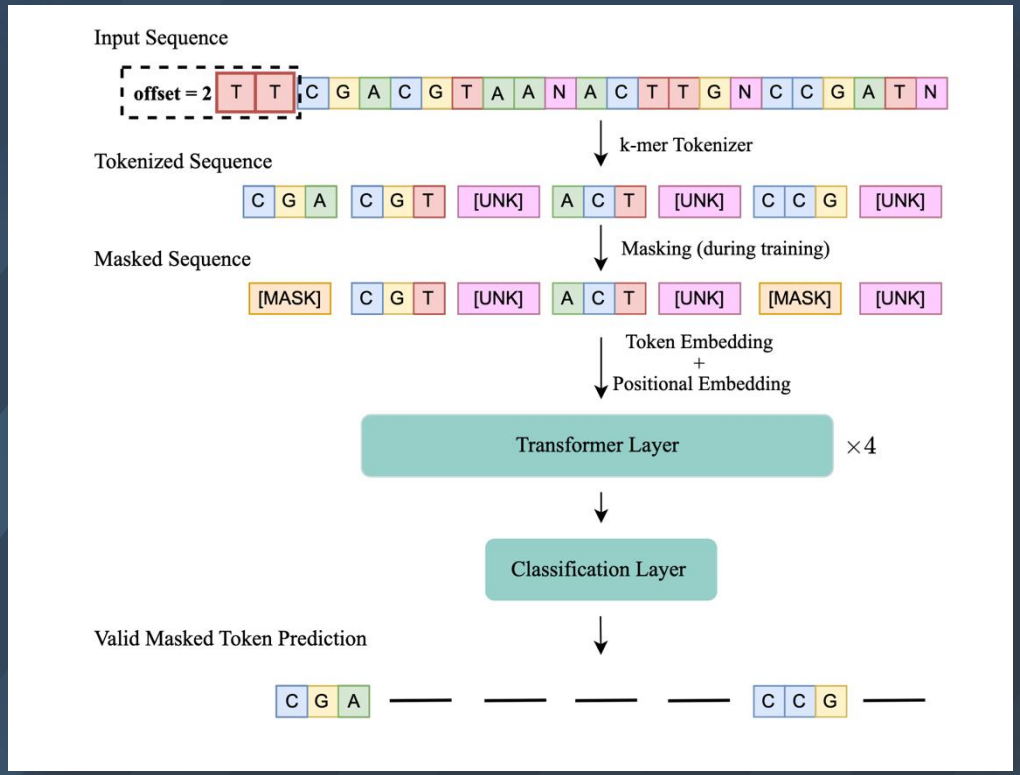
*Joint first author
†Joint second author
‡Joint senior author
♯Corresponding authors: gwtaylor@uguelph.ca, lila@uwaterloo.ca

A transformer-based language model from NLP

BarcodeBERT: Transformers for Biodiversity Analyses

- A short, standardized DNA sequence
- Works like a biological “ID code”
- Predictive of species

Architecture of BarcodeBERT



Input: DNA barcode sequence

k-mer Tokenization = making "DNA words"

Masked tokens = the blanks the model must learn to fill

Transformer layers = learning context and structure

Output: predicted DNA tokens

Performance Comparison of BarcodeBERT and Baseline Models

Model	#Param.	TPS (seq/s)	Species-level acc (%) of seen species			Genus-level 1-NN probe of unseen species		BIN reconstruction accuracy (%)
			Finetuned	Linear probe	Dur (s)	Acc (%)	Dur (s)	ZSC probe
BLAST	N/A	N/A	**99.7***		1495	**83.9**	602	N/A
CNN encoder	1.8 M	934	98.2	51.8	13	47.0	55	26.8
DNABERT	88.1 M	50	(k=6) 99.5	(k=4) 47.1	248	(k=6) 48.1	1021	**79.3**
DNABERT-2	118.9 M	134	**99.7**	87.2	101	23.5	381	38.1
DNABERT-S	117.1 M	134	**99.7**	93.1	101	30.6	381	62.7
HyenaDNA-tiny	1.6 M	**1167**	99.2	93.5	**11**	37.5	**44**	25.8
Nucleotide Transformer	55.9 M	95	99.5	65.1	140	40.1	536	22.4
BarcodeBERT (4-4-4)	29.1 M	484	**99.7**	**99.0**	27	78.5	108	73.2

55× faster

BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research

Tiancheng Gao[1,2] **Scott C. Lowe**[2] **Brendan Furneaux**[3] **Angel X. Chang**[4,5]
Graham W. Taylor[1,2]*

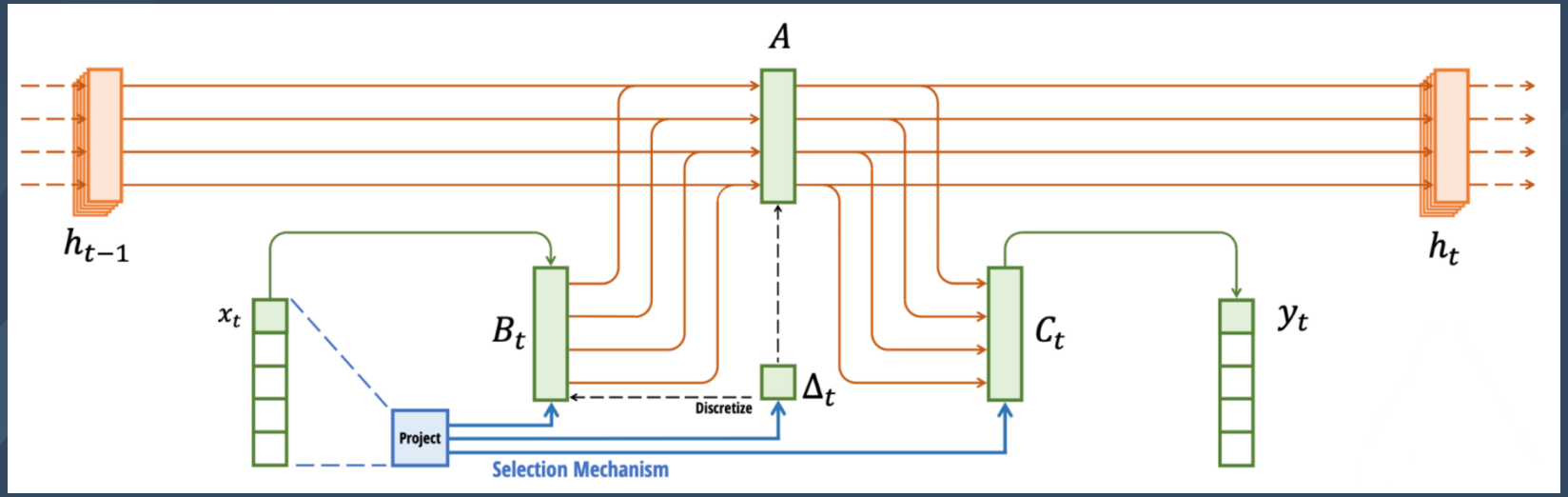
[1]University of Guelph [2]Vector Institute [3]University of Jyväskylä
[4]Simon Fraser University [5]Amii

BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research

Tiancheng Gao[1,2] **Scott C. Lowe**[2] **Brendan Furneaux**[3] **Angel X. Chang**[4,5]
Graham W. Taylor[1,2]*

[1]University of Guelph [2]Vector Institute [3]University of Jyväskylä
[4]Simon Fraser University [5]Amii

SSMs achieve linear runtime



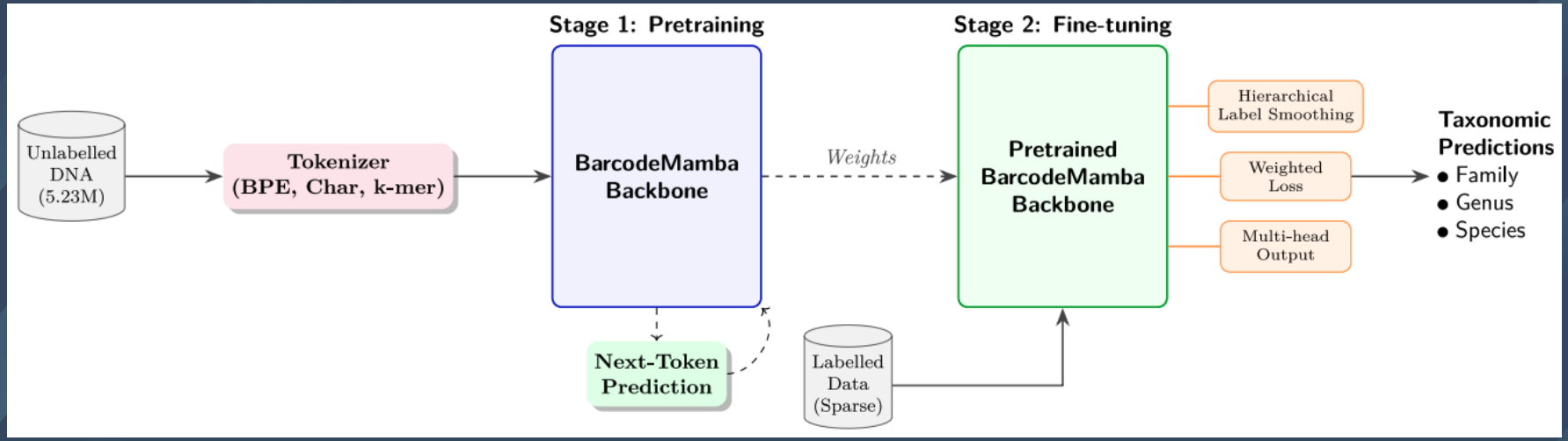
$$h_t = A_t h_{t-1} + B_t x_t$$

$$y_t = C_t^T h_t$$

Scale better with sequence length relative to attention/transformers

Gao, Tiancheng, et al. ”BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research." *ICLR* (2025).

BarcodeMamba+ learns a representation through NTP



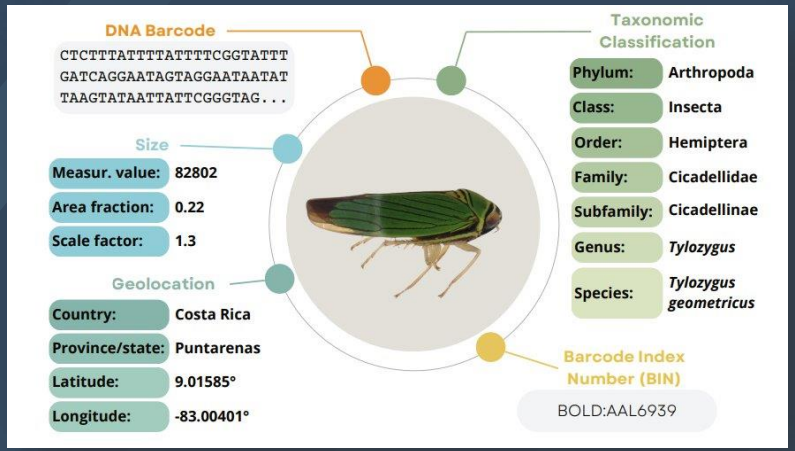
Gao, Tiancheng, et al. ”BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research." *ICLR* (2025).

BarcodeMamba+ achieves superior performance

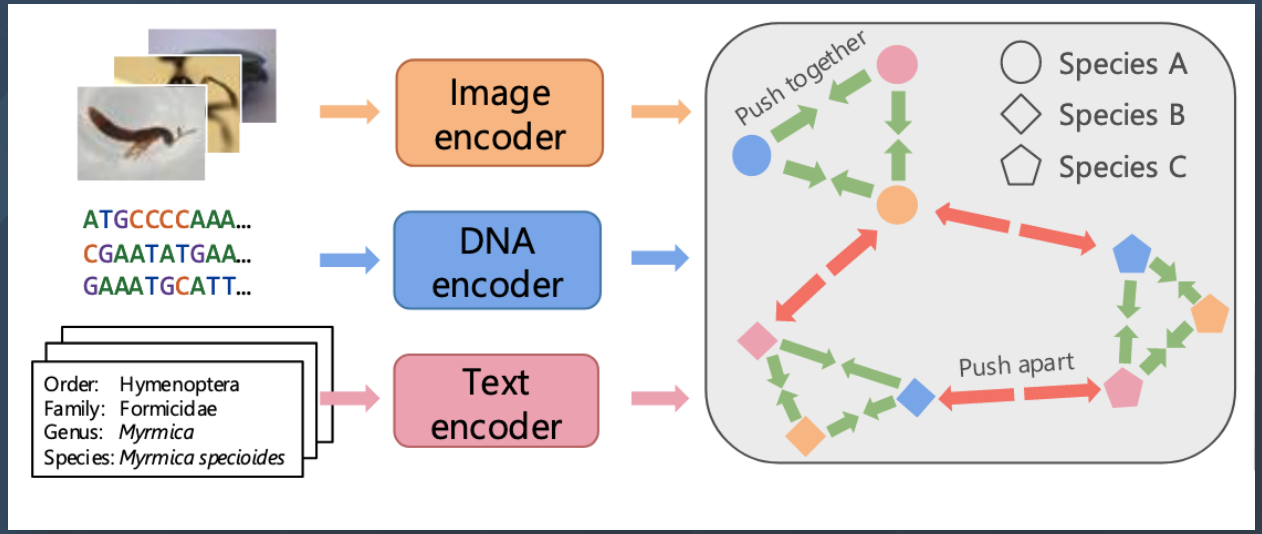
	Yeast Acc. (%)↑			Filamentous Acc. (%)↑			MycoAI Acc. (%)↑				
Model	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Size ↓	Time ↓
BLAST	86.6	92.9	75.4	81.4	71.5	33.4	94.7	93.1	55.0	N/A	208.6 ms
MycoAI-CNN (Vu)	90.5	86.4	60.0	84.1	69.8	28.2	93.9	87.8	57.1	11.6 M	11.8 ms
MycoAI-BERT (base)	88.9	75.7	33.5	85.1	60.8	16.6	93.2	80.3	39.3	18.4 M	4.5 ms
CNN Encoder	94.1	88.3	67.6	84.5	69.1	31.4	97.5	93.6	72.6	12.1 M	5.8 ms
BarcodeBERT	95.4	88.6	59.1	87.8	70.2	27.7	97.8	92.0	58.9	44.6 M	8.8 ms
BarcodeMamba+	98.7	95.3	80.6	**92.6**	81.1	46.5	99.0	96.5	81.7	12.1 M	8.0 ms
BarcodeMamba+ (large)	**98.8**	**95.9**	**83.6**	92.5	**81.6**	**50.4**	**99.3**	**97.7**	**88.9**	49.2 M	14.7 ms

Gao, Tiancheng, et al. ”BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research." *ICLR* (2025).

Multimodal Learning

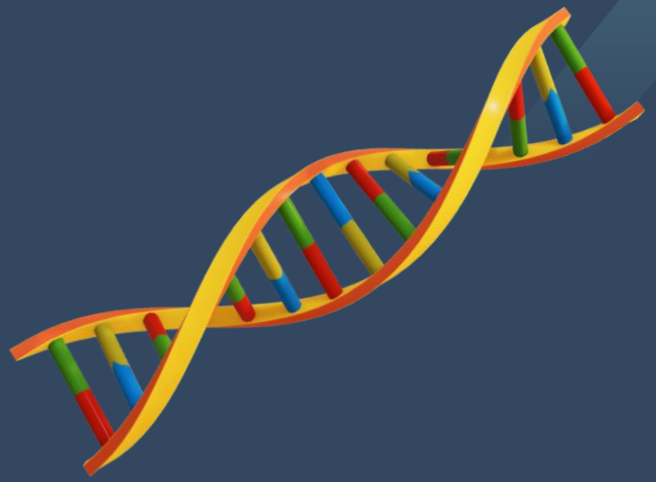


CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale



Gong, ZeMing, et al. "CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale." *ICLR* (2025).

Conclusion



1. **DNA is a language**
2. **Language models like BarcodeBERT** can learn this language
3. **Multimodal AI** can leverage information across multiple modalities to form a better understanding of insects.