BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity

Zahra Gharaee[3*] Scott C. Lowe[5*] ZeMing Gong[4*] Pablo Millan Arias[3*]
Nicholas Pellegrino[3] Austin T. Wang[4] Joakim Bruslund Haurum[7] Iuliia Zarubiieva[2,5] Lila Kari[3]
Dirk Steinke[1,2†] Graham W. Taylor[2,5†] Paul Fieguth[3†] Angel X. Chang[4,6†]

[1]Centre for Biodiversity Genomics [2]Uni. of Guelph [3]Uni. of Waterloo [4]Simon Fraser Uni. [5]Vector Institute [6]Amii [7]Aalborg Uni. and Pioneer Centre

As part of ongoing **worldwide** efforts to comprehend and **monitor insect biodiversity**, this paper presents the BIOSCAN-5M Insect dataset to the machine learning community and establishes several benchmark experiments.

BIOSCAN-5M: Key distinguishing features

Data Volume	**Multimodal** information for over **5 million insect** specimens, significantly expanding existing image-based biological datasets.
RGB Images	High-resolution **microscopy images** (120 pixels per mm).
DNA Barcodes	Unique 660-base-pair sequences for automated species identification, with each image **paired** to a DNA barcode.
Taxonomic Labels	**Hierarchical** annotations to **fine-grained** species level, plus subspecies-like barcode index numbers (BINs).
Geographic Data	Specimen collection sites enable **species distribution modelling**.
Size Data	Supporting **biodiversity analysis** and ML models for classification and **species comparison**.
Unlabelled Data	Partial species-level annotations make the dataset well-suited for **semi-supervised learning**.

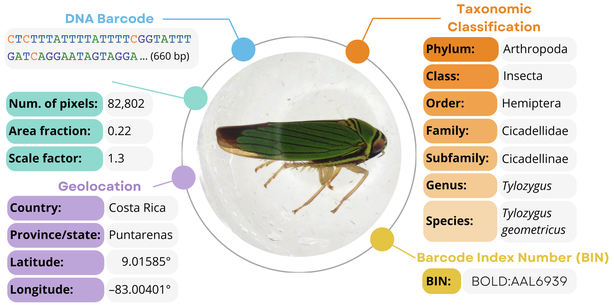


Figure 1: Multiple features associated with a BIOSCAN-5M dataset record.

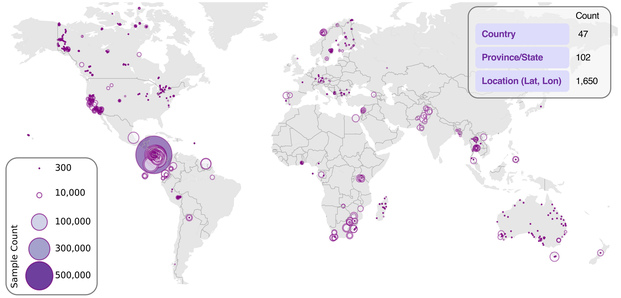


Figure 2: Geographic Data. Distribution of sample collection locations.

BIOSCAN-5M: Data partitioning

We partition the data into seen and unseen species, used for **closed-world** and **open-world** settings. New species are frequently discovered—it is estimated that 80% of insect species are as yet undiscovered—making open-world categorization highly relevant. **Seen:** samples labelled with an established scientific species name. **Unseen:** samples with a placeholder species name.

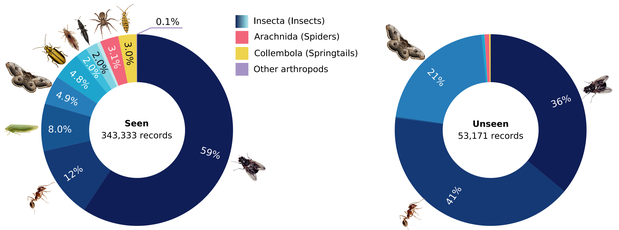
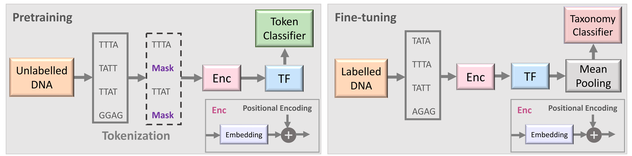


Figure 3: Distribution of class- and (Insecta) order-level taxa for seen and unseen data partitions. The distributions reflect that of known and newly-discovered species, respectively.

Task I: DNA Barcode taxonomic classification



Pretraining: DNA sequences are tokenized into non-overlapping k-mers, with 50% masked for a *masked language modelling* task. Tokens are encoded via a transformer model, and the output embeddings are used for token-level classification.

Downstream tasks: Tokenized DNA sequences (w/o masking) are processed by the pretrained transformer. Mean-pooled token embeddings provide representations for taxonomic classification, refined with fine-tuning.

Task I: Results

Table 1: Performance of DNA-based sequence models in closed- and open-world settings. For the closed-world setting, we show the species-level accuracy (%) for predicting seen species, for open-world the genus-level accuracy (%) for unseen species, while using seen species to fit the model. Best and second best results are highlighted.

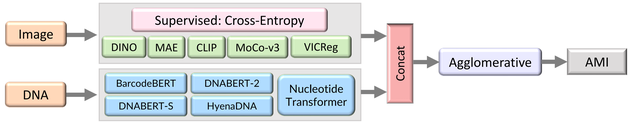
				Seen: Species		Unseen: Genus
Model	Architecture	SSL-Pretraining	Tokens seen	Fine-tuned	Linear probe	1NN-Probe
CNN baseline	CNN	–	–	97.70	–	29.88
NT	Transformer	Multi-Species	300 B	98.99	52.41	21.67
DNABERT-2	Transformer	Multi-Species	512 B	99.23	67.81	17.99
DNABERT-S	Transformer	Multi-Species	∼1,000 B	98.99	95.50	17.70
HyenaDNA	SSM	Human DNA	5 B	98.71	54.82	19.26
BarcodeBERT	Transformer	DNA barcodes	5 B	98.52	91.93	23.15
Ours	Transformer	DNA barcodes	7 B	99.28	94.47	47.03

BIOSCAN-5M: Data statistics

Table 2: Dataset statistics. For each taxonomic rank, we show three key statistics. First, *labelling* statistics—the number and percentage of records labelled to a rank. Second, the *category* metrics—the num. unique labels for a rank, and imbalance ratio (IR) highlighting disparities between most and least populated groups. Lastly, *diversity* insights show the num. unique DNA barcodes across samples labelled to a rank, and average Shannon diversity indices (SDI), calculated as the entropy of the distribution of DNA barcodes across groups.

	Labelled		Categories		Unique DNA	
Level	Count	%	Count	IR	Count	Avg SDI
Phylum	5,150 k	100.0	1	1	2,486 k	19.78
Class	5,146 k	99.9	10	719 k	2,482 k	8.56
Order	5,134 k	99.7	55	3,675 k	2,474 k	7.05
Family	4,932 k	95.8	934	938 k	2,231 k	5.42
Subfamily	1,472 k	28.6	1,542	323 k	657 k	4.28
Genus	1,226 k	23.8	7,605	200 k	531 k	2.63
Species	473 k	9.2	22,622	7,690	202 k	1.46
BIN	5,137 k	99.7	324,411	35 k	2,474 k	1.29

Task II: Zero-shot transfer learning



How well can pretrained models cluster the data samples without retraining on our data? We embed images or DNA barcodes with a pretrained network (i.e. trained on ImageNet), reduce to 50-d with UMAP, then cluster with Agglomerative Clustering. This experiment lays an initial baseline for models aiming to categorize unseen species.

Task II: Results

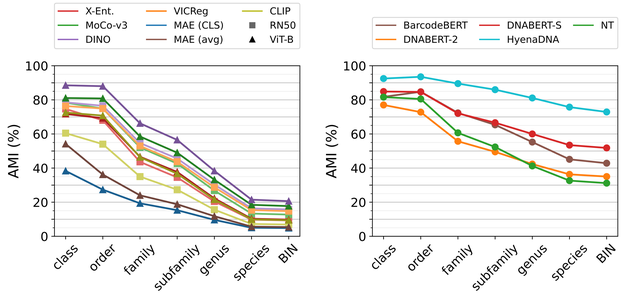
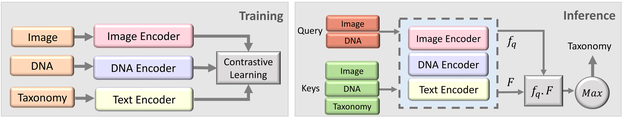


Figure 4: Zero-shot clustering AMI (%) performance across taxonomic ranks. For images (left), pretrained encoders only capture coarse-grained information, but with DNA barcodes (right), clustering yields high performance to species-level, even without model retraining.

Task III: Multimodal retrieval learning



Training: Multimodal features—RGB images, textual taxonomy, and DNA sequences—are independently encoded and aligned with a CLIP-style contrastive loss. The encoder for each modality is initialized using a pretrained unimodal model, which is fine-tuned to achieve cross-modal alignment.

Inference: Image or DNA embeddings serve as queries, compared against a database of image, DNA, and text embeddings (keys). Classification is performed by retrieving the closest key via cosine similarity and returning its taxonomic label.

Task III: Results

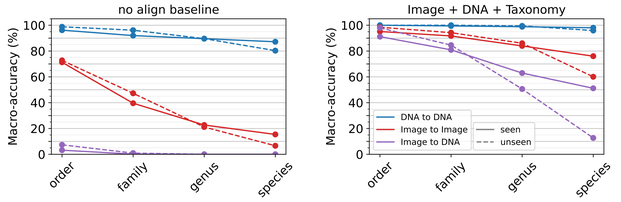


Figure 5: Multimodal retrieval accuracy (%) on seen and unseen species across different methods of retrieval (image-to-image, image-to-DNA, and DNA-to-DNA). Left: retrieval accuracy before alignment of encoders. Right: retrieval accuracy after aligning images, DNA, and taxonomic labels.

Explore more

- **Webpage:** https://biodiversitygenomics.net/5M-insects/
- **Github:** https://github.com/bioscan-ml/BIOSCAN-5M
- **Dataset Browser:** https://bioscan-browser.netlify.app

Dataset quick start

```
!pip install bioscan-dataset
from bioscan dataset import BIOSCAN5M
ds = BIOSCAN5M("~/Datasets/bioscan-5m", download=True)
```

Acknowledgements

BIOSCAN is supported in part by funding from the Government of Canada's New Frontiers in Research Fund.
FR: BIOSCAN repose sur des recherches financées par le fonds Nouvelles frontières en recherche du gouvernement du Canada.