DNA:

A New Language for LLMs to Learn

Presenter: Chuanqi Tang
Date: March 26, 2025

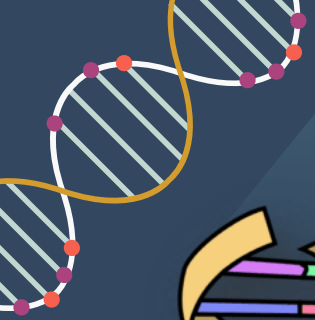
Gemini

Claude

LLMs can understand English, Chinese, code…

Can it understand

life?



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**

DNA is a language — and we're teaching machines to read it.

If LLMs Could Understand DNA…

What Could We Do? 🤯

🧬 **Predict what parts of DNA do**

Functional annotation of unknown regions

🌳 **Model how species are related**

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

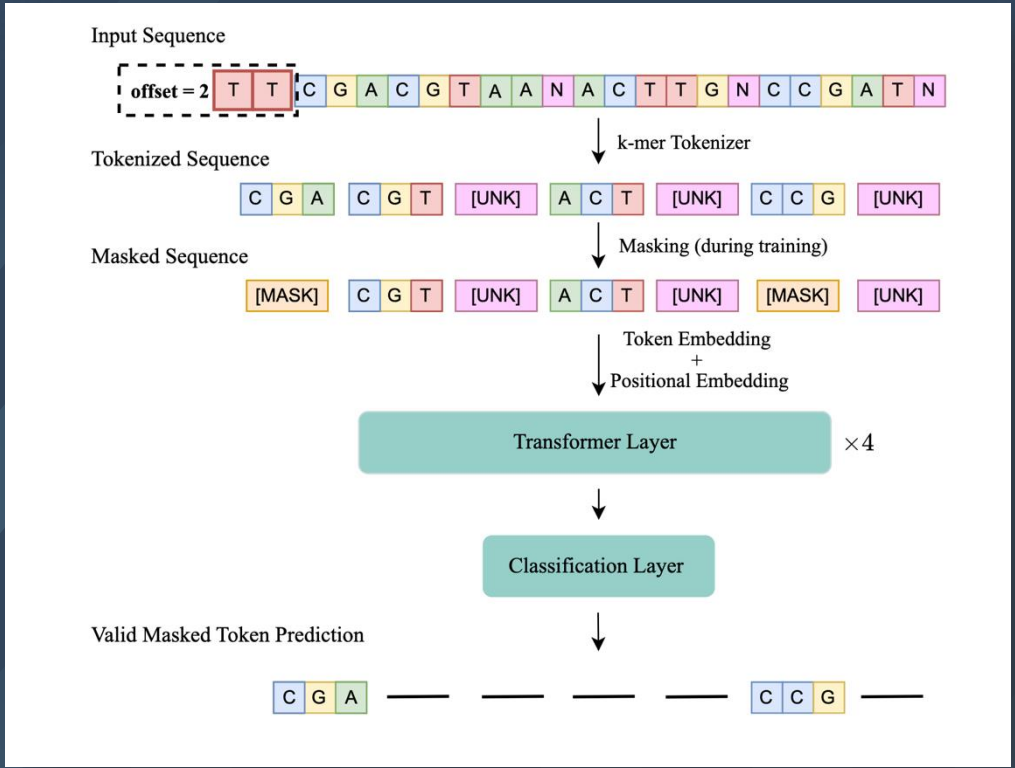
BarcodeBERT: Transformers for Biodiversity Analyses

A transformer-based language model from NLP

BarcodeBERT: Transformers for Biodiversity Analyses

- A short, standardized DNA sequence
- Works like a biological “ID code”
- Used to identify 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

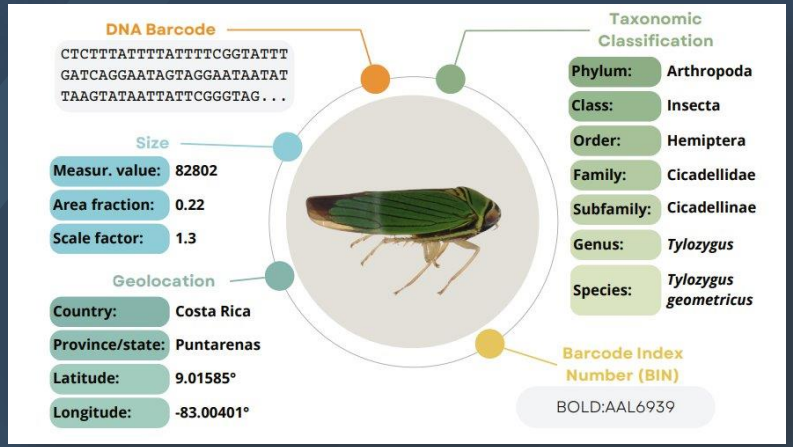
Output: predicted DNA tokens

Performance Comparison of BarcodeBERT and Baseline Models

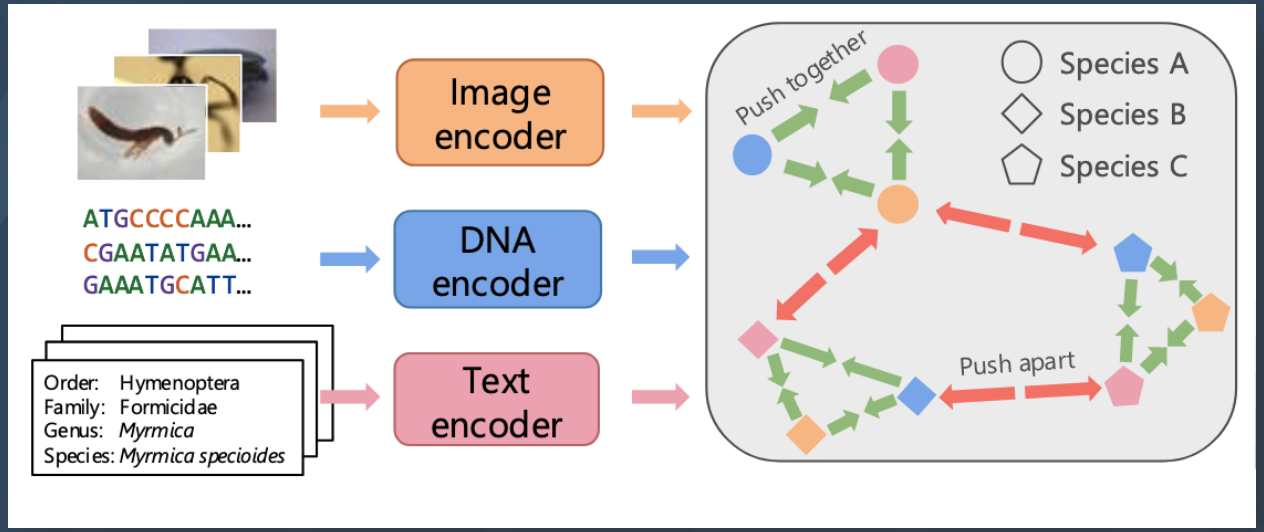
			Species-level acc (%) of seen species			Genus-level 1-NN probe of unseen species		BIN reconstruction accuracy (%)
Model	#Param.	TPS (seq/s)	Finetuned	Linear probe	Dur (s)	Acc (%)	Dur (s)	ZSC probe
BLAST	N/A	N/A	**99.7***		1495	**83.9**	602	N/A
CNN encoder	1.8 M	934	98.2	51.8	13	47.0	55	26.8
DNABERT	88.1 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

Multimodal Learning

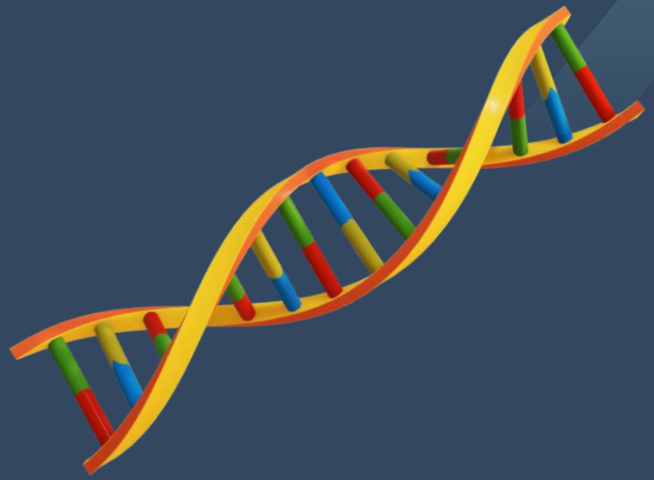


CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale



Gong, ZeMing, et al. "CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale." *arXiv preprint arXiv:2405.17537* (2024).

Conclusion



1. **DNA is a language**
2. **LLMs like BarcodeBERT** can learn this language
3. **Multimodal AI** takes it further

Thanks!

Do you have any questions?