BIOSCAN-5M
A Multimodal Dataset for Insect Biodiversity

Zahra Gharaee* and Scott C. Lowe* and ZeMing Gong* and Pablo Millan Arias*

Nicholas Pellegrino and Austin T. Wang and Joakim Bruslund Haurum

Iuliia Zarubiieva and Lila Kari

Dirk Steinke Δ and Graham W. Taylor Δ and Paul Fieguth Δ and Angel X. Chang Δ

* Joint first author. Δ Joint last / senior author.

Zahra Gharaee
Vision and Image Processing Lab (VIP),
Systems Design Engineering, University of Waterloo
Waterloo, Canada
zahra.gharaee@gmail.com

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Introduction to BIOSCAN-5M Dataset

A New Frontier in Insect Biodiversity Exploration

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Taxonomy

Phylum	Arthropoda
Class	Insecta
Order	Hymenoptera
Family	Formicidae
Subfamily	Dolichoderinae
Genus	*Tapinoma*
Species	*Tapinoma sessile*
BIN	BOLD:AAA3908

DNA Barcode

Thymine	T
Cytosine	C
Guanine	G
Cytosine	C
Thymine	T
Adenine	A
.	
.	
.	

Geographic Data

Country	United States
Province/State	California
Latitude	40.10132
Longitude	-122.05354

Size Data

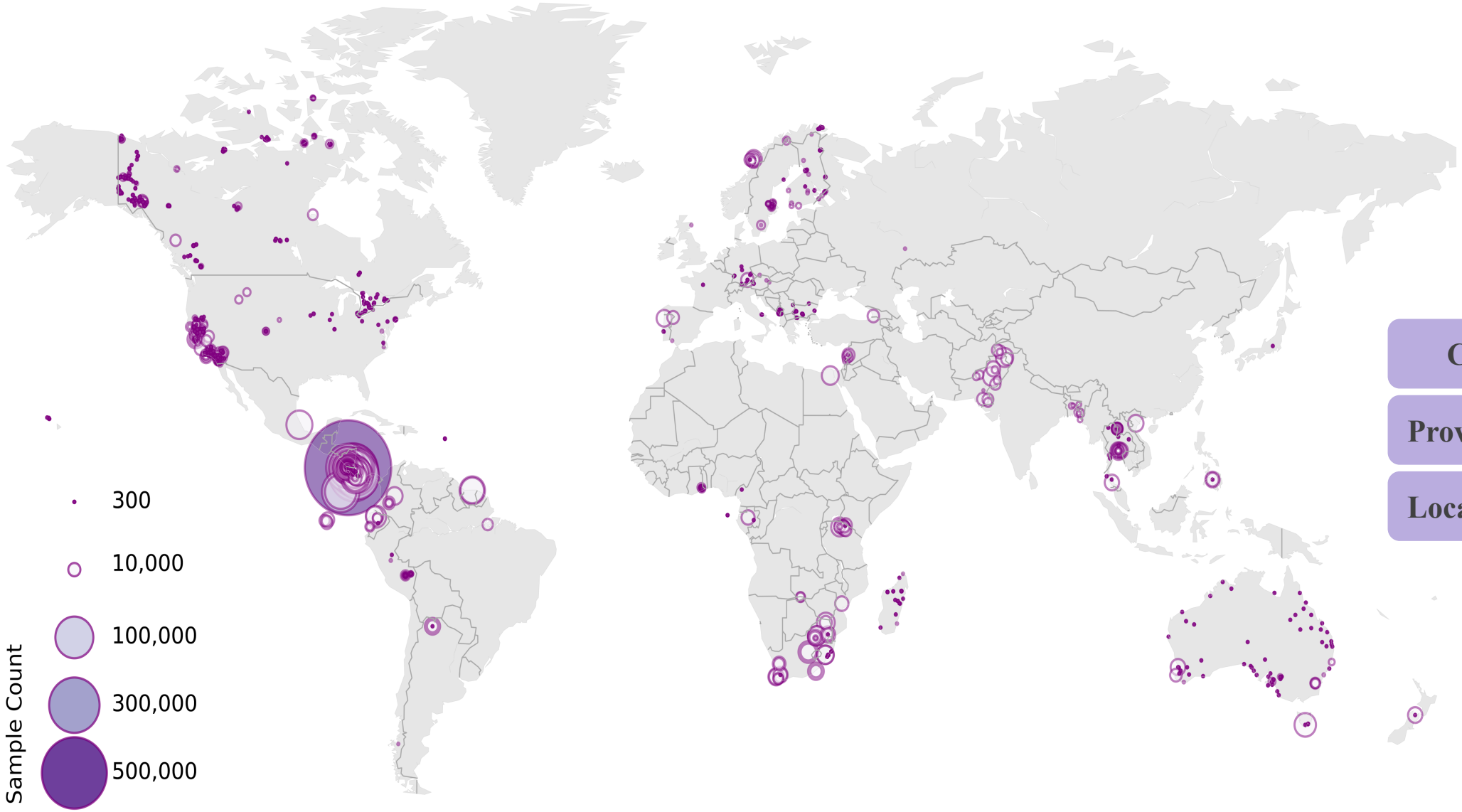
Meas.Value	64,331
Area Fraction	0.40
Scale Factor	2.08

Image Data

Introduction to BIOSCAN-5M Dataset

A New Frontier in Insect Biodiversity Exploration

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024



	Count
Country	47
Province/State	102
Location (Lat, Lon)	1,650

Introduction to BIOSCAN-5M Dataset

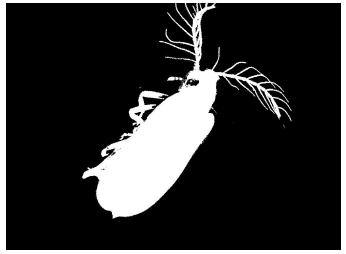
A New Frontier in Insect Biodiversity Exploration

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Meas.Value

$$f_p = \sum_{n=1}^{N} p_i$$

$$p_i = \begin{cases} 1 & \text{If pixel } i \text{ occupied} \\ 0 & \text{Otherwise} \end{cases}$$

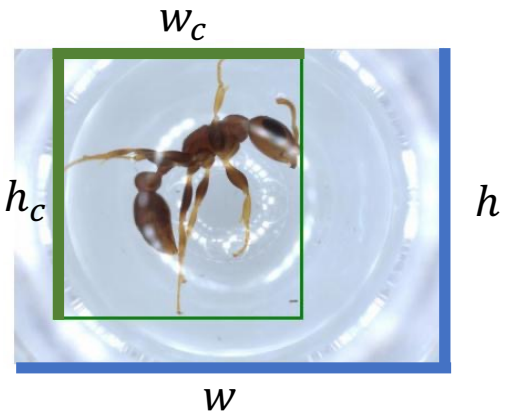



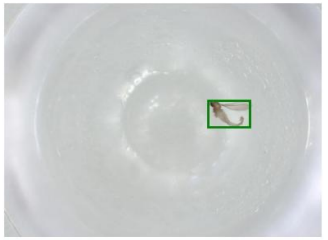
Area Fraction

$$f_a = \frac{w_c h_c}{w\ h}$$

Scale Factor

$$f_s = \frac{\min(w_c,\ h_c)}{256}$$




Introduction to BIOSCAN-5M Dataset

A New Frontier in Insect Biodiversity Exploration

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Statistical Insights into BIOSCAN-5M

Uncovering Patterns in Insect Biodiversity

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

	Labelled Records		Categories				Unique DNA	
			Count	Imbalance Ratio	Most Populated	Least Populated	Count	Avg SDI
Phylum	5,150 k	100%	1	1	5,150 k	5,150 k	2,486 k	19.78
Class	5,146 k	99.9%	10	719 k	5,038 k	7	2,482 k	8.56
Order	5,134 k	99.7%	55	3,675 k	3,675 k	1	2,474 k	7.05
Family	4,932 k	95.8%	934	938 k	938 k	1	2,321 k	5.42
Subfamily	1,472 k	28.6%	1,542	323 k	323 k	1	657 k	4.28
Genus	1,226 k	23.8%	7,605	200 k	200 k	1	531 k	2.63
Species	473 k	9.2%	22,622	7,694	7,694	1	202 k	1.46

Practical Insights into BIOSCAN-5M

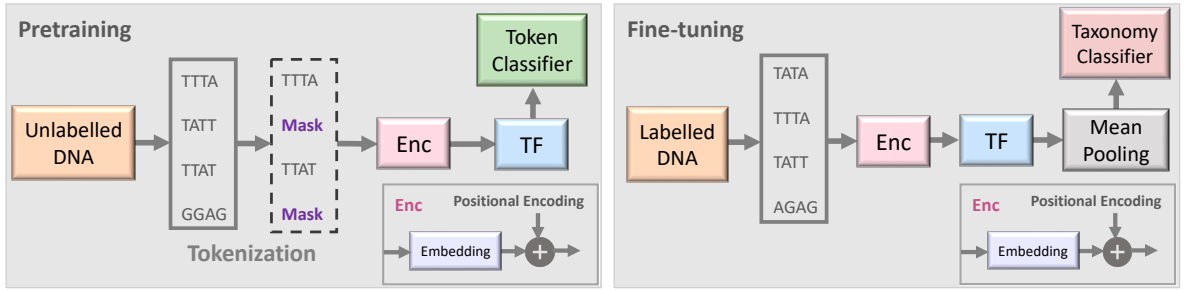
Revealing Biodiversity Patterns for Real-World Impact

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

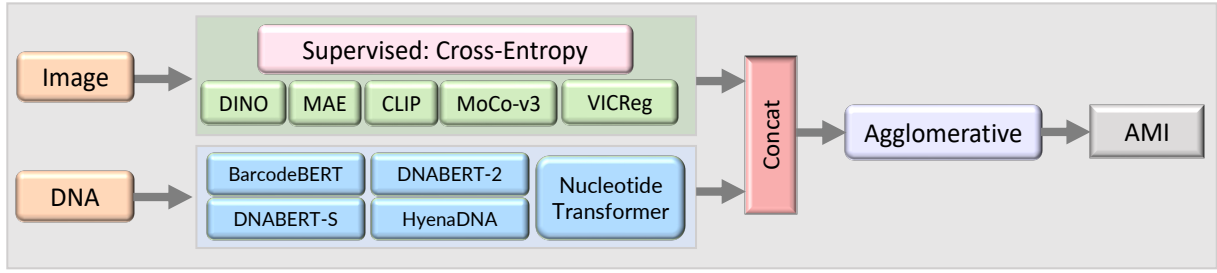
Data partitioning supporting both closed- and open world settings

Benchmark experiments

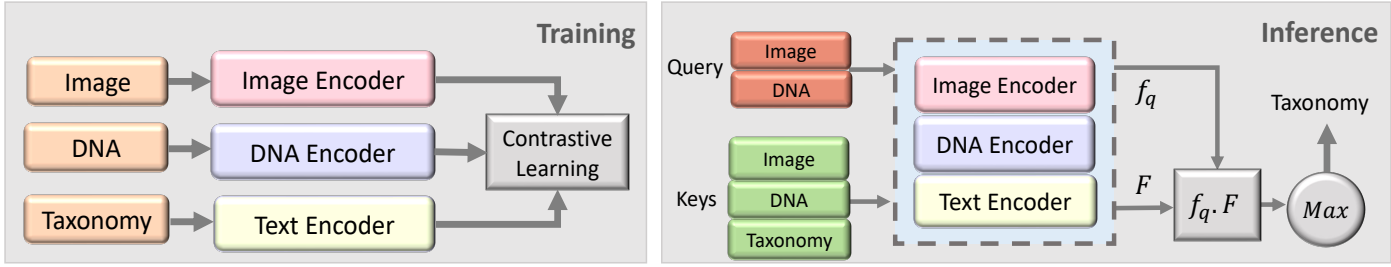
- DNA based taxonomic classification



- Zero-shot transfer learning



- Multimodal retrieval learning

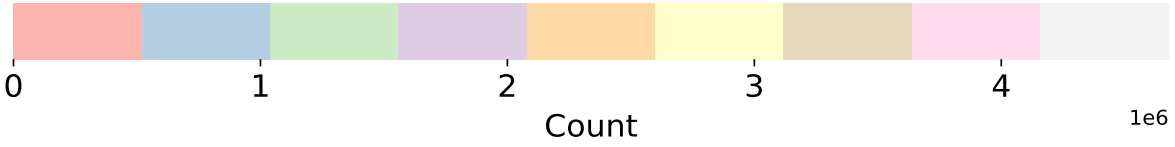


From Data to Discovery

Data Partitioning for Bridging Known and Novel Species

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Species	Split sets	Applications	Samples	Barcodes	Species
Unknown	pretrain	Self- and semi-supervised learning	4678 k	2284 k	0
Seen	train	Supervised learning, retrieval keys	289 k	118 k	12 k
	val	Model development; retrieval queries	15 k	7 k	3 k
	test	Final evaluation; retrieval queries	39 k	18 k	3 k
Unseen	key_unseen	Retrieval keys	36 k	12 k	914
	val_unseen	Model development; retrieval queries	9 k	2 k	903
	test_unseen	Final evaluation; retrieval queries	8 k	3 k	880
Heldout	other_heldout	Novelty detector training	77 k	41 k	10 k



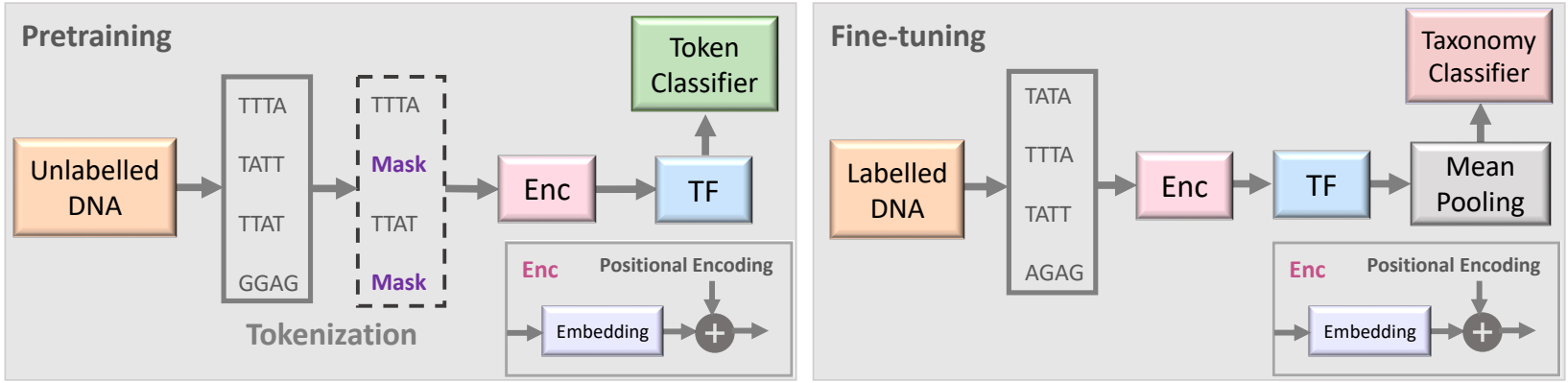
Real-World Applications of BIOSCAN-5M

DNA Based Taxonomic Classification

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Best Model

Second Best Model

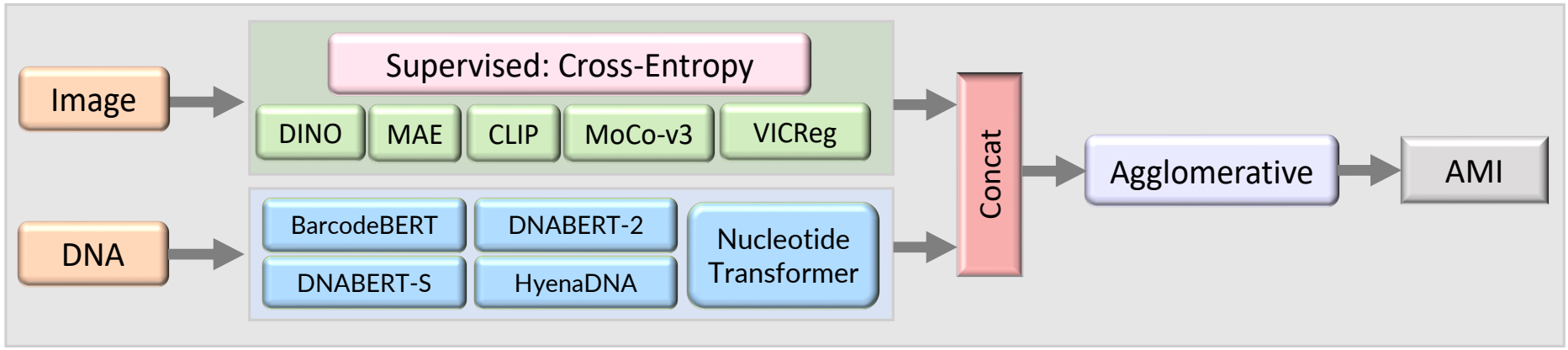


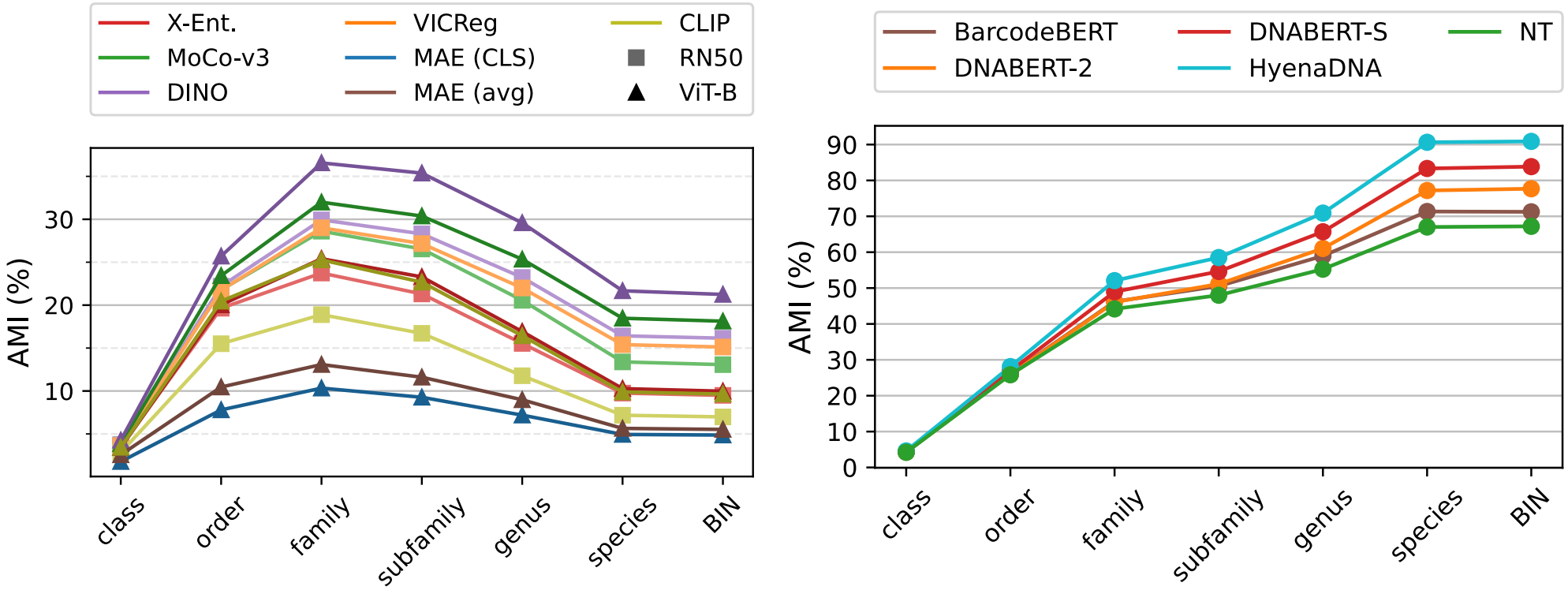
				Seen: Species		Unseen: Genus
Model	**Architecture**	**SSL-Pretraining**	**Token seen**	**Fine-tuned**	**Linear probe**	**1NN-Probe**
CNN Baseline	CNN	-	-	97.70	-	29.88
NT	Transformer	Multi-Species	300 B	98.99	52.41	21.67
DANBERT-2	Transformer	Multi-Species	512 B	99.23	67.81	17.99
DANBERT-S	Transformer	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

Real-World Applications of BIOSCAN-5M

Zero-shot Transfer Learning

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024





Real-World Applications of BIOSCAN-5M

Multimodal Retrieval Learning

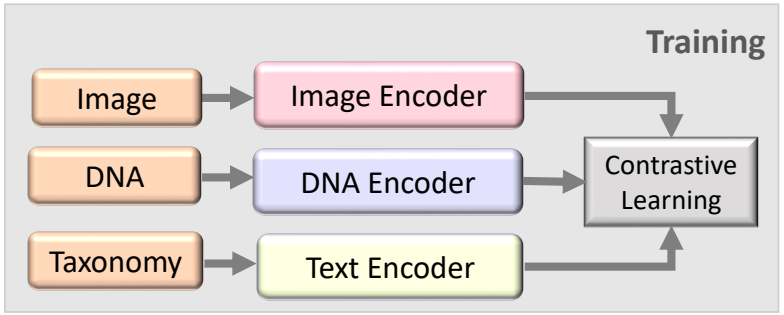
NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

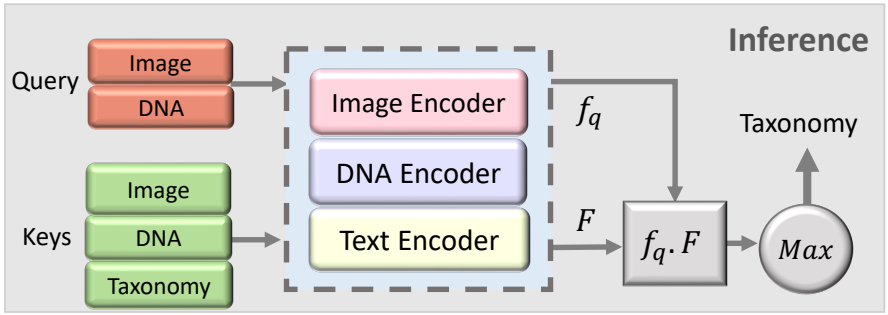
Taxonomy Prediction: **Species**

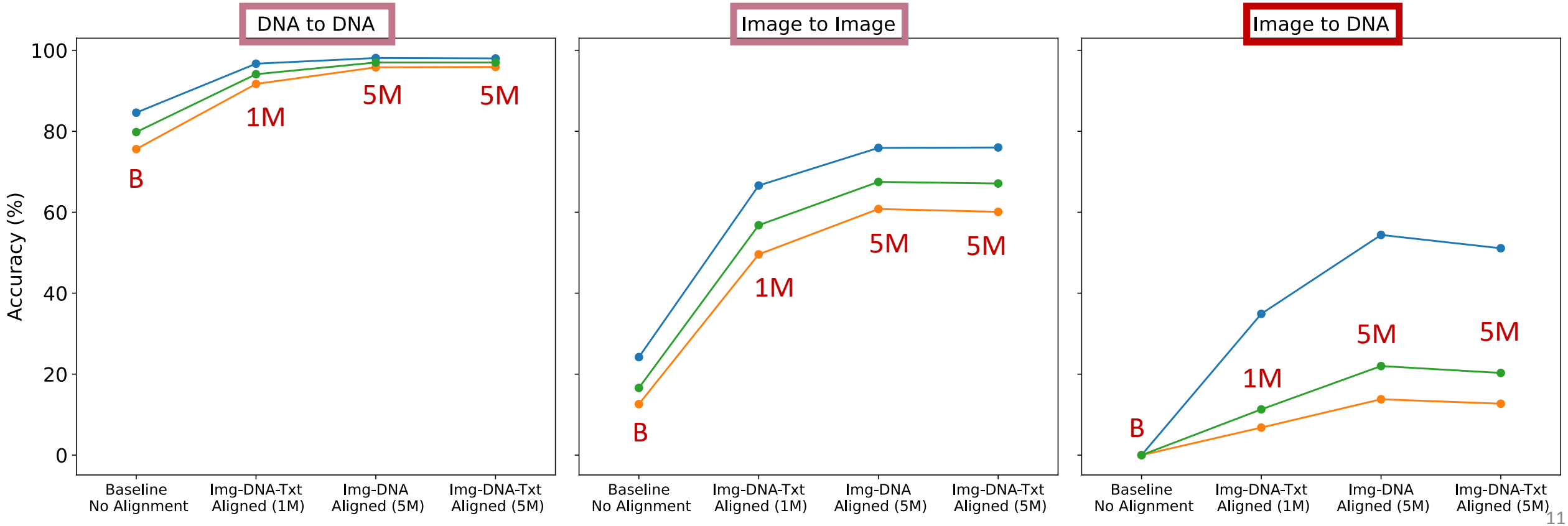
Seen

Unseen

Harmonic Mean of seen and unseen







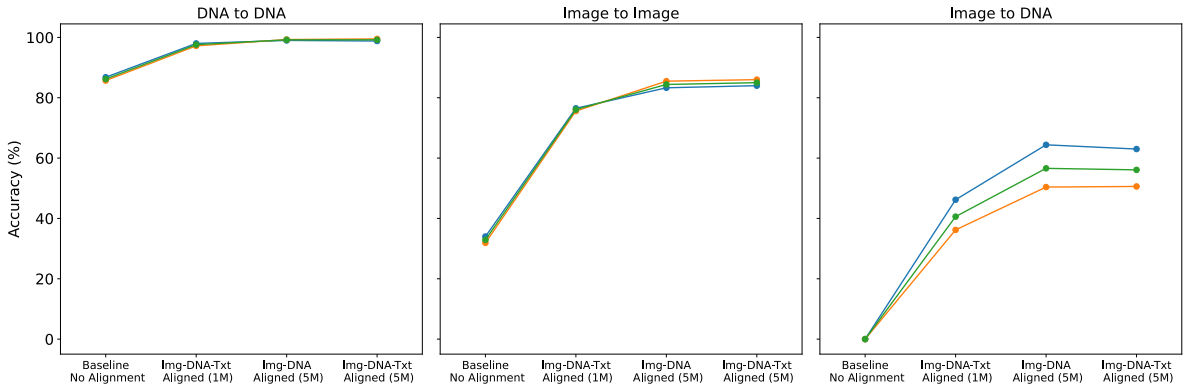
Real-World Applications of BIOSCAN-5M

Multimodal Retrieval Learning

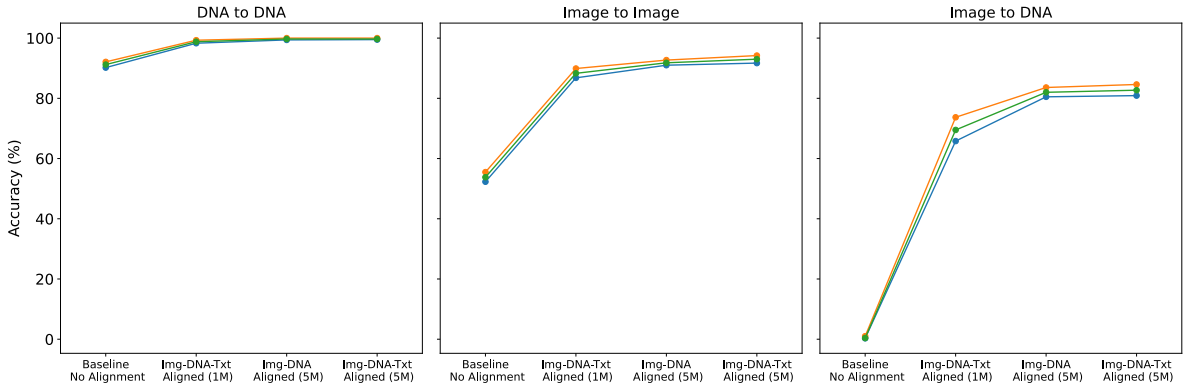
NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

Seen
Unseen
Harmonic Mean of seen and unseen

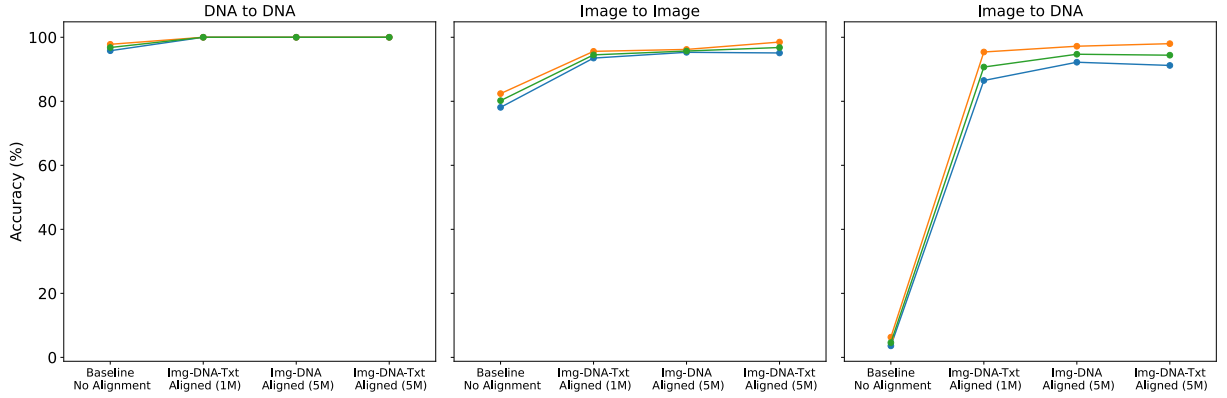
Genus



Family



Order



Key Takeaways from BIOSCAN-5M Research

Conclusion & Resources for Continued Exploration

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024

BIOSCAN-5M offers

- A comprehensive resource with **over 5 million** arthropod specimens.
- **Multimodal** data including images, DNA barcodes, and taxonomic text annotations.
- **Data partitions** for both **closed-** and **open-world** machine learning settings.
- **Benchmark** experiments for fine-grained taxonomic classification and zero-shot clustering.

Future plans

- Advancing **biodiversity monitoring** by developing models for **species categorization** and **novel species discovery**.
- Fostering **collaborative** projects and initiatives inspired by BIOSCAN-5M to map and preserve **global biodiversity through machine learning.**

Preprint	https://arxiv.org/abs/2406.12723
Github	https://github.com/bioscan-ml/BIOSCAN-5M
GoogleDrive	https://drive.google.com/drive/u/0/folders/1Jc57eKkeiYrnUBc9WlIp-ZS_L1bVlT-0
Zenodo	https://zenodo.org/records/11973457
Kaggle	https://www.kaggle.com/datasets/zahragharaee/bioscan-5m
HuggingFace	https://huggingface.co/datasets/Gharaee/BIOSCAN-5M
Contact	zahra.gharaee@gmail.com

BIOSCAN-5M
A Multimodal Dataset for Insect Biodiversity

https://biodiversitygenomics.net/projects/5m-insects/

Thank you for your engagement.

Thanks to my collaborators, the BIOSCAN team, and everyone who made this research possible.

Zahra Gharaee
Vision and Image Processing Lab (VIP),
Systems Design Engineering, University of Waterloo
Waterloo, Canada
zahra.gharaee@gmail.com

NEURAL INFORMATION PROCESSING SYSTEMS
10 - 15 December 2024