# BarcodeBERT: Transformers for biodiversity analysis

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

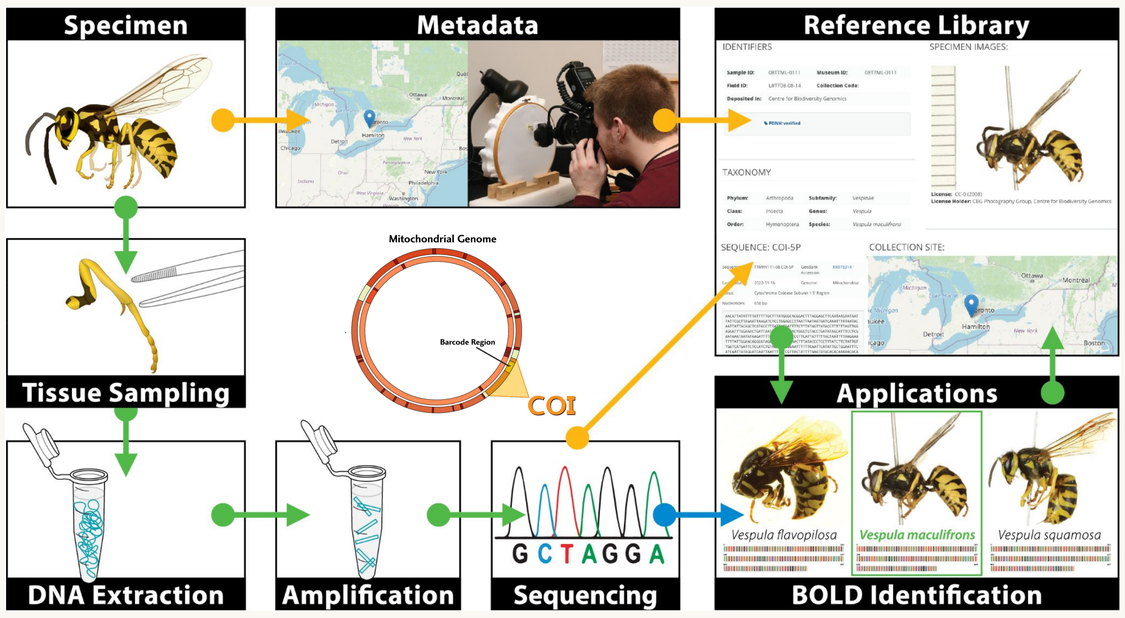
[1]University of Waterloo, [2]University of Guelph, [3]Simon Fraser University, [4]Vector Institute for AI,

[5]Alberta Machine Intelligence Institute (Amii), [6]Aalborg University and Pioneer Centre for AI

[*]Joint first author. [†]Joint second author. Correspondence: Graham W. Taylor (gwtaylor@uoguelph.ca)

NEURAL INFORMATION PROCESSING SYSTEMS

## 1. Introduction



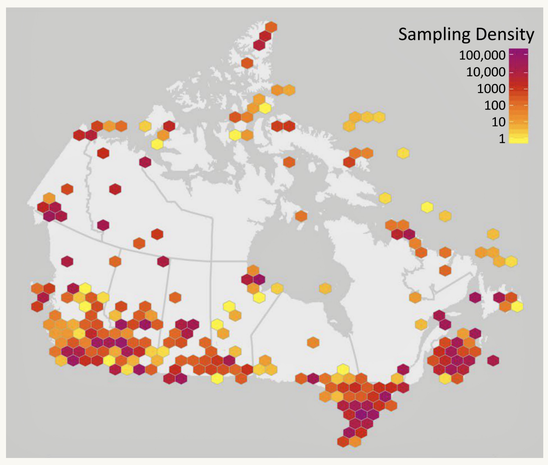

**DNA barcodes** play a crucial role in the global effort to understand **biodiversity**, particularly among **invertebrates**, a diverse and under-explored group that poses unique taxonomic challenges. Here, we explore several machine learning approaches, comparing supervised CNNs, fine-tuned foundation models, and a DNA barcode-specific pretrained model across different classification tasks. While supervised CNNs excel on simpler datasets, challenging taxonomic identification demands a shift to self-supervised pretrained models. We propose BarcodeBERT, the first self-supervised method for biodiversity analysis, leveraging a 1.5 M invertebrate DNA barcode reference library. **This work emphasizes the impact of considering particular genic regions and broader taxonomic coverage on the performance of genomic tasks.** Specific self-supervised pretraining proves crucial for achieving high-accuracy DNA barcode-based identification. Notably, without fine-tuning, BarcodeBERT pretrained on a large DNA barcode dataset outperforms DNABERT (Ji et al., 2021) and DNABERT-2 (Zhou et al., 2023) on multiple downstream classification tasks.

## 2. Motivation

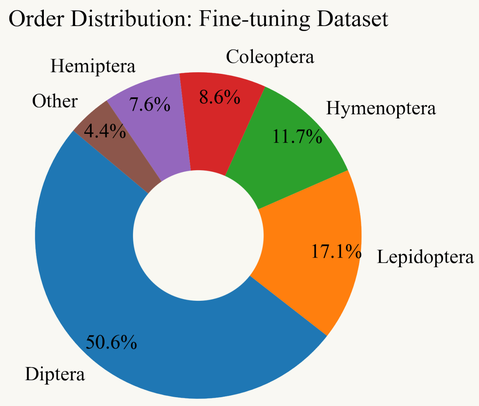
- Aligned with BIOSCAN's ambitious goal, we join researchers across disciplines to contribute to **cataloging all multicellular life** on Earth.
- The Barcode of Life Database (BOLD) provides unparalleled biodiversity data for neural algorithm development.
- Leveraging BOLD's diverse data modalities, (Badirli et al., 2021) performs **Bayesian zero-shot learning over insect images using DNA embeddings**, we explore self-supervised methods for encoder training.

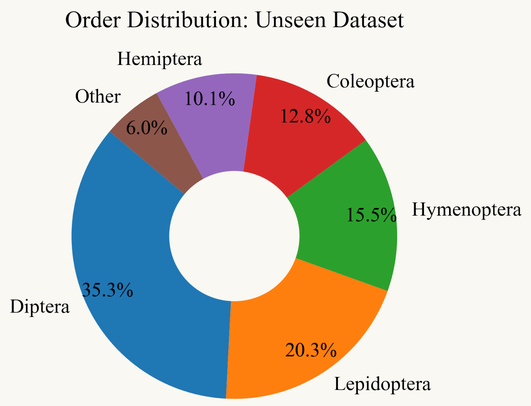
## 3. Dataset

Our primary data source was the Reference Library for Canadian Invertebrates (deWaard et al., 2019). We also used the INSECT dataset (Badirli et al., 2021) to benchmark against prior work.




The library contains 1.5 M DNA samples from across Canada.



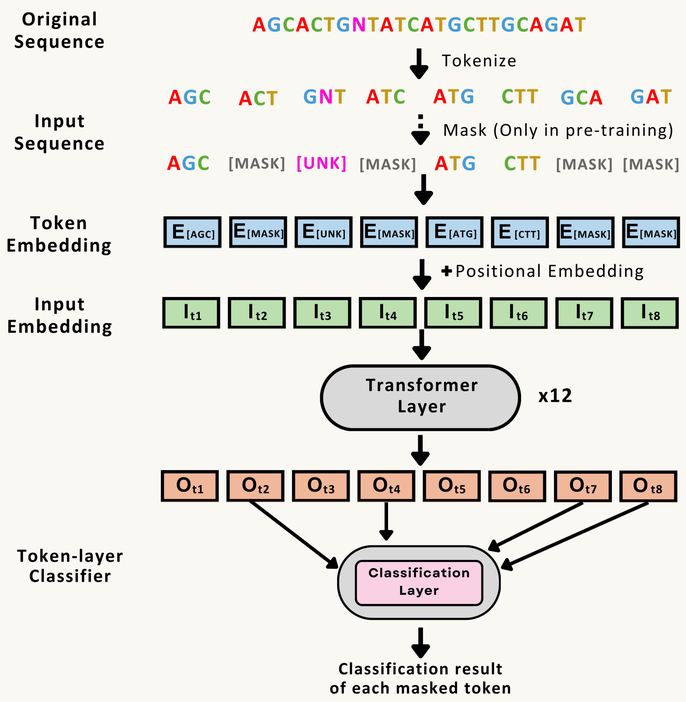




A fine-tuning subset was selected for classification of known species; the Unseen subset mimics species unknown to science, testing model adaptability to novel taxa.

## 4. Transformer models and DNA

Self-supervised learning (SSL) has led to the development of various models for genomic data analysis:

- **DNABERT:** Pretrained on the human genome dataset, using mask (15% of tokens) and same-segment prediction tasks.
- **DNABERT-2:** Adopts Byte Pair Encoding for variable-length tokenization and is pretrained on a large-scale multi-species genome dataset.
- Our proposed model **BarcodeBERT** (below) performs masked pretraining on barcodes:
  - Non-overlapping tokenizer
  - Increased masking ratio (50%)
  - Elimination of same-segment prediction task




## 5. Evaluation of embedding quality (DNA only)

We assess the performance of each model against the CNN baseline (Badirli et al., 2021) on several tasks:

- Fine-tuning for species-level classification
- 1-NN probing at the genus level of unseen species
- Linear probing for species-level classification

| | Species-level acc (%) of seen species | | | | | | Genus-level acc (%) of unseen species | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Fine-tuned | | | Linear-probe | | | 1-NN probe | | |
| CNN baseline | 98.2 | | | 51.8 | | | 47.0 | | |
| DNABERT-2 | **98.3** | | | 87.2 | | | 40.9 | | |
| $k$-mer length | $k=4$ | $k=5$ | $k=6$ | $k=4$ | $k=5$ | $k=6$ | $k=4$ | $k=5$ | $k=6$ |
| DNABERT | 96.3 | 96.9 | 97.4 | 47.1 | 38.4 | 41.2 | 38.2 | 41.6 | 48.5 |
| BarcodeBERT (ours) | 97.6 | 97.0 | 98.1 | **93.0** | 88.6 | 84.0 | 49.0 | **58.4** | 57.6 |

BarcodeBERT and DNABERT-2 outperform the baseline and DNABERT across all tasks. For models with variable stride length, we present results across multiple $k$-mer lengths.

## 6. Bayesian Zero-Shot Learning downstream task (Multimodal)

BarcodeBERT embeddings on the INSECT dataset were assessed using K-nearest seen classes in DNA space with image-based priors. The results were compared to the baselines.

| | Data sources | | Species-level acc (%) | | |
|---|---|---|---|---|---|
| Model | SSL pretraining | Fine-tuning | Seen | Unseen | Harmonic Mean |
| CNN encoder | – | Insect | 38.3 / 39.4 | 20.8 / 18.9 | **27.0** / 25.5 |
| DNABERT | Human | – | 35.0 | 10.3 | 16.0 |
| DNABERT | Human | Insect | **39.8** | 10.4 | 16.5 |
| DNABERT-2 | Multi-species | – | 36.2 | 10.4 | 16.2 |
| DNABERT-2 | Multi-species | Insect | 30.8 | 8.6 | 13.4 |
| BarcodeBERT (ours) | Arthropod | – | 38.4 | 16.5 | 23.1 |
| BarcodeBERT (ours) | Arthropod | Insect | 37.3 | **20.8** | **26.7** |

BarcodeBERT excels in zero-shot insect species prediction, surpassing DNABERT and matching the CNN baseline, original paper result (left) and reproduced result (right).

## 7. Conclusions

- **Pretraining** masked language models on DNA barcode data is **effective and essential** for insect species identification.
- **Diversification** of large pretraining datasets **beyond human DNA** sequences is crucial to advancing biodiversity genomics.
- Strides have been made in improving classification from DNA sequences and images, but untapped data sources, like the 1.5 M DNA barcodes in the BOLD dataset, offer significant potential.
- Findings highlight the need for **continuous augmentation** with data from both previously **seen and unseen** species.

**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).