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]University of Guelph, [3]University of Waterloo,
[4]Simon Fraser University, [5]Vector Institute, [6]Alberta Machine Intelligence Institute (Amii),
[7]Aalborg University and Pioneer Centre for AI
https://biodiversitygenomics.net/5M-insects/

Abstract

As part of an ongoing worldwide effort to comprehend and monitor insect biodiversity, this paper presents the BIOSCAN-5M Insect dataset to the machine learning community and establish several benchmark tasks. BIOSCAN-5M is a comprehensive dataset containing multi-modal information for over 5 million insect specimens, and it significantly expands existing image-based biological datasets by including taxonomic labels, raw nucleotide barcode sequences, assigned barcode index numbers, geographical, and size information. We propose three benchmark experiments to demonstrate the impact of the multi-modal data types on the classification and clustering accuracy. First, we pretrain a masked language model on the DNA barcode sequences of the BIOSCAN-5M dataset, and demonstrate the impact of using this large reference library on species- and genus-level classification performance. Second, we propose a zero-shot transfer learning task applied to images and DNA barcodes to cluster feature embeddings obtained from self-supervised learning, to investigate whether meaningful clusters can be derived from these representation embeddings. Third, we benchmark multi-modality by performing contrastive learning on DNA barcodes, image data, and taxonomic information. This yields a general shared embedding space enabling taxonomic classification using multiple types of information and modalities. The code repository of the BIOSCAN-5M Insect dataset is available at https://github.com/bioscan-ml/BIOSCAN-5M.

1 Introduction

Biodiversity plays a multifaceted role in sustaining ecosystems and supporting human well-being. Primarily, it serves as a cornerstone for ecosystem stability and resilience, providing a natural defence against disturbances such as climate change and invasive species (Cardinale et al., 2012). Additionally, biodiversity serves as a vital resource for the economy, supplying essentials like food, medicine, and genetic material (Sala et al., 2000). Understanding biodiversity is paramount for sustainable resource management, ensuring the availability of these resources for future generations (Duraiappah et al., 2005). To understand and monitor biodiversity, Gharaee et al. (2023) introduced the BIOSCAN-1M Insect dataset, which pairs DNA with images, as a stepping stone to developing AI tools for automatic classification of organisms.

[*]Joint first author. [†]Joint senior/last author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 1: The BIOSCAN-5M Dataset provides taxonomic labels, a DNA barcode sequence, barcode index number, a high-resolution image along with its cropped and resized versions, as well as size and geographic information for each sample.

However, that work only investigated image classification down to the family level, focusing on the Diptera order, and did not fully utilize the multimodal nature of the dataset. In addition, BIOSCAN-1M was limited to specimen collected from just 3 countries and the *Insecta* `class`. Expanding upon BIOSCAN-1M, we introduce the BIOSCAN-5M dataset—a comprehensive repository of multi-modal information (see Figure 1) on over 5 million arthropod specimens (98% insects), with 1.2 million labelled to `genus` or `species` taxonomic ranks. Compared to its predecessor, the BIOSCAN-5M dataset offers a significantly larger volume of high-resolution microscope images and DNA barcodes along with critical annotations, including taxonomic ranks, size, and geographical information. Additionally, we performed data cleaning to resolve inconsistencies and provide more reliable labels.

The multimodal characteristics of BIOSCAN-5M are not only essential for biodiversity studies, but also facilitate further innovation in machine learning and AI. In this paper, we conduct experiments that leverage the multimodal aspects of BIOSCAN-5M, extending its application beyond the image-only modality used in Gharaee et al. (2023). Here, we train the masked language model (MLM) proposed in BarcodeBERT (Millan Arias et al., 2023) on the DNA barcodes of the BIOSCAN-5M dataset and demonstrate the impact of using this large reference library on species- and genus-level classification. We achieve an accuracy higher than that of state-of-the-art models pretrained on more general genomic datasets, especially in the 1NN-probing task of assigning samples from unseen species to seen genera. Next, we perform a zero-shot transfer learning task (Lowe et al., 2024a) through zero-shot clustering representation embeddings obtained from encoders trained with self-supervised paradigms. This approach demonstrates the effectiveness of pretrained embeddings in clustering data, even in the absence of ground-truth. Finally, as in CLIBD (Gong et al., 2024), we learn a shared embedding space across three modalities in the dataset—high-quality RGB images, textual taxonomic labels, and DNA barcodes—for fine-grained taxonomic classification.

2 Related work

2.1 Datasets for taxonomic classification

Biological datasets are essential for advancing our understanding of the natural world, with uses in genomics (Network et al., 2013), proteomics (Kim et al., 2014), ecology (Kattge et al., 2011), evolutionary biology (Flicek et al., 2014), medicine (Jensen et al., 2012), and agriculture (Lu & Young, 2020; Xu et al., 2023; Galloway et al., 2017; He et al., 2024). Table 1 compares biological datasets used for taxonomic classification. Many of these datasets feature fine-grained classes and exhibit a long-tailed class distribution, making the recognition task challenging for machine learning (ML) methods that do not account for these properties. While many datasets provide images, they do not include other attributes such as DNA barcode, or geographical locations. Most relevant to our work is BIOSCAN-1M Insect (Gharaee et al., 2023), which introduced a dataset of 1.1 M insect images paired with DNA barcodes and taxonomic labels.

DNA barcodes are short, highly descriptive DNA fragments that encode sufficient information for species-level identification. For example, a DNA barcode of an organism from Kingdom Animalia (Hebert et al., 2003; Braukmann et al., 2019) is a specific 648 bp sequence of the cytochrome c oxidase I (COI) gene from the mitochondrial genome, used to classify unknown individuals and discover new species (Moritz & Cicero, 2004). DNA barcodes have been successfully applied to taxonomic identification and classification, ecology, conservation, diet analysis, and food safety (Ruppert et al., 2019; Stoeck et al., 2018), offering faster and more accurate results than traditional meth-

Table 1: **Summary of fine-grained and long-tailed biological datasets.** The "Taxa" column indicates the taxonomic scope of each dataset. The "IR" column is the class imbalance ratio, computed as the ratio of the number of samples in the largest category to the smallest category.

Dataset	Reference	Year	Images	IR	Taxa	Rank	Categories	Taxon	BIN	DNA	Geography	Size
LeafSnap	Kumar et al. (2012)	2012	31 k	8	Plants	Species	184	✗	✗	✗	✗	✗
NA Birds	Van Horn et al. (2015)	2015	48 k	15	Birds	Species	400	✗	✗	✗	✗	✗
Urban Trees	Wegner et al. (2016)	2016	80 k	7	Trees	Species	18	✗	✗	✗	✗	✗
DeepWeeds	Olsen et al. (2019)	2019	17 k	9	Plants	Species	9	✗	✗	✗	✓	✗
IP102	Wu et al. (2019)	2019	75 k	14	Insects	Species	102	✓	✗	✗	✗	✗
Pest24	Wang et al. (2020)	2020	25 k	494	Insects	Species	24	✗	✗	✗	✗	✗
Pl@ntNet-300K	Garcin et al. (2021)	2021	306 k	3,604	Plants	Species	1,000	✗	✗	✗	✗	✗
iNaturalist (2021)	Van Horn et al. (2021)	2021	2,686 k	2	All	Species	10,000	✓	✗	✗	✗	✗
iNaturalist-Insect	Van Horn et al. (2021)	2021	663 k	2	Insects	Species	2,526	✓	✗	✗	✗	✗
Species196-L	He et al. (2024)	2023	19 k	351	Various	Mixed	196	✓	✗	✗	✗	✗
CWD30	Ilyas et al. (2023)	2023	219 k	61	Plants	Species	30	✓	✗	✗	✗	✗
BenthicNet	Lowe et al. (2024b)	2024	1,429 k	22,394	Aquatic	Mixed	791	✓	✗	✗	✓	✗
Insect-1M	Nguyen et al. (2024b)	2024	1,017 k	N/A	Arthropods	Species	34,212	✓	✗	✗	✗	✗
BIOSCAN-1M	Gharaee et al. (2023)	2023	1,128 k	12,491	Insects	BIN*	90,918	✓	✓	✓	✗	✗
BIOSCAN-5M	Ours	2024	5,150 k	35,458	Arthropods	BIN*	324,411	✓	✓	✓	✓	✓

* For datasets that include Barcode Index Numbers (BINs) annotations, we present BINs, which serve as a (sub)species proxy for organisms and offer a viable alternative to Linnean taxonomy.

ods (Pawlowski et al., 2018). Barcodes can also be grouped together based on sequence similarity into clusters called Operational Taxonomic Units (OTUs) (Sokal & Sneath, 1963; Blaxter et al., 2005), each assigned a Barcode Index Number (BIN) (Ratnasingham & Hebert, 2013). In general, biological datasets may also incorporate other data such as labels for multi-level taxonomic ranks, which can offer valuable insights into the evolutionary relationships between organisms. However, datasets with hierarchical taxonomic annotations (He et al., 2024; Ilyas et al., 2023; Liu et al., 2021; Wu et al., 2019; Gharaee et al., 2023) are relatively scarce.

2.2 Self-supervised learning

Self-supervised learning (SSL) has recently gained significant attention for its ability to leverage vast amounts of unlabelled data, producing versatile feature embeddings for various tasks (Balestriero et al., 2023). This has driven the development of large-scale language models (Brown et al., 2020) and computer vision systems trained on billions of images (Goyal et al., 2021). Advances in transformers pretrained with SSL at scale, known as foundation models (Ji et al., 2021; Zhou et al., 2023; Dalla-Torre et al., 2023; Zhou et al., 2024; Chia et al., 2022; Gu et al., 2021), have shown robust performance across diverse tasks.

Recent work has leveraged these advances for taxonomic classification using DNA. Since the introduction of the first DNA language model, DNABERT (Ji et al., 2021), which mainly focused on human data, multiple models with different architectures and tokenization strategies have emerged (Mock et al., 2022; Zhou et al., 2023, 2024; Millan Arias et al., 2023; Nguyen et al., 2024a) with some incorporating data from multiple species during pretraining and allowing for species classification (Zhou et al., 2023, 2024; Millan Arias et al., 2023). These models are pretrained to be task-agnostic, and are expected to perform well after fine-tuning in downstream tasks. Yet, their potential application for taxonomic identification of arbitrary DNA sequences or DNA barcodes has not been extensively explored. One relevant approach, BERTax (Mock et al., 2022), pretrained a BERT (Dosovitskiy et al., 2021b) model for hierarchical taxonomic classification on broader ranks such as kingdom, phylum, and genus. For DNA barcodes specifically, BarcodeBERT (Millan Arias et al., 2023) was developed for species-level classification of insects, with assignment to genus for unknown species.

Although embeddings from SSL-trained feature extractors exhibit strong performance on downstream tasks post fine-tuning, their utility without fine-tuning remains underexplored. Previous studies (Vaze et al., 2022; Zhou & Zhang, 2022) suggest that SSL feature encoders produce embeddings conducive to clustering, albeit typically after fine-tuning. A recent study (Lowe et al., 2024a) has delved into whether SSL-trained feature encoders *without* fine-tuning can serve as the foundation for clustering, yielding informative clusters of embeddings on real-world datasets unseen during encoder training.

2.3 Multimodal Learning

There has been a growing interest in exploring multiple data modalities for biological tasks (Ikezogwo et al., 2024; Lu et al., 2023; Zhang et al., 2023). Badirli et al. (2021) introduced a Bayesian zero-shot

learning approach, leveraging DNA data to model priors for species classification based on images. Those authors also employed Bayesian techniques (Badirli et al., 2023), combining image and DNA embeddings in a unified space to predict the genus of unseen species.

Recent advances in machine learning allowed scalable integration of information across modalities. For example, CLIP (Radford et al., 2021) used contrastive learning to encode text captions and images into a unified space for zero-shot classification. BioCLIP (Stevens et al., 2024) used a similar idea to align images of organisms with their common names and taxonomic descriptions across a dataset of 10 M specimens encompassing plants, animals, and fungi. CLIBD (Gong et al., 2024) used a contrastive loss to align the three modalities of RGB images, textual taxonomic labels, and DNA barcodes. By aligning these modalities, CLIBD can use either images or DNA barcodes for taxonomic classification and learn from incomplete taxonomic labels, making it more flexible than BioCLIP (Stevens et al., 2024), which requires full taxonomic annotations for each specimen.

3 Dataset

The BIOSCAN-5M dataset is derived from Steinke et al. (2024) and comprises 5,150,850 `arthropod` specimens, with insects accounting for about 98% of the total. The diverse features of this dataset are described in this section. BIOSCAN-5M is a superset of the BIOSCAN-1M Insect dataset (Gharaee et al., 2023), providing more samples and additional metadata such as geographical location.

Images. The BIOSCAN-5M dataset provides specimen images at 1024×768 pixels, captured using a Keyence VHX-7000 microscope. Figure 2 showcases the diversity in organism morphology across the dataset. The images are accessed via the `processid` field of the metadata as `{processid}.jpg`. Following BIOSCAN-1M Insect (Gharaee et al., 2023), the images are cropped and resized to 341×256 pixels to facilitate model training. We fine-tuned DETR (End-to-End Object Detection with Transformers) for image cropping. For BIOSCAN-1M Insect, the cropping model was trained using 2 k insect images. Building on the BIOSCAN-1M Insect cropping tool checkpoint, we fine-tuned the model for BIOSCAN-5M using the same 2 k images and an additional 837 images that were not well-cropped previously. This fine-tuning process followed the same training setup, including batch size, learning rate, and other hyper parameter settings (see Appendix Q.1 for details). The bounding box of the cropped region is provided as part of the dataset release.

Figure 2: Samples of original full-size images of distinct organisms in the BIOSCAN-5M dataset.

Genetic-based indexing. The genetic information of the BIOSCAN-5M dataset described in §2 is represented as the raw nucleotide barcode sequence, under the `dna_barcode` field, and the Barcode Index Number under `dna_bin` field. Independently, the field `processid` is a unique number assigned by BOLD (International Barcode of Life Consortium, 2024) to each record, and `sampleid` is an identifier given by the collector.

Biological taxonomic classification. Linnaean taxonomy is a hierarchical classification system instigated by Linnaeus (1758) for organizing living organisms which has been developed over several hundred years. It categorizes `species` based on shared characteristics and establishes a standardized naming convention. The hierarchy includes several taxonomic ranks, such as `domain`, `kingdom`, `phylum`, `class`, `order`, `family`, `genus`, and `species`, allowing for a structured approach to studying biodiversity and understanding the relationships between different organisms.

The dataset samples undergo taxonomic classification using a hybrid approach involving an AI-assisted tool proposed by Gharaee et al. (2023) and human taxonomic experts. After DNA barcoding and sequence alignment, the taxonomic levels derived from both the AI tool and DNA sequencing are compared. Any discrepancies are then reviewed by human experts. Importantly, assignments to deeper taxonomic levels, such as `family` or lower, rely entirely on human expertise. Labels at seven taxonomic ranks are used to represent individual specimens, denoted by fields `phylum`, `class`, `order`, `family`, `subfamily`, `genus`, and `species`.

Table 2: Summary statistics of dataset records by taxonomic rank.

	BIOSCAN-5M (Ours)				**BIOSCAN-1M** (Gharaee et al., 2023)		
Attributes	IR	Categories	Labelled	Labelled (%)	Categories	Labelled	Labelled (%)
`phylum`	1	1	5,150,850	100.0	1	1,128,313	100.0
`class`	719,831	10	5,146,837	99.9	1	1,128,313	100.0
`order`	3,675,317	55	5,134,987	99.7	16	1,128,313	100.0
`family`	938,928	934	4,932,774	95.8	491	1,112,968	98.6
`subfamily`	323,146	1,542	1,472,548	28.6	760	265,492	23.5
`genus`	200,268	7,605	1,226,765	23.8	3,441	254,096	22.5
`species`	7,694	22,622	473,094	9.2	8,355	84,397	7.5
`dna_bin`	35,458	324,411	5,137,441	99.7	91,918	1,128,313	100.0
`dna_barcode`	3,743	2,486,492	5,150,850	100.0	552,629	1,128,313	100.0

In the source data, we found identical DNA nucleotide sequences labelled differently at some taxonomic levels, which was likely due to human error (e.g. typos) or disagreements in the taxonomic labelling. To address this, we checked and cleaned the taxonomic labels to address typos and ensure consistency across DNA barcodes (see Appendix Q.4 for details). We note that some of the noisy species labels are placeholder labels that do not correspond to well-established scientific taxonomic species names. In our data, the placeholder `species` labels are identified by `species` labels that begin with a lowercase letter, contain a period, contain numerals, or contain "malaise".

Statistics for BIOSCAN-5M are given in Table 2 for the seven taxonomic ranks along with the BIN and DNA nucleotide barcode sequences. For each group, we report the number of categories, and the count and fraction labelled. We compute the class imbalance ratio (IR) as the ratio of the number of samples in the largest category to the smallest category, reflecting the class distribution within each group. For more detailed statistical analysis, see Appendix L.

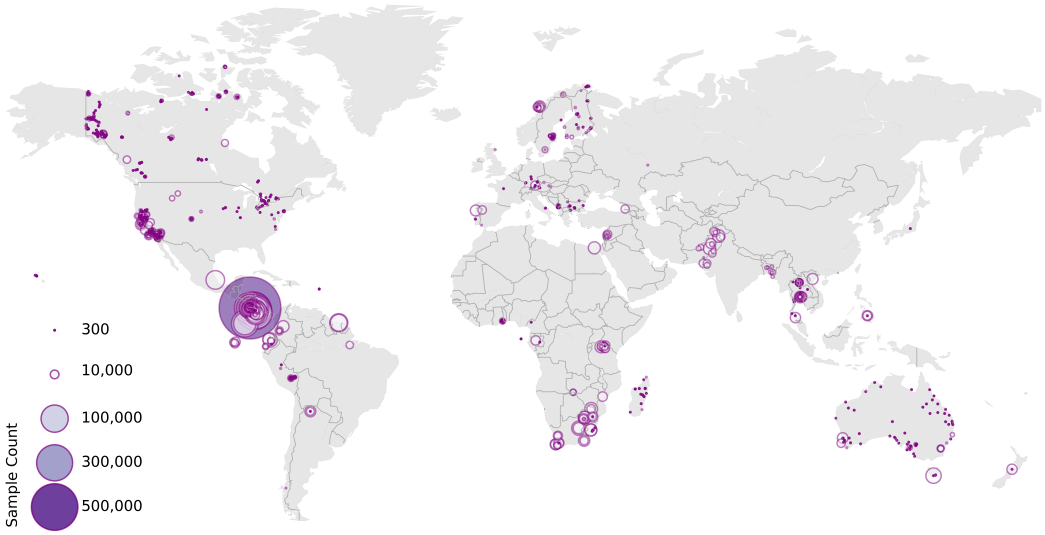


Figure 3: Geographical locations obtained from latitude and longitude coordinates of the regions where the samples of the BIOSCAN-5M dataset were collected.

Geographic location. The BIOSCAN-5M dataset includes geographic location information, detailing the country and province or state where each specimen is collected, along with the latitude and longitude coordinates of each collection site. This information is detailed in the fields `country`, `province_state`, `coord-lat` and `coord-lon`. The distribution of specimen collection sites are shown on a world map in Figure 10.

Challenges. The BIOSCAN-5M dataset faces two key challenges: First, there exists a sampling bias as a result of the locations where and the methods through which specimens are collected. Second, the number of labelled records sharply declines at deeper taxonomic levels, especially beyond the family rank, which makes fine-grained classification tasks more challenging.

4 Benchmark experiments and results

In real-world insect biodiversity monitoring, it is common to encounter both species which are already known to science, and samples whose species is novel. Thus, to excel in biodiversity monitoring, a model must correctly categorize instances of known species, and identify novel species outside the existing taxonomy, grouping together samples of the same new species. In our experiments, we explore three methods which offer utility in these regards, evaluated in two settings: closed-world and open-world. In the closed-world setting, the task is to accurately identify species from a predefined set of existing labels. In the open-world setting, the task is to group together samples of novel species.

4.1 Data partitioning

Species sets. We first partition records based on their species label into one of four categories, with all samples bearing the same species label being placed in the same species set. *Seen*: all samples whose species label is an established scientific name of a species. *Unseen*: labelled with an established scientific name for the genus, and a uniquely identifying placeholder name for the species. *Heldout*: labelled with a placeholder genus and species name. *Unknown*: samples without a species label (note: these may truly belong in any of the other three categories).

Table 3: Statistics and purpose of our data partitions.

Species set	Split	Purpose	# Samples	# Barcodes	# Species
unknown	`pretrain`	self- and semi-sup. training	4,677,756	2,284,232	—
seen	`train`	supervision; retrieval keys	289,203	118,051	11,846
	`val`	model dev; retrieval queries	14,757	6,588	3,378
	`test`	final eval; retrieval queries	39,373	18,362	3,483
unseen	`key_unseen`	retrieval keys	36,465	12,166	914
	`val_unseen`	model dev; retrieval queries	8,819	2,442	903
	`test_unseen`	final eval; retrieval queries	7,887	3,401	880
heldout	`other_heldout`	novelty detector training	76,590	41,250	9,862

Splits. Using the above species sets, we establish partitions for our experiments (Table 3). The *unknown* samples are all placed into a `pretrain` split for use in self-supervised pretraining and/or semi-supervised learning. As some DNA barcodes are common to multiple samples, for *seen* and *unseen* records we split the records by placing all samples with the same barcode in the same partition, to ensure there is no repetition of barcodes across splits. For the closed-world setting, we use the *seen* records to establish `train`, `val`, `test` splits. To ensure that the `test` set is not too imbalanced in species distribution, we place samples in the `test` set with a flattened distribution. We sample records from species with at least two unique barcodes and eight samples, and the number of samples placed in the `test` set scales linearly with the total number of samples for the species, until reaching a cap of 25 samples. We sample 5% of the remaining *seen* data to form the `val` partition, but in this case match the imbalance of the overall dataset. The remaining samples then form the `train` split, with a final split distribution of 84.2 : 4.3 : 11.5. Following standard practice, the `val` set is for model evaluation during development and hyperparameter tuning, and the `test` set is for final evaluation. In the retrieval setting, the `train` split should additionally be used as a database of *keys* to retrieve over, and the `val` and `test` split as queries. For additional details on the partitioning method and statistical comparisons between the partitions, please see Appendix R.

For the open-world scenario, we use a similar procedure to establish `val_unseen` and `test_unseen` over the *unseen* records. After creating `test_unseen` with the same methodology as `test`, we sample 20% of remaining *unseen* species records to create `val_unseen`. The remaining *unseen* species samples form the `keys_unseen` set. In the retrieval setting, `keys_unseen` is used to form the database of *keys* to retrieve, and the `val_unseen` and `test_unseen` splits act as queries. The *heldout* species samples form a final `other_heldout` partition. As these species are in neither *seen* nor *unseen*, this split can be used to train a novelty detector without using any *unseen* species.

4.2 DNA-based taxonomic classification

In this section, we demonstrate the utility of the BIOSCAN-5M dataset for DNA-based taxonomic classification. Due to their standardized length, DNA barcodes are ideal candidates as input to CNN- and transformer-based architectures for supervised taxonomic classification. However, as noted by Millan Arias et al. (2023), a limitation of this approach is the uncertainty in species-level labels for a substantial portion of the data. This uncertainty, partly due to the lack of consensus among researchers and the continuous discovery of new species, may render supervised learning suboptimal for this task. We address this issue by adopting a semi-supervised learning approach. Specifically, we train a model using self-supervision on unlabelled sequences from the `pretrain` split and the `other_heldout` split, followed by fine-tuning on sequences from the `train` split, which includes high-quality labels. The same pretrained model can be used to produce embeddings for sequences from unseen taxa to address tasks in the open-world setting. Consequently, we use these embeddings to perform non-parametric taxonomic classification at a higher (less specific) level in the taxonomic hierarchy for evaluation.

Experimental setup. Although there has been a growing number of SSL DNA language models proposed in the recent literature, the results obtained by the recently proposed BarcodeBERT (Millan Arias et al., 2023) model empirically demonstrate that training on a dataset of DNA barcodes can outperform more sophisticated training schemes that use a diverse set of non-barcode DNA sequences, such as DNABERT (Ji et al., 2021) and DNABERT-2 (Zhou et al., 2023). In this study, we selected BarcodeBERT as our reference model upon which to investigate the impact of pretraining on the larger and more diverse DNA barcode dataset BIOSCAN-5M. See Appendix A for pretraining details.

We compare our pretrained model against four pretrained transformer models: BarcodeBERT (Millan Arias et al., 2023), DNABERT-2 (Zhou et al., 2023), DNABERT-S (Zhou et al., 2024), and the nucleotide transformer (NT) (Dalla-Torre et al., 2023); one state space model, HyenaDNA (Nguyen et al., 2024a); and a CNN baseline following the architecture introduced by Badirli et al. (2021).

As an additional assessment of the impact of BIOSCAN-5M DNA data during pretraining, we use the different pretrained models as feature extractors and evaluate the quality of the embeddings produced by the models on two different SSL evaluation strategies (Balestriero et al., 2023). We first implement genus-level 1-NN probing on sequences from unseen species, providing insights into the models' abilities to generalize to new taxonomic groups. Finally, we perform species-level classification using a linear classifier trained on embeddings from the pretrained models. Note that for both probing tasks, all the embeddings produced by a single sequence are averaged across the token dimension to generate a token embedding for the barcode.

Results. We leverage the different partitions of the data and make a distinction between the experiments in the closed-world and open-world settings. In the closed-world setting, the task is species-level identification of samples from species that have been seen during training (Fine-tuned accuracy, Linear probing accuracy). For reference, BLAST (Altschul et al., 1990), an algorithmic sequence alignment tool, achieves an accuracy of 99.78% in the task (not included in Table 4 as it is not a machine learning model). In fine-tuning, our pretrained model with a 8-4-4 architecture achieves the highest accuracy with 99.28%, while DNABERT-2 achieves 99.23%, showing competitive performance. Overall, all models demonstrate strong performance in this task, showcasing the effectiveness of DNA barcodes in species-level identification. For linear probing accuracy, DNABERT-S outperforms others with 95.50%, followed by our model (8-4-4) with 94.47%. BarcodeBERT (k=4) and DNABERT-S also show strong performance with 91.93% and 91.59% respectively (see Table 4).

In the open-world setting, the task is to assign samples from unseen species to seen categories of a coarser taxonomic ranking (1NN-genus probing). In this task, BLAST achieves an accuracy of 58.74%

Table 4: **Performance of DNA-based sequence models** in closed- and open-world settings. For the closed-world, we show the species-level accuracy (%) for predicting seen species (`test`), for open-world the genus-level accuracy (%) for `test_unseen` species while using seen species to fit the model. Bold indicates highest accuracy, underlined denotes second highest.

				Seen: Species		Unseen: Genus
Model	Architecture	SSL-Pretraining	Tokens seen	Fine-tuned	Linear probe	1NN-Probe
CNN baseline	CNN	–	–	97.70	–	<u>29.88</u>
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 (8-4-4)	Transformer	DNA barcodes	7 B	**99.28**	<u>94.47</u>	**47.03**

(not in the table), and our model (8-4-4) performs notably well with an 47.03% accuracy, which is significantly higher than the other transformer models. The CNN baseline and HyenaDNA show lower accuracies of 29.88% and 19.26%, respectively. The use of DNA barcodes for pretraining in our models and BarcodeBERT demonstrates effectiveness in both seen and unseen species classification tasks. One limitation of the comparison is the difference in the dimension of the output space of the different models (128 for HyenaDNA, *vs*. 512 for NT and 768 for the BERT-based models). The selection of our model (8-4-4) as the best-performing configuration was done after performing a hyperparameter search to determine the optimal value of k for tokenization, as well as the optimal number of heads and layers in the transformer model. To do that, after pretraining, we fine-tuned the model for species-level identification and performed linear- and 1NN- probing on the `validation` split (see Table 6). We finally note that our pretrained model outperforms BarcodeBERT, the other model trained exclusively trained on DNA barcodes, across all tasks.

4.3 Zero-shot transfer-learning

Recently, Lowe et al. (2024a) proposed the task of *zero-shot clustering*, investigating how well unseen datasets can be clustered using embeddings from pretrained feature extractors. Lowe et al. (2024a) found that BIOSCAN-1M images were best clustered taxonomically at the family rank while retaining high clustering performance at species and BIN labels. We replicate this analysis using BIOSCAN-5M and extend the modality space to include both image and DNA barcodes.

Experimental setup. We follow the experimental setup of Lowe et al. (2024a). (1) Take pretrained encoders; (2) Extract feature vectors from the stimuli by passing them through an encoder; (3) Reduce dimensions to 50 using UMAP (McInnes et al., 2018); (4) Cluster the reduced embeddings with Agglomerative Clustering (L2, Ward's method) (Everitt et al., 2011); (5) Evaluate against the ground-truth annotations with Adjusted Mutual Information (AMI) score (Vinh et al., 2010).

For the image encoders, we consider ResNet-50 (He et al., 2016a) and ViT-B (Dosovitskiy et al., 2021a) models, each pretrained on ImageNet-1K (Russakovsky et al., 2015) using either cross-entropy supervision (X-ent.), or SSL methods (MAE: He et al., 2022; VICReg: Bardes et al., 2022; DINO-v1: Caron et al., 2021; MoCo-v3: Chen et al., 2021). We also considered the CLIP (Radford et al., 2021) encoder, which was pretrained on an unspecified, large dataset of captioned images. To cluster the DNA barcodes, we used recent pretrained models (see §4.2 and Appendix A.2), which feature a variety of model architectures, pretraining datasets, and training methodologies: BarcodeBERT (Millan Arias et al., 2023), DNABERT-2 (Zhou et al., 2023), DNABERT-S (Zhou et al., 2024), the nucleotide transformer (NT) (Dalla-Torre et al., 2023), and HyenaDNA (Nguyen et al., 2024a).

We only cluster samples from the `test` and `test_unseen` splits. None of the image or DNA pretraining datasets overlap with BIOSCAN-5M, so all samples are "unseen". However, we note that there is a greater domain shift from the image pretraining datasets than the DNA pretraining datasets.

Results. Similar to Lowe et al. (2024a), we find (Figure 4) the clustered images best agreed with the taxonomic labels at the family rank, and the best-performing image encoder was DINO, followed by other SSL methods VICReg and MoCo-v3, with (larger) ViT-B models outperforming ResNet-50 models. The performance gradually declined when considering more fine-grained taxa. This

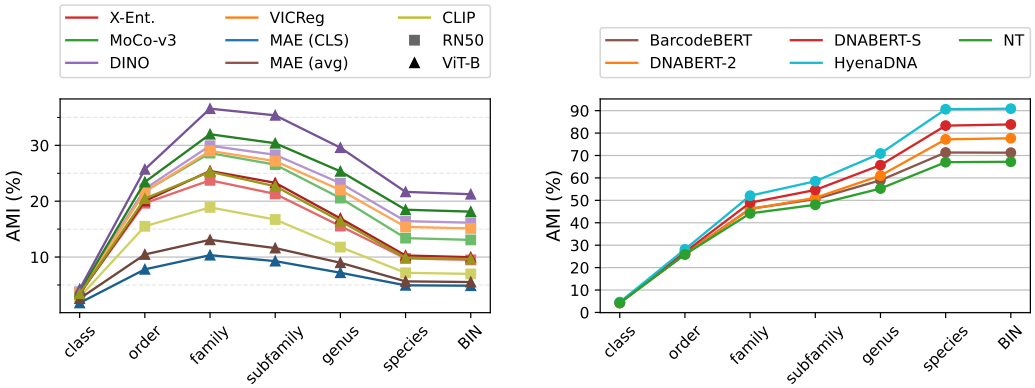


Figure 4: **Zero-shot clustering AMI (%) performance** across taxonomic ranks. Left: Image encoders. Right: DNA encoders.

difference from the findings of Lowe et al. (2024a) may be due to the flatter data balance compared with the BIOSCAN-1M splits. We found the performance of the DNA encoders exceeded that of the image encoders across all taxonomic levels, increasing monotonically as granularity becomes finer. HyenaDNA provided the best performance, with in excess of 90% agreement between its clusterings and the GT species and DNA BIN annotations. These results suggest that DNA barcodes are highly informative about species identity (which is unsurprising as it is the reason this barcode is used), and unseen samples can be readily grouped together using off-the-shelf DNA models.

We also considered the zero-shot clustering of the concatenated image and DNA representations, detailed in Appendix B.3. Due to the high performance of the DNA features, adding image features to the embeddings decreased the performance compared to using DNA embeddings alone. For additional details and analysis, see Appendix B.

4.4 Multimodal retrieval learning

Lastly, we demonstrate the importance of a multimodal dataset through alignment of image, DNA, and taxonomic label embeddings using CLIBD (Gong et al., 2024) to improve taxonomic classification. By learning a shared embedding space across modalities, we can query between modalities and leverage the information across them to achieve better performance in downstream tasks. We are able to incorporate a diversity of samples into training toward taxonomic classification, even with incomplete taxonomic labels.

Experimental setup. We follow the model architecture and experimental setup of CLIBD (Gong et al., 2024). We start with pretrained encoders for each modality and perform full-tuning with NT-Xent loss (Sohn, 2016). Our image encoder is a ViT-B (Dosovitskiy et al., 2021a) pretrained on ImageNet-21k and fine-tuned on ImageNet-1k (Deng et al., 2009). For DNA barcodes, we use BarcodeBERT (Millan Arias et al., 2023) with 5-mer tokenization, pretrained on about 893 k DNA barcodes from the Barcode of Life Data system (BOLD) (International Barcode of Life Consortium, 2024), and for text, we use BERT-small (Turc et al., 2019). We train on the `pretrain` and `train` splits using the Adam (Kingma & Ba, 2014) optimizer for 20 epochs until convergence with a learning rate of 1e-6 and batch size of 2000. Training took roughly 29 hours on four 80GB A100 GPUs. To evaluate the performance of our models, we report micro (see Appendix C) and macro top-1 accuracy for taxonomic classification at different levels. To determine the taxonomic labels for a new query, we encode the sample image or DNA and find the closest matching embedding in a set of labelled samples (keys). For efficient lookup, we use the FAISS (Johnson et al., 2019) library with exact search (`IndexFlatIP`).

We compared the performance for the initial pretrained (unimodal) encoders to our models fine-tuned on either the full `pretrain` and `train` partitions from BIOSCAN-5M, or on a random 1 million sample subset of these partitions. The 1M image subset contained 20% of the images, 27% of the barcodes, and 47% of the BINs of the 5 M image training dataset. We evaluated these using image-to-image, DNA-to-DNA, and image-to-DNA embeddings as queries and keys.

Table 5: Top-1 macro accuracy (%) on the test set for using different amount of pre-training data (1 million vs 5 million records from BIOSCAN-5M) and different combinations of aligned embeddings (image, DNA, text) during contrastive training. We show results for using image-to-image, DNA-to-DNA, and image-to-DNA query and key combinations. As a baseline, we show the results prior to contrastive learning (no alignment). We report the accuracy for seen and unseen species, and the harmonic mean (H.M.) between these (bold: highest acc.).

		Aligned embeddings			DNA-to-DNA			Image-to-Image			Image-to-DNA		
Taxon	# Records	Img	DNA	Txt	Seen	Unseen	H.M.	Seen	Unseen	H.M.	Seen	Unseen	H.M.
Order	—	✗	✗	✗	95.8	97.8	96.8	78.1	82.4	80.2	3.6	6.3	4.6
	1M	✓	✓	✓	**100.0**	**100.0**	**100.0**	93.5	95.6	94.5	86.5	95.4	90.7
	5M	✓	✓	✗	**100.0**	**100.0**	**100.0**	**95.3**	96.2	95.7	**92.2**	97.2	**94.7**
	5M	✓	✓	✓	**100.0**	**100.0**	**100.0**	95.1	**98.5**	**96.8**	91.2	**98.0**	94.4
Family	—	✗	✗	✗	90.2	92.1	91.2	52.3	55.5	53.8	0.3	1.0	0.4
	1M	✓	✓	✓	98.3	99.3	98.8	86.8	89.9	88.3	65.8	73.7	69.5
	5M	✓	✓	✗	99.4	**100.0**	**99.7**	91.0	92.7	91.8	80.5	83.6	82.0
	5M	✓	✓	✓	**99.5**	**100.0**	**99.7**	**91.7**	**94.2**	**93.0**	**80.9**	**84.6**	**82.7**
Genus	—	✗	✗	✗	86.8	85.7	86.2	34.0	31.9	32.9	0.0	0.0	0.0
	1M	✓	✓	✓	98.0	97.2	97.6	76.5	75.6	76.1	46.2	36.2	40.6
	5M	✓	✓	✗	**99.0**	99.3	**99.2**	83.3	85.5	84.4	**64.4**	50.4	**56.6**
	5M	✓	✓	✓	98.8	**99.5**	**99.2**	**84.0**	**86.0**	**85.0**	63.0	**50.6**	56.1
Species	—	✗	✗	✗	84.6	75.6	79.8	24.2	12.6	16.6	0.0	0.0	0.0
	1M	✓	✓	✓	96.7	91.7	94.1	66.6	49.6	56.8	34.9	6.8	11.3
	5M	✓	✓	✗	**98.1**	95.8	**97.0**	75.9	**60.8**	**67.5**	**54.4**	**13.8**	**22.0**
	5M	✓	✓	✓	98.0	**95.9**	**97.0**	**76.0**	60.1	67.1	51.1	12.7	20.3

Results. We compare CLIBD trained on the full BIOSCAN-5M training set against models trained on a randomly selected subset of 1 million records and the initial pretrained encoders before multimodal contrastive learning. Our results, shown in Table 5, demonstrate that our full model improves classification accuracy for same-modality queries and enables cross-modality queries. By aligning to DNA, our image embeddings are able to capture finer details. We likewise see improvements in alignment among DNA embeddings. Additionally, we observe that increasing the training dataset size from 1 million to 5 million records leads to better models with more accurate results across all studied taxa for both image and DNA modalities, indicating there are still benefits from dataset scale at this size. By including the text modality, we further improve accuracy at the higher taxa levels. Interestingly, including the text modality results in slightly lower performance at the species level. This is likely due to the sparse availability of species labels in the training data, as only 9% of records having species labels. For additional details and analysis, see Appendix C.

5 Conclusion

We present the BIOSCAN-5M dataset, a valuable resource for the machine learning community containing over 5 million arthropod specimens. To highlight the dataset's multimodal capabilities, we conducted three benchmark experiments that leverage images, DNA barcodes, and textual taxonomic annotations for fine-grained taxonomic classification and zero-shot clustering.

An open problem for biodiversity monitoring systems is handling novel species. To facilitate research in this space, our dataset includes partitions for both closed-world and open-world settings. Furthermore, we provide three distinct benchmark tasks, each evaluated down to species-level, demonstrating the real-world applicability of BIOSCAN-5M's multimodal features. These tasks include fine-grained taxonomic classification using DNA sequences, multimodal classification combining DNA, images, and taxonomic labels, and clustering of learned DNA and image embeddings.

We believe that the BIOSCAN-5M dataset will serve as a catalyst for further machine learning research in biodiversity, fostering innovations that can enhance our understanding and preservation of the natural world. By providing a curated multi-modal resource, we aim to support further initiatives in the spirit of TreeOfLife-10M (Stevens et al., 2024) and contribute to the broader goal of mapping and preserving global biodiversity. This dataset not only facilitates advanced computational approaches but also underscores the crucial intersection between technology and conservation science, driving forward efforts to protect our planet's diverse ecosystems for future generations.

Acknowledgments and Disclosure of Funding

We acknowledge the support of the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFT-2020-00073]. This research is also supported by an NVIDIA Academic Grant. This research was enabled in part by support provided by Calcul Québec[1] and the Digital Research Alliance of Canada[2]. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring[3] the Vector Institute. Data collection was enabled by funds from the Walder Foundation, a New Frontiers in Research Fund (NFRF) Transformation grant, a Canada Foundation for Innovation's (CFI) Major Science Initiatives (MSI) Fund and CFREF funds to the Food from Thought program at the University of Guelph. The authors also wish to acknowledge the team at the Centre for Biodiversity Genomics responsible for preparing, imaging, and sequencing specimens used for this study. We also thank Mrinal Goshalia for assistance with the cropping tool and annotation of images.

Author contributions

DS provided the original dataset and provided guidance on subsequent processing steps. ZG curated the dataset, created primary metadata file and removed invalid images. ZG and SCL implemented data analytics and statistical processing pipelines. ZG calculated the specimen size information. ZMG improved the image cropping tool and conducted experiments for cropping images. SCL cleaned inconsistent taxonomic labels, with assistance from ZG, and they finalized the metadata file. SCL partitioned the data, with assistance from ZMG. PMA and SCL conducted the DNA-based taxonomic classification experiments. SCL conducted the zero-shot transfer-learning clustering experiments, with assistance from PMA and NP. ZMG and ATW conducted the multimodal retrieval learning experiments. ZG and AXC packaged and released the dataset. ZG, SCL, ZMG, PMA, NP, ATW, JBH, and IZ authored the manuscript text and figures. AXC, PF, GWT, LK, JBH, and SCL provided guidance on experimental design. All authors reviewed the manuscript.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi:10.1016/S0022-2836(05)80360-2.

Badirli, S., Akata, Z., Mohler, G., Picard, C., and Dundar, M. Fine-grained zero-shot learning with DNA as side information. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, December 2021. doi:10.48550/arXiv.2109.14133.

Badirli, S., Picard, C. J., Mohler, G., Richert, F., Akata, Z., and Dundar, M. Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, 14(6):1515–1530, 2023. doi:10.1111/2041-210X.14104.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. doi:10.48550/arxiv.2304.12210.

Bardes, A., Ponce, J., and LeCun, Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. doi:10.48550/arxiv.2105.04906.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, 2005. doi:10.1098/rstb.2005.1725.

[1] https://calculquebec.ca

[2] https://alliancecan.ca

[3] https://vectorinstitute.ai/partnerships/current-partners/

Braukmann, T. W., Ivanova, N. V., Prosser, S. W., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., and Hebert, P. D. Metabarcoding a diverse arthropod mock community. *Molecular ecology resources*, 19(3):711–727, 2019. doi:10.1111/1755-0998.13008.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. doi:10.48550/arXiv.2005.14165.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., et al. Biodiversity loss and its impact on humanity. *Nature*, 486 (7401):59–67, 2012. doi:10.1038/nature11148.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *Proc. of the European Conference on Computer Vision*, pp. 213–229. Springer, 2020. doi:10.1007/978-3-030-58452-8_13.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. doi:10.48550/arXiv.2104.14294.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. doi:10.48550/arxiv.2104.02057.

Chia, P. J., Attanasio, G., Bianchi, F., Terragni, S., Magalhães, A. R., Goncalves, D., Greco, C., and Tagliabue, J. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. doi:10.1038/s41598-022-23052-9.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023. doi:10.1101/2023.01.11.523679.

Damerau, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964. doi:10.1145/363958.363994.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009. doi:10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=YicbFdNTTy.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021b. doi:10.48550/arxiv.2010.11929.

Duraiappah, A. K., Naeem, S., Agardy, T., Ash, N. J., Cooper, H. D., Díaz, S., Faith, D. P., Mace, G., McNeely, J. A., Mooney, H. A., et al. Ecosystems and human well-being: biodiversity synthesis; a report of the Millennium Ecosystem Assessment. Technical report, World Resources Institute, 2005. URL https://www.millenniumassessment.org/documents/document.356.aspx.pdf.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. *Cluster Analysis*. Wiley, January 2011. doi:10.1002/9780470977811.

Flicek, P., Amode, M., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C., Gordon, L., et al. Ensembl 2014. *Nucleic Acids Res.*, 42:D749–55, 2014. doi:10.1093/nar/gkt1196.

Galloway, A., Taylor, G. W., Ramsay, A., and Moussa, M. The Ciona17 dataset for semantic segmentation of invasive species in a marine aquaculture environment. In *Proceedings of the Conference on Computer and Robot Vision (CRV)*, pp. 361–366. IEEE, 2017. doi:10.1109/CRV.2017.46.

Garcin, C., Joly, A., Bonnet, P., Lombardo, J.-C., Affouard, A., Chouet, M., Servajean, M., Lorieul, T., and Salmon, J. Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.

Gharaee, Z., Gong, Z., Pellegrino, N., Zarubiieva, I., Haurum, J. B., Lowe, S. C., McKeown, J. T. A., Ho, C. C. Y., McLeod, J., Wei, Y.-Y. C., Agda, J., Ratnasingham, S., Steinke, D., Chang, A. X., Taylor, G. W., and Fieguth, P. A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43593–43619. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/87dbbdc3a685a97ad28489a1d57c45c1-Paper-Datasets_and_Benchmarks.pdf.

Gong, Z., Wang, A. T., Haurum, J. B., Lowe, S. C., Taylor, G. W., and Chang, A. X. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv preprint arXiv:2405.17537*, 2024. doi:10.48550/arxiv.2405.17537.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. doi:10.48550/arxiv.2103.01988.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021. doi:10.1145/3458754.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a. doi:10.1109/CVPR.2016.90.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016b. doi:10.1109/CVPR.2016.90.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi:10.1109/CVPR52688.2022.01553.

He, W., Han, K., Nie, Y., Wang, C., and Wang, Y. Species196: A one-million semi-supervised dataset for fine-grained species recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

Hebert, P. D., Cywinska, A., Ball, S. L., and DeWaard, J. R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512): 313–321, 2003. doi:10.1098/rspb.2002.2218.

Hickling, R., Roy, D. B., Hill, J. K., Fox, R., and Thomas, C. D. The distributions of a wide range of taxonomic groups are expanding polewards. *Global change biology*, 12(3):450–455, 2006. doi:10.1111/j.1365-2486.2006.01116.x.

Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., Krishna, R., and Shapiro, L. Quilt-1M: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilyas, T., Arsa, D. M. S., Ahmad, K., Jeong, Y. C., Won, O., Lee, J. H., and Kim, H. CWD30: A comprehensive and holistic dataset for crop weed recognition in precision agriculture. *arXiv preprint arXiv:2305.10084*, 2023. doi:10.48550/arxiv.2305.10084.

International Barcode of Life Consortium. Barcode of Life Data System, 2024. URL https://boldsystems.org/.

Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395–405, 2012. doi:10.1038/nrg3208.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi:10.1093/bioinformatics/btab083.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi:10.1109/TBDATA.2019.2921572.

Katoh, K. and Standley, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013. doi:10.1093/molbev/mst010.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bonisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., et al. TRY – a global database of plant traits. *Global Change Biology*, 17(9):2905–2935, 2011. doi:10.1111/j.1365-2486.2011.02451.x.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., et al. A draft map of the human proteome. *Nature*, 509:575–581, 2014. doi:10.1038/nature13302.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi:10.48550/arxiv.1412.6980.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. B. Leafsnap: A computer vision system for automatic plant species identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 502–516, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33709-3. doi:10.1016/j.ecoinf.2017.05.005.

Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics Doklady*, 1966.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *Proc. of the European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi:10.1007/978-3-319-10602-1_48.

Linnaeus, C. *Systema Naturae per regna tria naturae, secundum classes, ordines, genera, species; cum characteribus, differentiis, synonymis, locis*, volume 1. Holmiae, Impensis Direct. Laurentii Salvii, 1758-1759, 10th edition, 1758.

Liu, X., Min, W., Mei, S., Wang, L., and Jiang, S. Plant disease recognition: A large-scale benchmark dataset and a visual region and loss reweighting approach. *IEEE Transactions on Image Processing*, 30:2003–2015, 2021. doi:10.1109/TIP.2021.3049334.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. doi:10.48550/arxiv.1711.05101. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lowe, S. C., Haurum, J. B., Oore, S., Moeslund, T. B., and Taylor, G. W. An empirical study into clustering of unseen datasets with self-supervised encoders. *arXiv preprint arXiv:2406.02465*, 2024a. doi:10.48550/arxiv.2406.02465.

Lowe, S. C., Misiuk, B., Xu, I., Abdulazizov, S., Baroi, A. R., Bastos, A. C., Best, M., Ferrini, V., Friedman, A., Hart, D., Hoegh-Guldberg, O., Ierodiaconou, D., Mackin-McLaughlin, J., Markey, K., Menandro, P. S., Monk, J., Nemani, S., O'Brien, J., Oh, E., Reshitnyk, L. Y., Robert, K., Roelfsema, C. M., Sameoto, J. A., Schimel, A. C. G., Thomson, J. A., Wilson, B. R., Wong, M. C., Brown, C. J., and Trappenberg, T. BenthicNet: A global compilation of seafloor images for deep learning applications. *arXiv preprint arXiv:2405.05241*, 2024b. doi:10.48550/arxiv.2405.05241.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L. P., et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023. doi:10.48550/arxiv.2307.12914.

Lu, Y. and Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178:105760, 2020. doi:10.1016/j.compag.2020.105760.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi:10.48550/arxiv.1802.03426.

Millan Arias, P., Sadjadi, N., Safari, M., Gong, Z., Wang, A. T., Lowe, S. C., Bruslund Haurum, J., Zarubiieva, I., Steinke, D., Kari, L., et al. BarcodeBERT: Transformers for biodiversity analysis. *arXiv preprint arXiv:2405.17537v1*, 2023. doi:10.48550/arxiv.2405.17537.

Mock, F., Kretschmer, F., Kriese, A., Böcker, S., and Marz, M. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences*, 119(35):e2122636119, 2022. doi:10.1073/pnas.2122636119.

Moritz, C. and Cicero, C. DNA barcoding: promise and pitfalls. *PLOS Biology*, 2(10):e354, 2004. doi:10.1371/journal.pbio.0020354.

Network, C. G. A. R., Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–20, 2013. doi:10.1038/ng.2764.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36, 2024a.

Nguyen, H., Truong, T., Nguyen, X., Dowling, A., Li, X., and Luu, K. Insect-Foundation: A foundation model and large-scale 1M dataset for visual insect understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21945–21955, Los Alamitos, CA, USA, Jun 2024b. IEEE Computer Society. doi:10.1109/CVPR52733.2024.02072.

Olsen, A., Konovalov, D. A., Philippa, B., Ridd, P., Wood, J. C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., et al. DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019. doi:10.1038/s41598-018-38343-3.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., et al. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, 637:1295–1310, 2018. doi:10.1016/j.scitotenv.2018.05.002.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023. doi:10.48550/arXiv.2306.15794.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. doi:10.48550/arXiv.2103.00020.

Ratnasingham, S. and Hebert, P. D. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLOS ONE*, 8(7):1–16, 2013. doi:10.1371/journal.pone.0066213.

Ruppert, K. M., Kline, R. J., and Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17:e00547, 2019. doi:10.1016/j.gecco.2019.e00547.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, April 2015. doi:10.1007/s11263-015-0816-y.

Sala, O. E., Stuart Chapin, F., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L. F., Jackson, R. B., Kinzig, A., et al. Global biodiversity scenarios for the year 2100. *Science*, 287(5459):1770–1774, 2000. doi:10.1126/science.287.5459.1770.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

Sheridan, J. A. and Bickford, D. Shrinking body size as an ecological response to climate change. *Nature climate change*, 1(8):401–406, 2011. doi:10.1038/nclimate1259.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of residual networks using large learning rates. *arXiv preprint*, arXiv:1708.07120, 2017. doi:10.48550/arxiv.1708.07120.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Sokal, R. and Sneath, P. *Principles of Numerical Taxonomy*. Series of books in biology. W. H. Freeman, 1963.

Steckel, R. H. Stature and the standard of living. *Journal of economic literature*, 33(4):1903–1940, 1995.

Steinke, D., Ratnasingham, S., Agda, J., Ait Boutou, H., Box, I. C. H., Boyle, M., Chan, D., Feng, C., Lowe, S. C., McKeown, J. T. A., McLeod, J., Sanchez, A., Smith, I., Walker, S., Wei, C. Y.-Y., and Hebert, P. D. N. Towards a taxonomy machine: A training set of 5.6 million arthropod images. *Data*, 9(11), 2024. ISSN 2306-5729. doi:10.3390/data9110122.

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., and Su, Y. BioCLIP: A vision foundation model for the tree of life. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi:10.1109/CVPR52733.2024.01836.

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I., and Pawlowski, J. Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin*, 127:139–149, 2018. doi:10.1016/j.marpolbul.2017.11.065.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019. doi:10.48550/arxiv.1908.08962.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015. doi:10.1109/CVPR.2015.7298658.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and MacAodha, O. Benchmarking representation learning for natural world image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12879–12888, 2021. doi:10.1109/CVPR46437.2021.01269.

Vaze, S., Hant, K., Vedaldi, A., and Zisserman, A. Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2022. doi:10.1109/CVPR52688.2022.00734.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. URL http://jmlr.org/papers/v11/vinh10a.html.

Wang, Q.-J., Zhang, S.-Y., Dong, S.-F., Zhang, G.-C., Yang, J., Li, R., and Wang, H.-Q. Pest24: A large-scale very small object data set of agricultural pests for multi-target detection. *Computers and Electronics in Agriculture*, 175:105585, 2020. ISSN 0168-1699. doi:10.1016/j.compag.2020.105585.

Wegner, J. D., Branson, S., Hall, D., Schindler, K., and Perona, P. Cataloging public objects using aerial and street-level images — urban trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6014–6023, 2016. doi:10.1109/CVPR.2016.647.

Wu, X., Zhan, C., Lai, Y.-K., Cheng, M.-M., and Yang, J. IP102: A large-scale benchmark dataset for insect pest recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8779–8788, 2019. doi:10.1109/CVPR.2019.00899.

Xu, D., Zhao, Y., Hao, X., and Meng, X. Pink-Eggs Dataset V1: A step toward invasive species management using deep learning embedded solutions. *arXiv preprint arXiv:2305.09302*, 2023. doi:10.48550/arxiv.2305.09302.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. doi:10.48550/arxiv.2303.00915.

Zhou, X. and Zhang, N. L. Deep clustering with features from self-supervised pretraining. *arXiv preprint arXiv:2207.13364*, 2022. doi:10.48550/arxiv.2207.13364.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023. doi:10.48550/arxiv.2306.15006.

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., Wang, Z., and Liu, H. DNABERT-S: Learning species-aware DNA embedding with genome foundation models. *arXiv preprint arXiv:2402.08777*, 2024. doi:10.48550/arxiv.2402.08777.

Appendices

A DNA-based Taxonomic Classification — Additional Experiments

As described in the main text (§4.2), we leverage all data splits in the BIOSCAN-5M dataset by adopting a semi-supervised learning approach. Specifically, we train a model using self-supervision on the unlabelled partition of the data, followed by fine-tuning on the train split. Our experimental setup is illustrated in Figure 5.

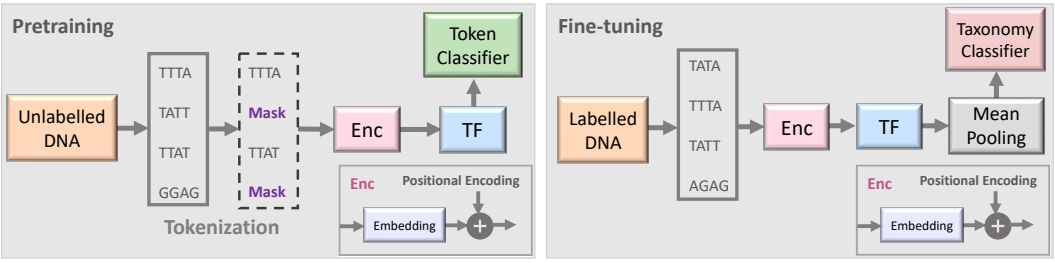


Figure 5: **DNA-based taxonomic classification methodology.** Two stages of the proposed semi-supervised learning set-up based on BarcodeBERT (Millan Arias et al., 2023). (1) Pretraining: DNA sequences are tokenized using non-overlapping k-mers and 50% of the tokens are masked for the MLM task. Tokens are encoded and fed into a transformer model. The output embeddings are used for token-level classification. (2) Fine-tuning: All DNA sequences in a dataset are tokenized using non-overlapping k-mer tokenization and all tokenized sequences, without masking, are passed through the pretrained transformer model. Global mean-pooling is applied over the token-level embeddings and the output is used for taxonomic classification.

A.1 Pretraining details

We pretrain the model on the 2,283,900 unique DNA sequences from the `pretrained` partition and the 41,232 unique sequences from the `other_heldout` partition, totalling 2,325,132 pretraining DNA samples. For all samples, trailing `N` characters are removed and all sequences are truncated at 660 nucleotides. Note that leading `N` characters are retained since they are likely to correspond to true unknown nucleotides in the barcode. The model was pretrained using the same MLM loss function and training configurations as in BarcodeBERT (Millan Arias et al., 2023). Specifically, we use a non-overlapping k-mer-based tokenizer and a transformer model with 12 transformer layers, each having 12 attention heads. However, we included a random offset of at most k nucleotides to each sequence as a data augmentation technique to enhance the sample efficiency. We use a learning rate of 2×10^{-4}, a batch size of 128, a OneCycle scheduler (Smith & Topin, 2017), and the AdamW optimizer (Loshchilov & Hutter, 2019), training the model for 35 epochs. In addition to using the architecture reported in BarcodeBERT, we performed a parameter search to determine the optimal k-mer tokenization length and model size, parameterized by the number of layers and heads in the transformer model, in order to identify an optimal architecture configuration. After pretraining, we fine-tuned the model with cross-entropy supervision for species-level classification. The pre-training stage takes approximately 50 hours using four Nvidia A40 GPUs and the fine-tuning stage of the 4-12-12 models takes 2.5 hours in four Nvidia A40 GPUs.

A.2 Baseline Models

There has been a growing number of SSL DNA language models proposed in recent literature, most of which are based on the transformer architecture and trained using the MLM objective (Ji et al., 2021; Zhou et al., 2023, 2024). These models differ in the details of their model architecture, tokenization strategies, and training data but the underlying principles remain somewhat constant. An exception to this trend is the HyenaDNA (Nguyen et al., 2024a) model, which stands out by its use of a state space model (SSM) based on the Hyena architecture (Poli et al., 2023) and trained for next token prediction. For evaluation, we utilized the respective pre-trained models from Huggingface ModelHub, specifically:

- DNABERT-2: zhihan1996/DNABERT-2-117M
- DNABERT-S: zhihan1996/DNABERT-S
- NT: InstaDeepAI/nucleotide-transformer-v2-50m-multi-species
- HyenaDNA: LongSafari/hyenadna-tiny-1k-seqlen

The BarcodeBERT implementation was taken from `https://github.com/Kari-Genomics-Lab/BarcodeBERT`. All the models, including our pretrained models, were fine-tuned for 35 epochs with a batch size of 32 or 128 and a learning rate of 1×10^{-4} per 64 samples in the batch with the OneCycle LR schedule (Smith & Topin, 2017).

A.3 Linear probe training

A linear classifier is applied to the embeddings generated by all the pretrained models for species-level classification. The parameters of the model are learned using stochastic gradient descent with a constant learning rate of 0.01, momentum $\mu = 0.9$ and weight $\lambda = 1 \times 10^{-5}$.

For the hyperparameter search, shown in Table 6, our linear probe is performed using the same methodology as the fine-tuning stage, except the encoder parameters are frozen.

Table 6: Search over the space of k-mer tokenization length and transformer architectures (number of layers and heads). For fine-tuned and linear probe, we show the class-balanced accuracy (%) on the closed-world `val` partition, and for 1-NN probe, we show the class-balanced accuracy on the `val_unseen` partition. Bold: architecture with highest accuracy for the row. Underlined: second highest accuracy.

	4 layers, 4 heads				6 layers, 6 heads				12 layers, 12 heads			
Evaluation	$k=2$	$k=4$	$k=6$	$k=8$	$k=2$	$k=4$	$k=6$	$k=8$	$k=2$	$k=4$	$k=6$	$k=8$
Fine-tuned	93.8	97.8	<u>98.7</u>	**98.9**	92.4	97.9	49.4	<u>98.7</u>	93.8	98.1	0.0	0.0
Linear probe	32.2	<u>79.8</u>	76.4	**97.1**	34.3	58.9	8.9	<u>79.7</u>	16.4	3.2	0.0	0.0
1-NN	43.1	**50.7**	35.0	<u>46.4</u>	46.2	37.2	23.4	37.9	29.1	28.3	0.0	0.1

B Zero-Shot Clustering — Additional Experiments

As described in §4.3, we performed a series of zero-shot clustering experiments to establish how pretrained image and DNA models could handle the challenge of grouping together repeat observations of novel/unseen species. Our methodology is illustrated in Figure 6.

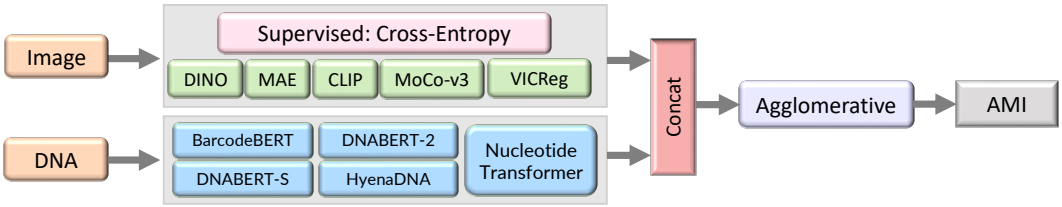


Figure 6: **Zero-shot clustering methodology.** Images and DNA are each passed through one of several pretrained encoders. These representations are clustered with Agglomerative Clustering.

B.1 Experiment resources

All zero-shot clustering experiments were performed on a compute cluster with the job utilizing two CPU cores (2x Intel Xeon Gold 6148 CPU @ 2.40GHz) and no more than 20 GB of RAM. The typical runtime per experiment was around 4.5 hours.

B.2 Accounting for Duplicated DNA Barcode Sequences

In our main experiments, we found the performance of DNA-embedding clusterings greatly outperformed that of image-embeddings. However, it is worth considering that there are fewer unique DNA barcodes than images. The mean number of samples per barcode is around two. This provides clustering methods using DNA with an immediate advantage as some stimuli compare as equal and are trivially grouped together, irrespective of the encoder.

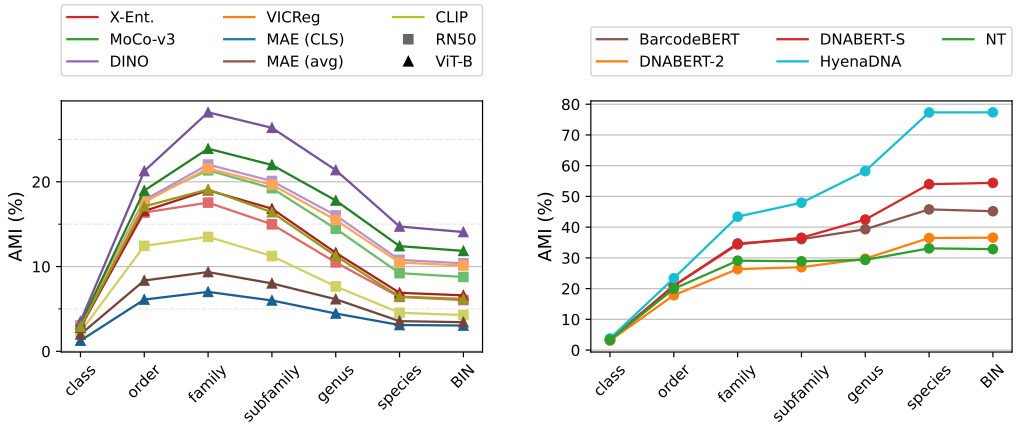


Figure 7: **Zero-shot clustering** AMI (%) performance across taxonomic ranks on `test` and `test_unseen` data, with **one sample per barcode**.

To account for this, we repeated our analysis with only one sample per barcode. Our results, shown in Figure 7, indicate that both image and DNA based clusterings are reduced in performance when the number of samples per barcode is reduced to one. This is explained in part by the fact that many species will be reduced to a single observation, which is challenging for clusterers to handle. We found that the performance of most DNA-encoders fell by around half (e.g. from 80% to 40% AMI) when the number of samples per barcode was reduced to one. However the best-performing DNA-encoder, HyenaDNA, was not greatly affected, with its performance reduced from 90% to 80% AMI on the harder clustering task.

Table 7: **Cross-modal zero-shot clustering** AMI (%) performance, on `test` and `test_unseen` data, with **one sample per barcode**.

			DNA encoder				
Architecture	Image encoder	*Image-only*	BarcodeBERT	DNABERT-2	DNABERT-S	HyenaDNA	NT
—	*DNA-only*	–	47	52	<u>63</u>	**81**	36
ResNet-50	X-Ent.	5	30	26	32	9	12
	MoCo-v3	8	29	23	27	11	11
	DINO	<u>11</u>	31	28	31	15	14
	VICReg	10	30	26	30	13	13
	CLIP	6	25	21	25	9	9
ViT-B	X-Ent.	7	33	35	42	13	14
	MoCo-v3	<u>13</u>	38	43	49	21	20
	DINO	**15**	38	45	<u>51</u>	23	21
	MAE (CLS)	5	33	33	40	10	13
	MAE (avg)	3	29	26	32	7	9
	CLIP	7	34	37	44	14	16

B.3 Cross-modal embedding clustering

We additionally considered the effect of clustering the embeddings from both modalities at once, achieved by concatenating an image embedding and a DNA embedding to create a longer feature vector per sample. As shown in Table 7, we find that combining image features with DNA features results in a worse performance at species-level clustering.

In preliminary experiments (not shown) we found that the magnitude of the vectors greatly impacted the performance, as large image embeddings would dominate DNA embeddings with a smaller magnitude. We considered standardizing the embeddings before concatenation with several methods (L2-norm, element-wise z-score, average z-score) and found element-wise z-score gave the best performance, a step which we include in these results. Even with this, the performance falls when we add image embeddings to the DNA embeddings. We note that the best DNA-only encoder, HyenaDNA, has the largest drop in performance, which we hypothesize is because it has the shortest embedding dimensions of 128-d compared with NT (512-d) and the BERT-based models (768-d).

C Multi-Modal Learning — Additional Experiments

As described in §4.4, we trained a multimodal model with an aligned embedding space across the images, DNA, and taxonomic labels. Our methodology is illustrated in Figure 8.

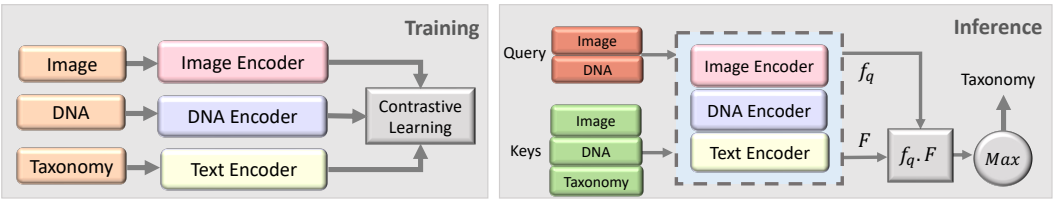


Figure 8: **Multi-modal learning methodology.** Our experiments using CLIBD (Gong et al., 2024) are conducted in two steps. (1) Training: Multiple modalities, including RGB images, textual taxonomy, and DNA sequences, are encoded separately, and trained using a contrastive loss function. (2) Inference: Image vs DNA embedding is used as a query, and compared to the embeddings obtained from a database of image, DNA and text (keys). The cosine similarity is used to find the closest key embedding, and the corresponding taxonomic label is used to classify the query.

C.1 Model training and inference

We illustrate our model training and inference methodology in Figure 8. For our multimodal model, we start with pretrained encoders for image, DNA, and taxonomic labels. We use contrastive learning to fine-tune the image, DNA, and text encoders. During inference, we compare the embedding of the query image or DNA input to a key database of embeddings from images, DNA, or taxonomy labels using cosine similarity, and we predict the query's taxonomy based on the taxonomy of the closest retrieved key embeddings.

C.2 Additional experiments

In the main paper, we reported the macro accuracy of our models. In Table 8, we report the micro accuracy to compare performance when averaged over individual samples rather than classes. The results show similar trends to the macro accuracy (Figure 8), with the model trained on the BIOSCAN-5M dataset performing best for broader taxa, especially in image-to-image and image-to-DNA inference setups. Results are more mixed at the species level due in part to the challenge of species classification, highlighting the importance of further research at this fine-grained level.

C.3 Retrieval examples

Figure 9 shows image retrieval examples using images as queries and DNA as keys. These demonstrate the ability of the model to classify taxonomy based on retrieval and the visual similarities of the retrieved images corresponding to the most closely matched DNA embeddings.

Table 8: Top-1 *micro* accuracy (%) on the test set for using different amount of training data (1 million vs 5 million records from BIOSCAN-5M) and different combinations of aligned embeddings (image, DNA, text) during contrastive training. We show results for using image-to-image, DNA-to-DNA, and image-to-DNA query and key combinations. As a baseline, we show the results prior to contrastive learning (no alignment). We report the accuracy for seen and unseen species, and the harmonic mean (H.M.) between these (bold: highest acc.).

		Aligned embeddings			DNA-to-DNA			Image-to-Image			Image-to-DNA		
Taxon	# Records	Img	DNA	Txt	Seen	Unseen	H.M.	Seen	Unseen	H.M.	Seen	Unseen	H.M.
Order	—	✗	✗	✗	98.9	99.3	99.1	94.2	97.0	95.6	18.3	14.7	16.3
	1M	✓	✓	✓	**100.0**	**100.0**	**100.0**	99.3	99.6	99.5	98.7	99.2	98.9
	5M	✓	✓	✗	**100.0**	**100.0**	**100.0**	**99.5**	**99.7**	**99.6**	**99.4**	99.5	**99.5**
	5M	✓	✓	✓	**100.0**	**100.0**	**100.0**	**99.5**	**99.7**	**99.6**	99.3	**99.6**	**99.5**
Family	—	✗	✗	✗	96.5	97.3	96.9	72.9	76.0	74.4	1.7	1.9	1.8
	1M	✓	✓	✓	99.8	99.8	99.8	95.5	96.8	96.2	90.6	89.1	89.9
	5M	✓	✓	✗	**99.9**	**100.0**	99.9	96.8	97.9	97.4	94.0	93.1	93.5
	5M	✓	✓	✓	**99.9**	**100.0**	**100.0**	**97.0**	**98.3**	**97.7**	**94.6**	**94.4**	**94.5**
Genus	—	✗	✗	✗	94.0	93.5	93.7	47.8	47.0	47.4	0.2	0.0	0.1
	1M	✓	✓	✓	99.3	98.8	99.0	86.0	85.9	86.0	68.1	52.3	59.2
	5M	✓	✓	✗	**99.6**	**99.8**	**99.7**	90.6	91.6	91.1	**79.5**	65.0	71.5
	5M	✓	✓	✓	**99.6**	**99.8**	**99.7**	**91.0**	**92.1**	**91.5**	79.3	**66.3**	**72.2**
Species	—	✗	✗	✗	91.6	84.8	88.1	31.9	19.1	23.9	0.0	0.0	0.0
	1M	✓	✓	✓	98.3	95.0	96.6	75.1	57.5	65.1	47.9	10.4	17.0
	5M	✓	✓	✗	**98.9**	97.4	98.2	82.7	**68.3**	**74.8**	**64.2**	**18.7**	**29.0**
	5M	✓	✓	✓	**98.9**	**97.7**	**98.3**	**82.8**	67.6	74.4	61.7	17.8	27.7

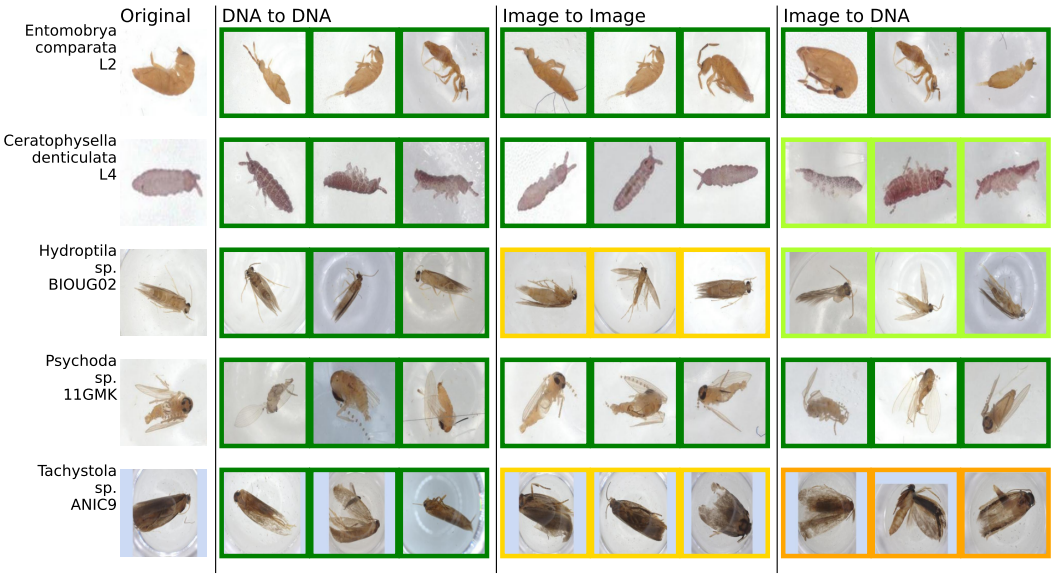


Figure 9: *Example query-key pairs.* Top-3 nearest specimens from the unseen validation-key dataset retrieved based on the cosine-similarity for DNA-to-DNA, image-to-image, and image-to-DNA retrieval. Box colour indicates whether the retrieved samples had the same species (green), genus (light-green), family (yellow), or order (orange) as the query.

D Dataset

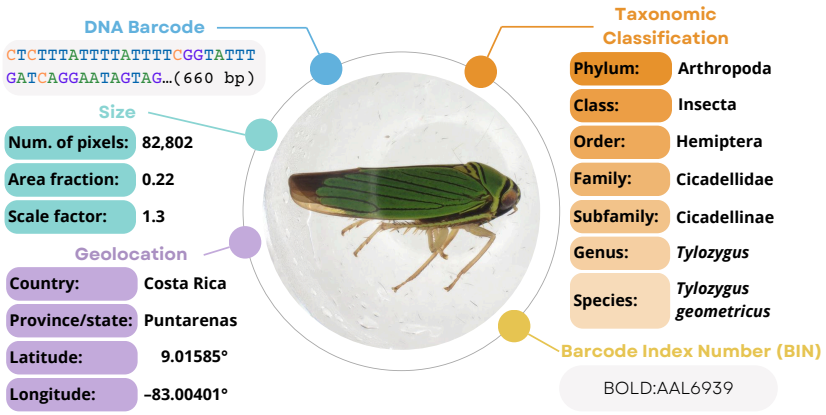


Figure 10: The BIOSCAN-5M Dataset provides taxonomic labels, a DNA barcode sequence, barcode index number, a high-resolution image along with its cropped and resized versions, as well as size and geographic information for each sample.

E Ethics and responsible use

The BIOSCAN project was instigated by the International Barcode of Life (iBOL) Consortium, which has collected a large dataset of manually-labelled images of organisms (International Barcode of Life Consortium, 2024; Steinke et al., 2024). As part of our project, we conducted a thorough review to identify any potential ethical issues related to the inclusion of our data sources. After careful evaluation, we did not find any ethical concerns. Therefore, we confirm that this work adheres to all relevant ethical standards and guidelines.

F Dataset availability and maintenance

To explore more about the BIOSCAN-5M dataset, kindly visit the following landing page: `https://biodiversitygenomics.net/5M-insects/`.

The BIOSCAN-5M dataset and all its contents are available in a GoogleDrive Folder. The Google Drive folder serves as the primary repository for the BIOSCAN-5M dataset, ensuring ongoing maintenance and the potential addition of new content as necessary. It will be gradually updated to address any data issues that may arise.

The Google Drive folder contains the following dataset contents:

- **BIOSCAN_5M_IMAGES**: This directory contains images:
  - `BIOSCAN_5M_original_full`: The original full-size images.
  - `BIOSCAN_5M_original_256`: The original images resized to 256 pixels on their shorter side.
  - `BIOSCAN_5M_cropped`: The cropped images.
  - `BIOSCAN_5M_cropped_256`: The cropped images resized to 256 pixels on their shorter side.
- **BIOSCAN_5M_METADATA**: This directory contains metadata:
  - `BIOSCAN_5M_Insect_Dataset_metadata_MultiTypes.zip`: A zip file containing both CSV and JSON formats of the metadata file.
- **BIOSCAN_5M_CropTool**: This directory contains our cropping tool components:
  - `bounding_box/BIOSCAN_5M_Insect_bbox.tsv`: A TSV file that includes bounding box information obtained from our cropping tool.

- `checkpoint/BIOSCAN_5M_Insect_cropping_tool.ckpt`: The model checkpoint used to crop the original full-size images, which generated the cropped images of the BIOSCAN-5M dataset.

Additionally, the dataset is released on several platforms, including Zenodo, Kaggle, and HuggingFace.

We provide a code repository for dataset manipulation, which supports tasks like reading images and metadata, cropping images, statistical processing, dataset splitting into pretrain, train, and evaluation, as well as running benchmark experiments presented in the BIOSCAN-5M paper. Additionally, it offers a Python package for working with the BIOSCAN-5M dataset, designed in the style of torchvision's `VisionDataset` class. To access the BIOSCAN-5M code repository, please visit `https://github.com/bioscan-ml/BIOSCAN-5M`.

G Licensing

Table 9 shows all the copyright associations related to the BIOSCAN-5M dataset with the corresponding names and contact information.

Table 9: Copyright associations related to the BIOSCAN-5M dataset

Copyright Associations	**Name & Contact**
Image Photographer	CBG Robotic Imager
Copyright Holder	CBG Photography Group
Copyright Institution	Centre for Biodiversity Genomics (email: CBGImaging@gmail.com)
Copyright License	Creative Commons Attribution 3.0 Unported (CC BY 3.0)
Copyright Contact	collectionsBIO@gmail.com
Copyright Year	2021

The authors state that they bear all responsibility in case of violation of usage rights.

H RGB images

The BIOSCAN-5M dataset comprises resized and cropped images, as introduced in BIOSCAN-1M Insect (Gharaee et al., 2023). We have provided various packages of the BIOSCAN-5M dataset, detailed in Table 10, each tailored for specific purposes.

- original_full: The raw images of the dataset, typically 1024×768 pixels.
- cropped: Images after cropping with our cropping tool (see §Q.1).
- original_256: Original images resized to 256 on their shorter side (most 341×256 pixels).
- cropped_256: Cropped images resized to 256 on their shorter side.

Among these, the original_256 and cropped_256 packages are specifically provided for experimentation as they are small and easy to work with. Therefore, using our predefined split partitions, we provide per-split experimental packages in addition to the packages with all the original_256 and cropped_256 images.

Accessing the dataset images is facilitated by the following directory structure used to organize the dataset images:

```
bioscan5m/images/[imgtype]/[split]/[chunk]/[processid.jpg]
```

where `[imgtype]` can be `original_full`, `cropped`, `original_256`, or `cropped_256`. The `[split]` values can be `pretrain`, `train`, `val`, `test`, `val_unseen`, `test_unseen`, `key_unseen`, or `other_heldout`. Note that the `val`, `test`, `val_unseen`, `test_unseen`, `key_unseen`, and `other_heldout` splits are within the evaluation partition of the original_256 and cropped_256 image packages.

Table 10: Various downloadable packages of the images comprising the BIOSCAN-5M dataset.

Image set	Package	Partition(s)	Size (GB)	# Parts
original_full	BIOSCAN_5M_original_full.zip	All	200	5
cropped	BIOSCAN_5M_cropped.zip	All	77.2	2
original_256	BIOSCAN_5M_original_256.zip	All	35.2	1
	BIOSCAN_5M_original_256_pretrain.zip	Pretrain	31.7	1
	BIOSCAN_5M_original_256_train.zip	Train	2.1	1
	BIOSCAN_5M_original_256_eval.zip	Evaluation	1.4	1
cropped_256	BIOSCAN_5M_cropped_256.zip	All	36.4	1
	BIOSCAN_5M_cropped_256_pretrain.zip	Pretrain	33.0	1
	BIOSCAN_5M_cropped_256_train.zip	Train	2.1	1
	BIOSCAN_5M_cropped_256_eval.zip	Evaluation	1.4	1

The `[chunk]` is determined by using the first one or two characters of the MD5 checksum (in hexadecimal) of the `processid`. This method ensures that the chunk name is purely deterministic and can be computed directly from the `processid`. As a result, the `pretrain` split organizes files into 256 directories by using the first two letters of the MD5 checksum of the `processid`. For the `train` and `other_heldout` splits, files are organized into 16 directories using the first letter of the MD5 checksum. The remaining splits do not use chunk directories since each split has less than 50 k images.

I Metadata file

To enrich the metadata of our published dataset, we provide integrated structured metadata conforming to Web standards. Our dataset's metadata file is titled **BIOSCAN_5M_Insect_Dataset_metadata**. We provide two versions of this file: one in CSV format (**.csv**) and the other in JSON-LD format (**.jsonld**). Accessing the dataset metadata files is facilitated by the following directory structure used to organize the dataset images:

```
bioscan5m/metadata/[type]/BIOSCAN_5M_Insect_Dataset_metadata.[type_extension]
```

In this structure, `[type]` refers to the file type of the metadata file, which can be either CSV or JSON-LD. The `[type_extension]` indicates the corresponding file extensions, which are `csv` for CSV files and `jsonld` for JSON-LD files.

Table 11 outlines the fields of the metadata file and the description of their contents.

J Comparison between BIOSCAN-5M and BIOSCAN-1M

The six key differences between BIOSCAN-1M and BIOSCAN-5M are as follows:

1. **Increased data volume:** BIOSCAN-5M contains five times as many samples as BIOSCAN-1M.
2. **Greater data diversity:** BIOSCAN-5M is collected from a broader range of geographic locations (3 countries in BIOSCAN-1M; 47 countries in BIOSCAN-5M) and encompasses a wider variety of insect life (1 class and 16 orders in BIOSCAN-1M; 10 classes and 55 orders in BIOSCAN-5M).
3. **Enhanced post-processing:** The taxonomic labels in BIOSCAN-5M underwent a rigorous data cleaning pipeline to identify and resolve inconsistencies in the original data, resulting in more reliable labels compared to those in BIOSCAN-1M.
4. **Geographic and specimen size data:** This information is available in BIOSCAN-5M but not in BIOSCAN-1M.
5. **Comprehensive partitioning support:** BIOSCAN-5M offers robust support for both closed-world and open-world tasks, whereas BIOSCAN-1M only supports closed-world partitioning.

Table 11: Table presents fields of the metadata file of BIOSCAN-5M dataset.

	Field	Description	Type
1	processid	A unique number assigned by BOLD (International Barcode of Life Consortium).	String
2	sampleid	A unique identifier given by the collector.	String
3	taxon	Bio.info: Most specific taxonomy rank.	String
4	phylum	Bio.info: Taxonomic classification label at phylum rank.	String
5	class	Bio.info: Taxonomic classification label at class rank.	String
6	order	Bio.info: Taxonomic classification label at order rank.	String
7	family	Bio.info: Taxonomic classification label at family rank.	String
8	subfamily	Bio.info: Taxonomic classification label at subfamily rank.	String
9	genus	Bio.info: Taxonomic classification label at genus rank.	String
10	species	Bio.info: Taxonomic classification label at species rank.	String
11	dna_bin	Bio.info: Barcode Index Number (BIN).	String
12	dna_barcode	Bio.info: Nucleotide barcode sequence.	String
13	country	Geo.info: Country associated with the site of collection.	String
14	province_state	Geo.info: Province/state associated with the site of collection.	String
15	coord-lat	Geo.info: Latitude (WGS 84; decimal degrees) of the collection site.	Float
16	coord-lon	Geo.info: Longitude (WGS 84; decimal degrees) of the collection site.	Float
17	image_measurement_value	Size.info: Number of pixels occupied by the organism.	Integer
18	area_fraction	Size.info: Fraction of the original image the cropped image comprises.	Float
19	scale_factor	Size.info: Ratio of the cropped image to the cropped_256 image.	Float
20	inferred_ranks	An integer indicating at which taxonomic ranks the label is inferred.	Integer
21	split	Split set (partition) the sample belongs to.	String
22	index_bioscan_1M_insect	An index to locate organism in BIOSCAN-1M Insect metadata.	Integer
23	chunk	The packaging subdirectory name (or empty string) for this image.	String

6. **Enhanced benchmarking experiments:** BIOSCAN-1M included a baseline with an image-only model evaluated at order and family ranks. In contrast, BIOSCAN-5M features three baselines that leverage the multimodal aspects of the dataset (including DNA barcode sequences, textual taxonomic labels, and RGB images), allowing for performance exploration in both closed- and open-world settings.

K Focus and objectives

We have released dataset splits for closed-world and open-world settings, using labelled species data for evaluation and reserving unlabelled data for pretraining. Our splitting approach and configurations offer valuable resources to the ML community. BIOSCAN-5M experiments evaluate down to the species level. Additionally, we benchmark three distinct tasks to showcase BIOSCAN-5M's multimodal utility in real-world applications: fine-grained taxonomic classification with DNA sequences, classification using DNA, images, and taxonomic labels, and clustering of DNA and image embeddings.

K.1 Leveraging unlabelled and multimodal data for enhanced taxonomic classification

It's important to note that taxonomic classification from images presents greater challenges compared to DNA barcodes, as illustrated by our clustering experiments; thus, paired data can be valuable even when unlabelled. Additionally, data not labelled at the species level remains useful for pretraining, highlighting the crucial role of unlabelled data in model development. In BIOSCAN-5M, we employ BERT-style masked sequence modelling to pretrain and encode DNA sequences, complemented by contrastive learning to align image and DNA embeddings. This pretraining approach enhances the model's ability to generalize across various applications.

K.2 Lack of utilization of geographic and size information in models

In BIOSCAN-5M, we focus on biological (taxonomic labels) and genetic (DNA barcode sequences and BIN) data for fine-grained taxonomic classification, intentionally excluding geographic and size information from our experiments. Our rationale is that while geographic and size data can help rule out certain species (e.g., knowing a sample was collected in North America excludes species not found there, and knowing a sample's size eliminates species that do not grow that large), they alone do not provide sufficient information for accurate species classification. In contrast, image and genetic data are often sufficient for accurate species-level predictions.

We believe that models incorporating geographic and size data will need to do so alongside image and genetic data. Therefore, models using only image and genetic information serve as valuable baselines for future work that combines these data types. Given the complexities of integrating geographic and size data into our models, we prioritized establishing a broad range of image and genetic baselines in this study and plan to explore the incorporation of geographic and size data in future research. We anticipate that effective use of this additional information will enhance model performance and look forward to the community's advancements in this area.

L Dataset features statistics

This section provides additional information regarding the dataset, including a detailed statistical analysis of its diverse multimodal data types and processing methods.

L.1 Geographical information

The detailed statistical analysis of the geographical locations where the organisms were collected is presented in Table 12. This table indicates the number of distinct regions represented by country, province or state, along with their corresponding latitude and longitude. Additionally, Table 12 provides the count of labelled versus unlabelled records, as well as the class imbalance ratio (IR) for each location group within the dataset.

Table 12: The statistics for the columns indicating geographical locations where the specimens are collected.

Geo locations	Categories	Labelled	Labelled (%)	Unlabelled	Unlabelled (%)	IR
`country`	47	5,150,842	100.00	8	0.00	325,631.6
`province_state`	102	5,058,718	98.21	92,132	1.79	1,243,427.0
`coord-lat`	1,394	5,149,019	99.96	1,831	0.04	556,352.0
`coord-lon`	1,489	5,149,019	99.96	1,831	0.04	618,931.0
Location (lat, lon)	1,650	5,149,019	99.96	1,831	0.04	520,792.0

The latitude and longitude coordinates indicate that the dataset comprises 1,650 distinct regions with unique geographical locations shown by Table 12. However, a significant portion of the organisms—approximately 73.36%—were collected from the top 70 most populated regions, which represent only 4.24% of the total regions identified by their coordinates.

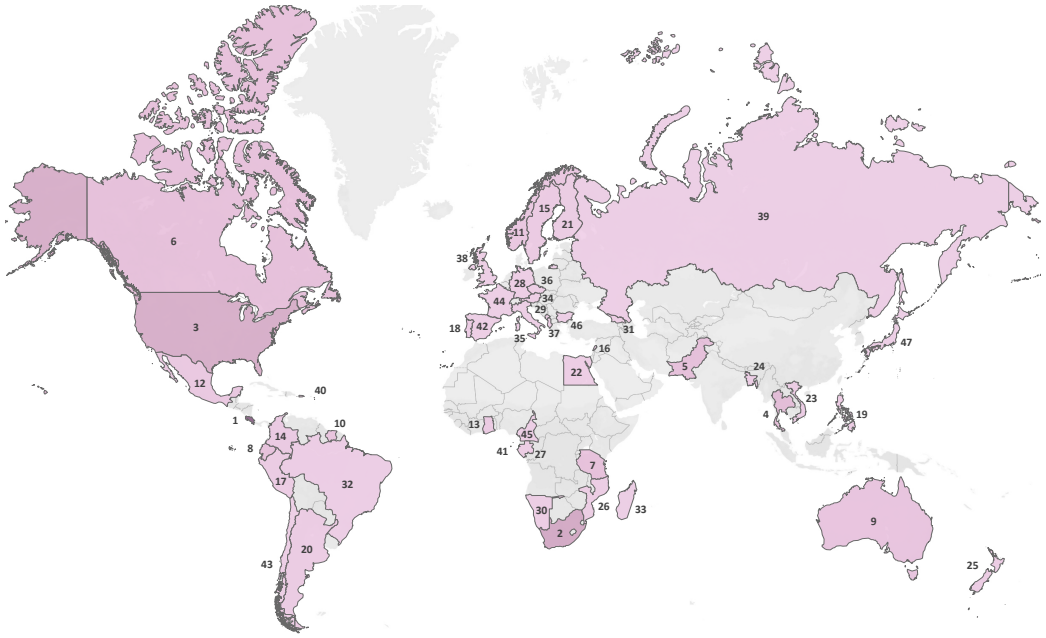
Figure 11 shows the distinct countries where the organisms were collected on the world map. The majority of the organisms, over 62%, were collected from Costa Rica.

L.2 Size information

Monitoring organism size is crucial as it can signal shifts in various factors affecting their lives, including food access, nutrition, and climate change (Sheridan & Bickford, 2011). For instance, in humans, limited access to nutrition correlates with a decrease in average height over generations (Steckel, 1995), reflecting environmental and economic changes. Tracking organism size offers insights into environmental shifts vital for biodiversity conservation (Hickling et al., 2006).

Pixel count. The raw dataset provides information about each organism's size by quantifying the total number of pixels occupied by the organism. This information is provided in the `image_measurement_value` field. Since the image capture settings are consistent for all images, irrespective of scale, as indicated by the organism's distance to the camera, the number of pixels occupied by the organism should approximate its size. Less than 1% of samples of the BIOSCAN-5M dataset do not have this information.

To provide a clearer understanding of the content in the `image_measurement_value` field, Figure 12 displays examples of original images along with their corresponding masks, highlighting the total number of pixels occupied by an organism. To determine the real-world size of the organism based on the number of pixels, it is also important to have the pixel to metric scaling factor. For the original



1. Costa Rica:	3,256,316	**9. Australia:**	90,664	**17. Peru:**	26,656	**25. New Zealand:**	14,184	**33. Madagascar:**	4,359	**41. Sao Tome and Principe:**	356
2. South Africa:	322,096	**10. Suriname:**	82,842	**18. Portugal:**	25,780	**26. Mozambique:**	12,217	**34. Austria:**	711	**42. Spain:**	345
3. United States:	281,411	**11. Norway:**	60,925	**19. Philippines:**	24,708	**27. Gabon:**	11,942	**35. Italy:**	701	**43. Chile:**	329
4. Thailand:	152,975	**12. Mexico:**	46,982	**20. Argentina:**	24,626	**28. Germany:**	11,310	**36. Czech Republic:**	515	**44. France:**	261
5. Pakistan:	126,990	**13. Ghana:**	38,256	**21. Finland:**	19,978	**29. Montenegro:**	10,869	**37. Albania:**	379	**45. Cameroon:**	203
6. Canada:	117,599	**14. Colombia:**	34,444	**22. Egypt:**	19,841	**30. Namibia:**	10,278	**38. United Kingdom:**	376	**46. Bulgaria:**	33
7. Tanzania:	108,945	**15. Sweden:**	27,912	**23. Vietnam:**	16,395	**31. Georgia:**	9,205	**39. Russia:**	361	**47. Japan:**	10
8. Ecuador:	104,676	**16. Lebanon:**	27,744	**24. Bangladesh:**	15,352	**32. Brazil:**	7,427	**40. Antigua and Barbuda:**	358		

Figure 11: Global distribution of sample collection efforts. The countries are ranked by the number of samples collected.

full sized images, most of the images are captured using a Keyence imaging system with a known pixel to millimetre scaling. See §Q.1 for details on the pixel scale and how to determine it for cropped and resized images.

Figure 12: Examples of original images of the BIOSCAN-5M dataset, along with their respective total number of pixels (size) that occupy the image. The top row shows original images and the bottom row shows masks.

M Dataset category distribution

Figure 13 illustrates the taxonomic `class` distribution within the rank `order`. For example, of the 99.9% of organisms labelled at the `class` level, approximately 71% are classified within the `order` *Diptera* of the `class` *Insecta*.

For detailed insights into the class distribution within the major categories of the BIOSCAN-5M dataset, Table 13 presents comprehensive statistics. This table provides the total number of categories

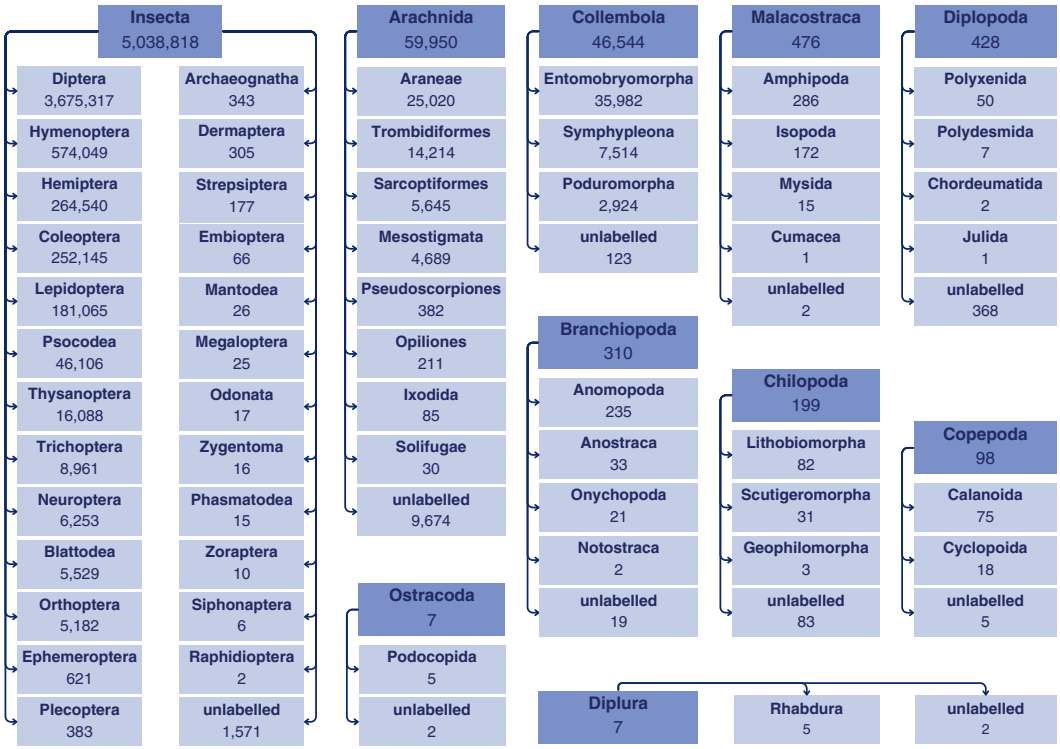


Figure 13: Distribution of taxonomic ranks in the BIOSCAN-5M dataset. Each darker cell represents a taxonomic `class`, while the lighter cells within each `class` represent the corresponding taxonomic `orders`. The numbers indicate the records belonging to each `class` and `order`. The *unlabeled* category denotes records assigned to a `class` but not to any specific `order`.

across 7 taxonomic group levels and BINs, highlighting both the most and least densely populated ones. Additionally, it includes calculated means, medians, and standard deviations of the population vectors of all subcategories of each attribute.

Table 13: BIOSCAN-5M taxonomic and BIN categories distribution. For each attribute, we show the value which occurs most often in the dataset and the least populated value (in the event of a tie, we show an exemplar selected at random).

		Most populated		Least populated				
Attributes	**Categories**	**Name**	**Size**	**Name**	**Size**	**Mean**	**Median**	**Std. Dev.**
`phylum`	1	Arthropoda	5,150,850	Arthropoda	5,150,850.0			
`class`	10	Insecta	5,038,818	Ostracoda	7	514,683.7	369.0	1,508,192.8
`order`	55	Diptera	3,675,317	Cumacea	1	93,363.4	172.0	495,969.5
`family`	934	Cecidomyiidae	938,928	Pyrgodesmidae	1	5,281.3	63.5	45,321.1
`subfamily`	1,542	Metopininae	323,146	Bombyliinae	1	953.7	23.0	9,092.8
`genus`	7,605	Megaselia	200,268	chalMalaise9590	1	161.3	6.0	2,492.2
`species`	22,622	Psychoda sp. 11GMK	7,694	Microcephalops sp. China3	1	20.9	2.0	139.5
`dna_bin`	324,411	BOLD:AEO1530	35,458	BOLD:ADT1070	1	15.8	2.0	146.4

N DNA barcode statistics

This section presents the DNA barcode statistics and analysis for the BIOSCAN-5M dataset. We provide several different statistics to show how the diversity of DNA barcodes varies across the different taxonomic levels. In Table 14, we report the number of distinct barcodes, the Shannon diversity index (e.g. entropy), and the average pairwise distances between barcodes at different taxonomic ranks. The different analysis all show that at higher levels of taxa, there are more distinct barcodes, and that at the genus and species level, the lexical distance between different barcodes are

much smaller than at the higher levels of taxa. Below we provide more details on how these statistics are computed.

N.1 Identical DNA barcodes: Statistical insights from the BIOSCAN-5M dataset

We compute and show in Table 14 the statistics for identical DNA barcode sequences across taxonomic ranks, including the total number of distinct barcodes per rank, as well as the average, median, and standard deviation of barcodes counts across subgroups within each rank.

Based on the statistics in Table 14, the total number of identical DNA barcode sequences within each subgroup of a specific taxonomic rank is lower than the total number of DNA sequences corresponding to the labelled samples in that subgroup. This indicates that some samples share identical DNA barcodes, possibly due to sequencing limitations. Since DNA barcodes are merely short snippets, they alone do not fully capture the unique genetic characteristics of individual samples.

Table 14: The DNA barcode statistics for various taxonomic ranks in the BIOSCAN-5M dataset. We indicate the total number of unique barcodes for the samples labelled to a given rank, and the mean, median, and standard deviation of the number of unique barcodes within the subgroupings at that rank. We also show the average across subgroups of the Shannon Diversity Index (SDI) for the DNA barcodes, measured in bits. We report the mean and standard deviation of pairwise DNA barcode sequence distances, aggregated across subgroups for each taxonomic rank.

		Unique Barcodes					Pairwise Distance	
Attributes	**Categories**	**Total**	**Mean**	**Median**	**Std. Dev.**	**Avg SDI**	**Mean**	**Std. Dev.**
`phylum`	1	2,486,492				19.78	158	42
`class`	10	2,482,891	248,289	177	725,237	8.56	166	103
`order`	55	2,474,855	44,997	57	225,098	7.05	128	53
`family`	934	2,321,301	2,485	46	19,701	5.42	90	46
`subfamily`	1,542	657,639	426	17	3,726	4.28	78	51
`genus`	7,605	531,109	70	5	1,061	2.63	50	39
`species`	22,622	202,260	9	2	37	1.46	17	18

N.2 Analyzing genetic diversity with the Shannon Index

Shannon Diversity Index (SDI). The Shannon Diversity Index (SDI) (Shannon, 1948), which measures the entropy within a group, is an effective metric for measuring genetic diversity as it considers both barcode richness (the number of distinct barcodes) and evenness (the distribution of samples among those barcodes). A high prevalence of identical barcodes leads to lower evenness and, consequently, a reduced SDI, indicating limited diversity and redundancy in genetic makeup.

Incorporating duplicated barcodes allows the SDI to capture the prevalence of specific barcodes within the subgroup. If certain barcodes are common across samples, the index may reflect a dominant genetic signature, resulting in a lower SDI and suggesting reduced diversity. Conversely, a greater presence of distinct barcodes with even distributions yields a higher SDI, indicating a more diverse subgroup structure. This dual focus on richness and evenness underscores the SDI's value in assessing genetic diversity and elucidating the genetic relationships within a subgroup.

We compute the Shannon Diversity Index (SDI) for each subgroup, T, within a taxonomic rank as

$$\text{SDI}_T = -\sum_{i=1}^{N} p_i \log_2(p_i), \tag{1}$$

where N is the number of unique DNA barcodes within a subgroup, and p_i is the fraction of samples in subgroup T which have the i-th barcode.

In Table 14, we report the average SDI (Avg SDI) for each taxonomic rank by computing SDI_T for each subgroup and then averaging these values across all subgroups within the respective rank. From the Table 14, the Avg SDI values indicate a high level of biodiversity at the `phylum` (19.78)

and class (8.56) levels, suggesting a rich community with a wide variety of taxa. However, as we move down the taxonomic hierarchy, the index values decline significantly, reaching the lowest point at the species level (1.46). This pattern suggests that while there is a diverse range of phyla and classes, the distribution of species within these groups is uneven, indicating the presence of a few dominant species or genera.

N.3 Pairwise distance analysis of identical DNA barcodes

Damerau-Levenshtein Distance. The Damerau-Levenshtein distance (Damerau, 1964) is a string-edit distance metric that measures the minimum number of operations required to transform one string into another. It is an extension of the standard Levenshtein distance (Levenshtein, 1966), which counts the number of single-character edits needed for transformation. The key difference is that the Damerau-Levenshtein distance also accounts for transpositions, i.e., when two adjacent characters are swapped. In the context of our DNA barcoding, it measures how similar or different two DNA sequences are by counting how many single-character changes (insertions, deletions, substitutions, or transpositions) are needed to make one sequence identical to another.

We report the average Damerau-Levenshtein pairwise distance between unique DNA barcodes at different taxonomic ranks in Table 14. To compute the statistics for the pairwise distances, we take each subgroup at every taxonomic rank, and only consider subgroups with sufficient number of distinct barcodes. For a given subgroup, if the total number of unique DNA barcode sequences is fewer than 4, the subgroup is not considered. If the total exceeds 1,000, up to 1,000 sequences are randomly sampled; otherwise, all sequences are included.

To compute the distances between barcodes, the sampled DNA barcode sequences are first aligned using the MAFFT alignment technique (Katoh & Standley, 2013). Next, the pairwise distances between aligned DNA barcodes are computed using the Damerau-Levenshtein metric, with a total of $n \times \frac{(n-1)}{2}$ comparisons (where n is the number of DNA barcodes). The mean and standard deviation of these distance values are then computed within each subgroup and subsequently aggregated using the mean function across subgroups at each taxonomic rank.

The statistics (Table 14, right columns) indicate that as we progress from higher to lower taxonomic ranks (e.g., from phylum and class to genus and species), both the mean and standard deviation of pairwise genetic distances decrease. This reduction indicates that genetic differences between organisms become smaller as we move down the taxonomic hierarchy, meaning organisms at lower ranks are more genetically similar to each other compared to those at higher ranks. For instance, species within the same genus tend to have much more similar DNA sequences than families within an order or orders within a class. This pattern aligns with the hierarchical structure of biological classification, where organisms are grouped based on increasing genetic relatedness as we move to finer taxonomic levels.

At the same time, the larger standard deviations observed at higher taxonomic ranks, such as class and order, reflect greater variability in genetic distances, suggesting a broader range of genetic diversity at these levels. Conversely, at lower ranks, such as genus and species, the smaller mean and standard deviation of pairwise distances highlight closer genetic relationships. However, these reduced distances can pose challenges for classification since the differences between closely related species become subtle.

This emphasizes the need for finer genetic markers or additional traits beyond pairwise distances to accurately distinguish between organisms, especially at the species level, where genetic distinctions can be minimal. Incorporating multimodal data, such as combining DNA sequences with images, can help address this challenge by providing complementary information. While DNA sequences offer insights into genetic differences, images capture morphological traits that may not be reflected in the genetic data. This multimodal approach can enhance classification accuracy, particularly when distinguishing between closely related species.

Figure 14, Figure 15 and Figure 16 provide a visual representation of the statistics of pairwise distances computed in Table 14 for taxonomic ranks class, order, and species, respectively. The Interquartile Range (IQR) is a measure of statistical dispersion that describes the range within which the central 50% of the pairwise distances lies. It is calculated as the difference between the third

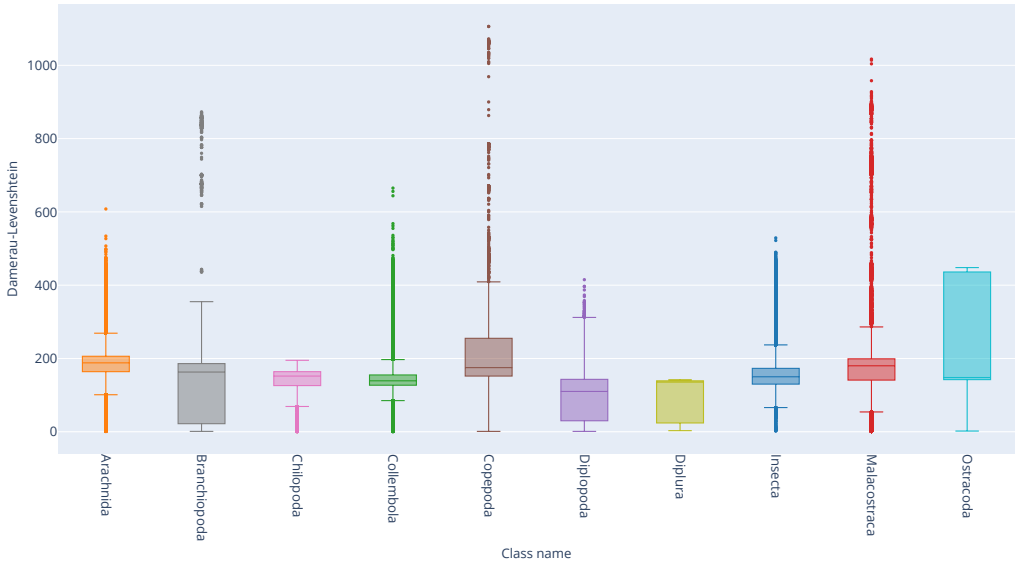


Figure 14: **Distribution of pairwise distances of subgroups of class.** The x-axis shows the subgroup categories sorted alphabetically.

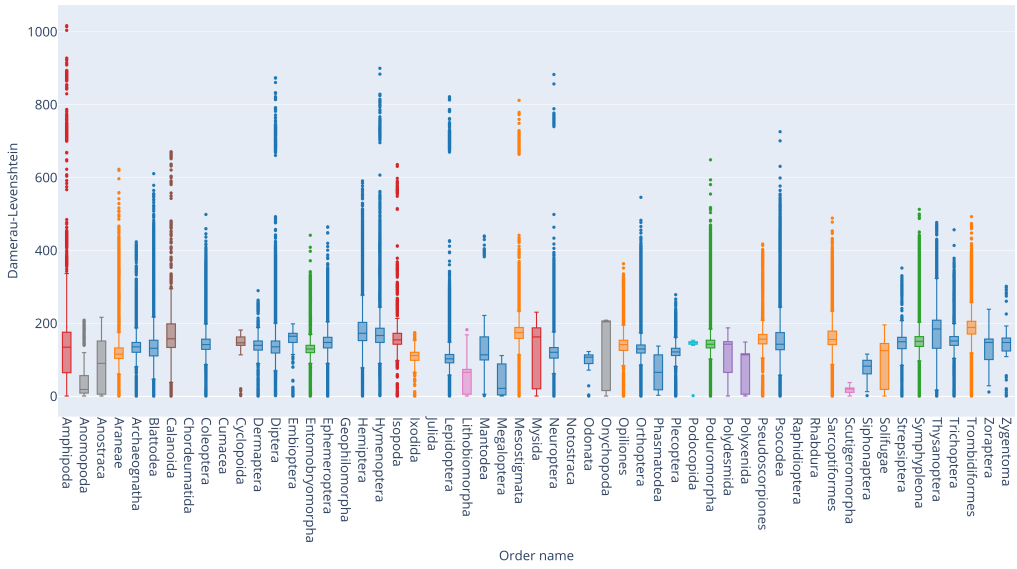


Figure 15: **Distribution of pairwise distances of subgroups of order.** The x-axis shows the subgroup categories sorted alphabetically.

quartile (Q_3) and the first quartile (Q_1) of the data,

$$\text{IQR} = Q_3 - Q_1,$$

where Q_1 is the 25th percentile of the data, and Q_3 is the 75th percentile. The line inside the box represents the median (Q_2) of the data. The height of the box illustrates the IQR. The lines extending from the box (whiskers) indicate the range of the data outside the IQR, typically extending up to 1.5 times the IQR from the quartiles, which help identify the spread of the data.

A small IQR (e.g., Collemboda in Figure 14) indicates that the pairwise distances among DNA barcode sequences within the group are tightly clustered around the median, suggesting that the

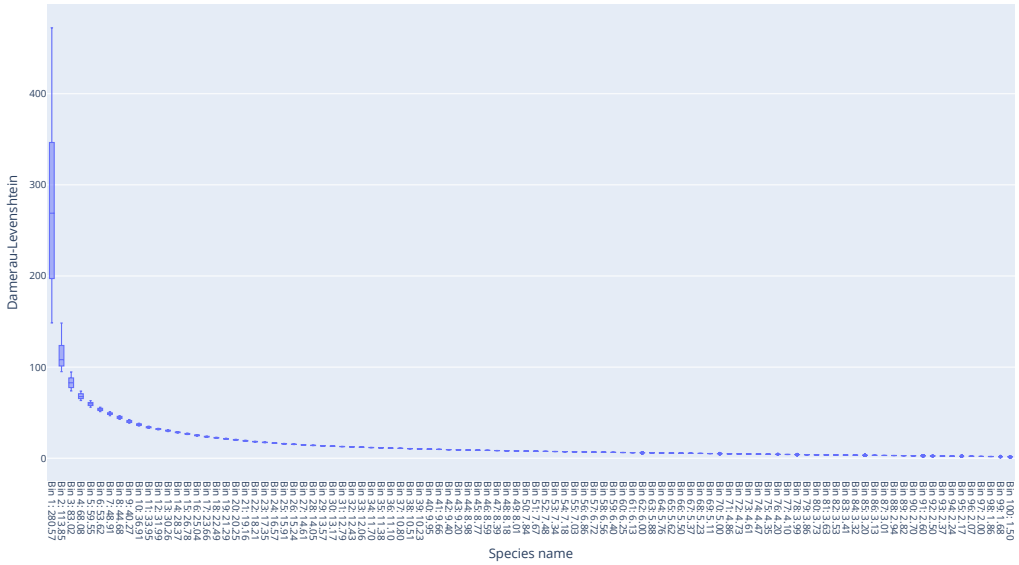


Figure 16: **Distribution of pairwise distances of subgroups of species.** Among the species, there are 8,372 distinct subgroups with sufficient identical barcodes for calculating pairwise distances, which makes visualization challenging. To address this, the groups are sorted in descending order based on their mean distances and partitioned into 100 bins. These bins are used to plot the distribution of pairwise distances within the species rank. The mean distance of each bin is displayed along the x-axis.

sequences are similar to one another. This homogeneity may imply that the groups consist of closely related species or individuals with minimal genetic divergence, possibly due to a recent common ancestor.

Conversely, a large IQR (e.g., Ostracoda in Figure 14) signifies significant variability in the pairwise distances among sequences within a group, indicating a wider range of genetic diversity. This heterogeneity suggests that the groups may encompass genetically diverse species or populations with notable evolutionary divergence. Additionally, the presence of a large IQR may point to potential outliers—sequences that differ substantially from the majority—which could warrant further investigation to understand the underlying genetic variations.

If the whiskers are long while the IQR is small (e.g., Malacostraca in Figure 14), it implies that there are outlier values or a wider distribution of data points beyond the central cluster, highlighting the presence of variability in the dataset that may be worth investigating further.

If the median Q_2 is closer to Q_1 (e.g., Copepoda in Figure 14), the distribution is positively skewed, with most data points concentrated at the lower end and fewer but larger values at the higher end. Conversely, if the median is closer to Q_3 (e.g., Branchiopoda in Figure 14), the distribution is negatively skewed, with more values at the higher end and fewer, smaller values at the lower end.

Note that in all taxonomic ranks except for `species`, a random selection of 1,000 records is made for subgroups with more than 1,000 samples. For the `species` rank, all subgroups with a large number of records are included in the pairwise distance calculations. Some taxonomic ranks contain extremely large subgroups, such as *Arthropoda* in `phylum` and *Insecta* in `class`, each with over 2 million unique DNA records. Consequently, the 1,000 selected records may not fully represent the pairwise distances within the large subgroups. Due to computational limitations—since 1,000 records result in about 500 k unique pairwise distance computation—we adhere to this rule of selecting a random subset of 1,000 records.

O Insect vs non-insect organisms

Focusing on *Insecta* as the most populous group at the `class` level, we present its detailed statistical records for DNA, BIN, and various taxonomic ranks in Table 15.

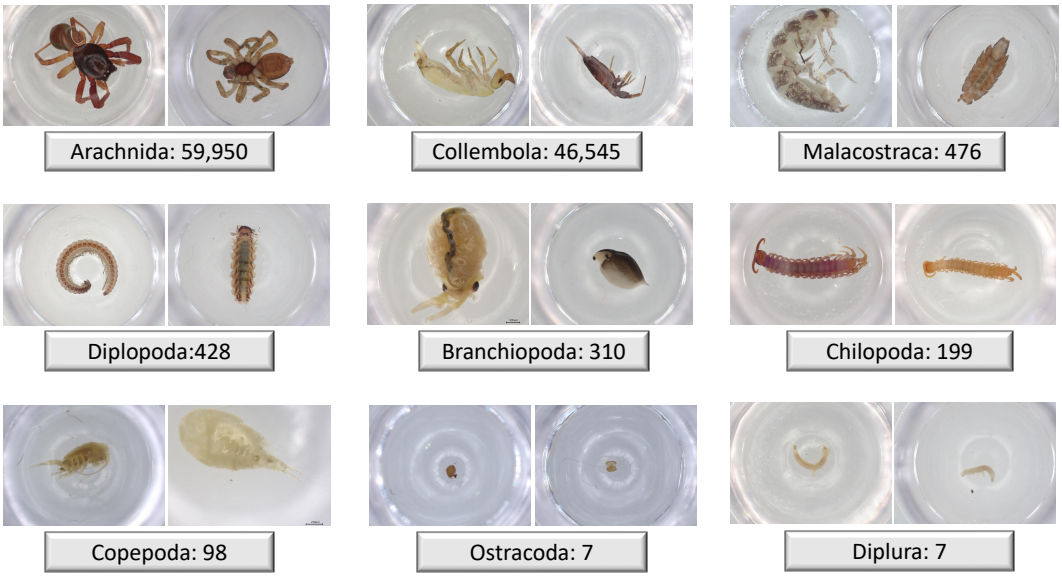


Figure 17: Examples of original images of non-insect organisms from the BIOSCAN-5M dataset. Below each image, the `class` name and its population within the BIOSCAN-5M dataset are displayed.

Figure 13 shows the class distribution within the taxonomic rank `class`, with 99.9% of organisms labelled at this level, of which 97.8% belong to the `class` *Insecta*. Figure 17 displays original images of non-insect taxonomic classes from the BIOSCAN-5M dataset, which includes a total of 137,479 organisms.

Table 15: Detailed statistical records for DNA, BIN and taxonomic ranks within `class` *Insecta* of the BIOSCAN-5M dataset.

Attributes	Categories	Labelled	Labelled (%)	Unlabelled	Unlabelled (%)	IR
`order`	25	5,037,247	99.97	1,571	0.03	1,837,658
`family`	681	4,853,383	96.32	185,435	3.68	938,928
`subfamily`	1,305	1,431,962	28.42	3,606,856	71.58	323,146
`genus`	6,897	1,188,043	23.58	3,850,775	76.42	200,268
`species`	21,512	450,215	8.93	4,588,603	91.07	7,694
`taxon`	26,603	5,038,818	100.00	0	0.00	925,520
`dna_bin`	311,743	5,025,921	99.74	12,897	0.26	35,458
`dna_barcode`	2,423,704	5,038,818	100.00	0	0.00	3,743

P Limitations and challenges

P.1 Fine-grained classification

The BIOSCAN-5M dataset offers detailed biological features for each organism by annotating images with multi-grained taxonomic ranks. The class imbalance ratio (IR) across taxonomic groups reveals significant disparities in sample sizes between the majority class (with the most samples) and minority classes (with fewer samples). Notably, among the 55 distinct `orders`, *Diptera* accounts for approximately 71% of the total organisms. Figure 18 illustrates various `species` within the `order`

Diptera, highlighting the high similarity among images of distinct categories, which poses additional challenges for downstream image classification tasks.

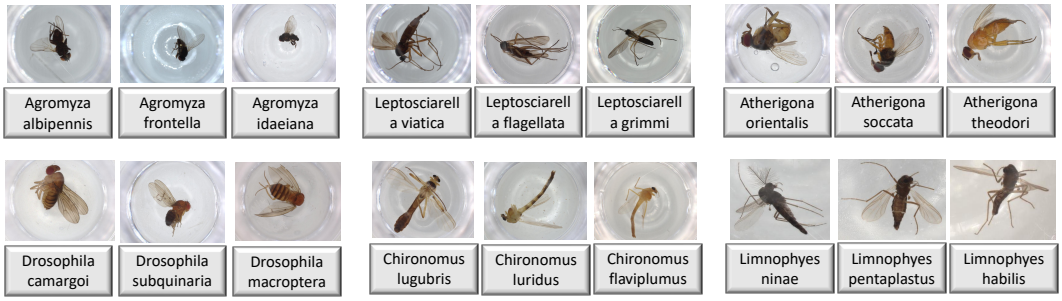


Figure 18: Sample images of distinct `species` from the `order` *Diptera*, which comprises about 71% of BIOSCAN-5M dataset. High similarity between samples of different `species` highlights significant image classification challenges.

P.2 Accessing ground-truth labels

The BIOSCAN-5M dataset exposes a limitation regarding labelling. The number of labelled records sharply declines as we delve deeper into taxonomic ranking groups, particularly when moving towards finer-grained taxonomic ranks beyond the `family` level. In fact, over 80% of the organisms lack taxonomic labels for ranks such as `subfamily`, `genus` and `species`. This circumstance poses a significant challenge for conducting taxonomic classification tasks. However, this limitation also opens doors to opportunities for research in various domains. The abundance of unlabelled data presents avenues for exploration in clustering, unsupervised, semi-supervised, and self-supervised learning paradigms, allowing for innovative approaches to data analysis and knowledge discovery.

P.3 Sampling Bias

The BIOSCAN-5M dataset also exposes a sampling bias as a result of the locations where and the methods through which organisms were collected, as depicted by Figure 11.

Q Data processing

To optimize our benchmark experiments using the BIOSCAN-5M dataset, we implemented two critical pre-processing steps on the raw dataset samples. These steps were necessary to enhance the efficiency and accuracy of our downstream tasks.

The first step involved image cropping and resizing. Due to the high resolution and large size of images in the dataset, processing the original images is both time-consuming and computationally expensive. Additionally, the area around the organism in each image is redundant for our feature extraction. To address these issues, we cropped the images to focus on the region of interest, specifically the area containing the organism. This step eliminated unnecessary background, reducing the data size and focusing the analysis on the relevant parts of the images. After cropping, we resized the images to a standardized resolution, further reducing the computational load and ensuring uniformity across all image samples.

The second step addressed inconsistencies in the taxonomic labels. In the raw dataset, we encountered identical DNA nucleotide sequences labelled differently at certain taxonomic levels, likely due to human error (e.g., typos) or disagreements in taxonomic naming conventions. Such discrepancies posed significant challenges for our classification tasks involving images and DNA barcodes. To address this, we implemented a multi-step cleaning process for the taxonomic labels. We identified and flagged inconsistent labels associated with identical DNA sequences and corrected typographical errors to ensure accurate and consistent naming.

We present additional details of our pre-processing steps in the following section.

Q.1 Image processing details

The BIOSCAN-5M dataset contains resized and cropped images following the process in BIOSCAN-1M Insect (Gharaee et al., 2023). We resized images to 256 px on the smaller dimension. As in BIOSCAN-1M, we opt to conduct experiments on the cropped and resized images due to their smaller size, facilitating efficient data loading from disk.

Cropping. Following BIOSCAN-1M (Gharaee et al., 2023), we develop our cropping tool by fine-tuning a DETR (Carion et al., 2020) model with a ResNet-50 (He et al., 2016b) backbone (pretrained on MSCOCO, Lin et al., 2014) on a small set of 2,837 insect images annotated using the Toronto Annotation Suite[4].

For BIOSCAN-1M, the DETR model was fine-tuned using 2,000 insect images (see Section 4.2 of Gharaee et al., 2023 for details). While the BIOSCAN-1M cropping tool worked well in general, there are some images for which the cropping was poor. Thus, we took the BIOSCAN-1M cropping tool checkpoint, and further fine-tuned the model for BIOSCAN-5M using the same 2,000 images and an additional 837 images that were not well-cropped previously. We followed the same training setup and hyperparameter settings as in BIOSCAN-1M and fine-tuned DETR on one RTX2080 Ti with batch size 8 and a learning rate of 0.0001.

Table 16: We compare the performance of the DETR model we used for cropping that was trained with the extra 837 images (NWC-837) that were previously not well-cropped to the model used for BIOSCAN-1M. We report the Average Precision (AP) and Average Recall (AR) computed on an additional validation set consisting of 100 images that were not-well cropped previously (NWC-100-VAL), as well as the images (IP-100-VAL + IW-150-VAL) used to evaluate the cropping tool's model used in BIOSCAN-1M. Our updated model performs considerably better on NWC-100-VAL, while given comparable performance on the original validation set of images.

		NWC-100-VAL		IP-100-VAL + IW-150-VAL	
Dataset	Training data	AP[0.75]	AR[0.50:0.95]	AP[0.75]	AR[0.50:0.95]
BIOSCAN-1M	IP-1000 + IW-1000	0.257	0.485	**0.922**	**0.894**
BIOSCAN-5M	IP-1000 + IW-1000 + NWC-837	**0.477**	**0.583**	0.890	0.886

Table 16 shows that our model with additional data achieves better cropping performance on an evaluation set of 100 images that were previously poorly cropped (NWC-100-VAL). Before cropping, we increase the size of the predicted bounding box by a fixed ratio $R = 1.4$ relative to the tight bounding box to capture some of the image background. If the bounding box extends beyond the image's edge, we pad the image with maximum-intensity pixels to align with the white background. These processes are the same as used by Gharaee et al. (2023). After cropping, we save the cropped-out bounding box.

Resizing. After cropping the image, we resize the image to 256 pixels on its smaller side while maintaining the aspect ratio ($r = \frac{w}{h}$). As nearly all original images are 1024×768 pixels, our resized images are (nearly all) 341×256 pixels.

Area fraction. The `area_fraction` field in the metadata file indicates the proportion of the original image represented by the cropped image. This factor is calculated using the bounding box information predicted by our cropping tool and serves as an indicator of the organism's size. Figure 19 displays the bounding boxes detected by our cropping model, which we used to crop images in the BIOSCAN-5M dataset. The area fraction factor is calculated as follows:

$$f_a = \frac{w_c\, h_c}{w\, h} \tag{2}$$

Scale factor. When capturing images of physical objects, such as medical scans or biological samples, it is essential to ensure that measurements derived from these images accurately represent the real objects. To compute real-world sizes from captured images, a consistent relationship between pixel size and physical size is necessary. Therefore, we introduced the `scale_factor` field in the metadata

[4] https://aidemos.cs.toronto.edu/annotation-suite/

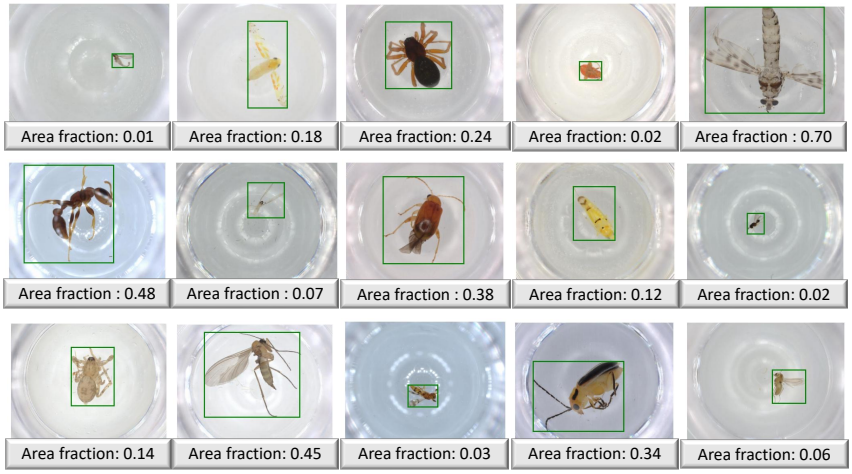


Figure 19: Examples of original images of organisms of the BIOSCAN-5M dataset with the bounding boxes detected by our cropping module. The area fraction value below each image shows how much of the original image is included in the crop.

file, which defines the ratio between the cropped image (`cropped`) and the cropped and resized image (`cropped_256`).

Assuming the original images (I) have constant dimensions, width (w) and height (h), the cropped images (I_c) are extracted using bounding box information from our cropping tool and have varying widths (w_c) and heights (h_c) proportional to the size of the organism. The resized images (I_r) are adjusted so that the shorter dimension, either width (w_r) or height (h_r), is set to a constant size of 256 px, while the other dimension is scaled proportionally to maintain the aspect ratio, resulting in a dimension greater than 256 px.

We calculated the scale-factor (f_s) as follows:

$$f_s = \frac{\min(w_c, h_c)}{256} \tag{3}$$

If we define the pixel scale as the number of millimeters per pixel, then the pixel scale of the cropped and resized image (`cropped_256`) is equivalent to the pixel scale of the original image multiplied by the scale factor:

$$\text{pixel\_scale}_{\text{cropped\_256}} = \text{pixel\_scale}_{\text{original}} \times f_s \tag{4}$$

Note the pixel scale of the original image remains unchanged during the cropping process, as cropping only involves cutting out areas around the region of interest (the organism) without scaling the image.

The original images were captured using a Keyence VHX-7000 Digital Microscope system imaging system at a resolution of 2880×2160 pixels. These images were then resized to a resolution of 1024×768 pixels to obtain the original images (`original_full`) of the BIOSCAN-5M dataset. Each pixel in the raw images represents a physical space of 2.95 µm by 2.95 µm. Using this pixel scale and the scale factor obtained from Equation 4, we can estimate the size of the object in the real-world.

Q.2 HDF5 file

To load data efficiently during the training of the CLIBD baseline, we also generated a 190 GB HDF5 file to store images and related metadata from the BIOSCAN-5M dataset. This file is structured to allow rapid access and processing of large-scale data.

At the top level, the file consists of a *group* of the following *datasets* representing different partitions of BIOSCAN-5M. Each partition includes keys or queries for one or all of the splits (pretraining, validation, or test).

For more information or to download the HDF5 file, the instructions are found at the CLIBD GitHub repo: https://github.com/bioscan-ml/clibd.

Q.3 Taxa of unassigned placement

Some taxonomic labels had "holes" in them due to the complexities of the definition of taxonomic labels. Established taxonomic labels for some species can omit taxonomic ranks because there is currently no scientific need to define a grouping at that taxonomic level.

In particular, we found there were 1,448 genera which were missing a subfamily label because their genus had not been grouped into a subfamily by the entomological community. Note that these genera might at some point in the future be assigned a subfamily, if a grouping of genera within the same family becomes apparent.

This situation of mixed rankings creates a complexity for hierarchical modelling, which for simplicity typically assumes a rigid structure of level across the labelling tree for each sample. We standardized this by adding a placeholder subfamily name where there was a hole, equal to "unassigned <Family_name>". For example, for the genus Alpinosciara, the taxonomic label was originally:

```
Arthropoda > Insecta > Diptera > Sciaridae > [none] > Alpinosciara
```

and after filling the missing subfamily label, it became:

```
Arthropoda > Insecta > Diptera > Sciaridae > unassigned Sciaridae > Alpinosciara
```

This addition ensures that the mapping from genus to subfamily is injective, and labels which are missing because they are not taxonomically defined are not confused with labels which are missing because they have not been identified. Furthermore, this ensures that each subsequent rank in the taxonomic labels provides a partitioning of each of the labels in the rank that proceeds it.

Q.4 Taxonomic label cleaning

The taxonomic labels were originally entered into the BOLD database by expert entomologists using a drop-down menu for existing species, and typed-in manually for novel species. Manual data entry can sometimes go awry. We were able to detect and resolve some typographical errors in the manual annotations, as described below.

Genus and species name comparison. Since species names take the form "<Genus_name> <species_specifier>", the genus is recorded twice in samples which possess species labels. This redundancy provides an opportunity to provide a level of quality assurance on the genus-level annotations A few samples (82 samples across 13 species) had a species label but no genus label; for these we used the first word of the species label as the genus label. For the rest of the samples with a species label, we compared the first word of the species label with the genus label, and resolved 166 species where these were inconsistent. These corrections also uncovered cases where the genus name was entered incorrectly more broadly, and we were able to correct genera values which were entered incorrectly even in cases where they were consistent with their species labels or had no species labels.

Conflicted annotations for the same barcode. We found many DNA barcodes were repeated across the dataset, with multiple images bearing the same barcode. Overall, there were on average around two repetitions per unique barcode in the dataset. It is already well-established that the COI mitochondrial DNA barcode is a (sub)species-level identifier, i.e. same barcode implies same species, and different species implies different barcodes (Moritz & Cicero, 2004; Sokal & Sneath, 1963; Blaxter et al., 2005). Hence we have a strong prior that samples with the same barcode should be samples of the same species. This presents another opportunity to provide quality assurance on the data, by comparing the taxonomic annotations across samples which shared a DNA barcode. Differences can either arise by typographical errors during data entry, or by differences of opinion between annotators.

We investigated cases where completed levels of the taxonomic annotations differed for the same barcode. This indicated some common trends as values often compared as different due to stylistic differences, where one annotation differed only by casing, white-space, the absence of a 0 padding digit to an identifier code number, or otherwise misspellings. We resolved some such disagreements automatically, by using the version more common across the dataset.

The majority of placeholder genus and species names follow one of a couple of formats such as "<Genus_name> Malaise1234", e.g. "*Oxysarcodexia Malaise4749*". Comparing different taxonomic

annotations of the same barcode only allows us to find typos where a barcode has been annotated more than once. However, there are of course more barcodes than species and so there may remain some typos which make two samples of the same species with different barcodes compare as different when they should be the same. To address this, we found labels which deviated from the standardized placeholder name formats and modified them to fit the standardized format. Examples of these corrections include adding missing zero-padding on digits, fixing typos of the word "Malaise", and inconsistent casing. In this way, we renamed the species of 6,756 samples and genus of 3,675 across 7,673 records.

We resolved the remaining conflicts between differently annotated samples of the same barcode as follows. We considered each taxonomic rank one at a time. In cases where there was a conflict between the annotations, we accepted the majority value if at least 90% of the annotations were the same. If the most common annotation was less prevalent than this, we curtailed the annotation at the preceding rank. Curtailed annotations which ended at a filler value (i.e. a subfamily name of the format "unassigned <Family_name>") were curtailed at the last completed rank instead. In total, we dropped at least one label from 3,478 records.

Next, we considered barcodes whose multiple annotations differed in their granularity. In such cases, we inferred the annotations for missing taxonomic ranks from the samples that were labelled to a greater degree of detail. In total, we inferred at least one label for 172,895 records. We believe these inferred labels are unlikely to have an error rate notably higher than that of the rest of the data. Even so, we provide details about which ranks were inferred in the metadata field `inferred_ranks` in case the user wishes to exclude the inferred labels. This field takes the following values:

- 0 — Original label only (nothing inferred).
- 1 — Species label was copied. (Sample was originally labelled to genus-level.)
- 2 — Genus and (if present) species labels were copied.
- 3 — Subfamily, and every rank beneath it, were copied.
- 4 — Family, and every rank beneath it, were copied.
- 5 — Order, and every rank beneath it, were copied.
- 6 — Class, and every rank beneath it, were copied.

Non-uniquely identifying species names. Finally, we noted that some species names were not unique identifiers for a species. Theses cases arise where an annotator has used *open nomenclature* to indicate a suspected new species, e.g. "*Pseudosciara sp.*", "*Olixon cf. testaceum*", and "*Dacnusa nr. faeroeensis*". Since this is not a uniquely identifying placeholder name for the species, it is unclear whether two instances with the same label are the same new species or different new species. For example, there were 1,247 samples labelled as "*Pseudosciara sp.*", and these will represent a range of new species within the Pseudosciara genus, and not repeated observations of the same new species. Consequently, we removed such species annotations which did not provide a unique identifier for the species. In total, 198 such species values were removed from 5,101 samples.

Conclusion. As a result of this cleaning process we can make the following claims about the dataset, with a high degree of confidence:

- All records with the same barcode have the same annotations across the taxonomic hierarchy.
- If two samples possess a species annotation and their species annotation is the same, they are the same species. (Similarly for genus level annotations, etc.)
- If two samples possess a species annotation and their species annotations differ, they are not the same species. (Similarly for genus level annotations, etc.)

R Dataset partitioning — Additional details

R.1 Species sets

As summarized in §4.1, we first partitioned the data based on their species label into four categories as follows:

Table 17: Example species from each species set.

			Number of samples			
Species set	**Genus**	**Species**	**All**	**Train/ Keys**	**Val**	**Test**
seen	Aacanthocnema	Aacanthocnema dobsoni	3	3	0	0
	Glyptapanteles	Glyptapanteles meganmiltonae	65	45	2	18
	Megaselia	Megaselia lucifrons	699	640	34	25
	Pseudomyrmex	Pseudomyrmex simplex	378	335	18	25
	Rhopalosiphoninus	Rhopalosiphoninus latysiphon	148	116	7	25
	Stenoptilodes	Stenoptilodes brevipennis	16	10	1	5
	Zyras	Zyras perdecoratus	10	6	0	4
unseen	Anastatus	Anastatus sp. GG28	42	24	6	12
	Aristotelia	Aristotelia BioLep531	87	51	13	23
	Glyptapanteles	Glyptapanteles Whitfield155	11	6	1	4
	Megaselia	Megaselia BOLD:ACN5814	24	13	3	8
	Orthocentrus	Orthocentrus Malaise5315	39	23	5	11
	Phytomyptera	Phytomyptera Janzen3550	14	8	1	5
	Zatypota	Zatypota alborhombartaDHJ03	9	8	1	0
heldout	Basileunculus	Basileunculus sp. CR3	268			
	Cryptophilus	Cryptophilus sp. SAEVG Morph0281	55			
	Glyptapanteles	Glyptapanteles Malaise2871	1			
	Odontofroggatia	Odontofroggatia corneri-MIC	13			
	Palmistichus	Palmistichus ixtlilxochitliDHJ01	416			
	gelBioLep01	gelBioLep01 BioLep3792	16			
	microMalaise01	microMalaise01 Malaise1237	13			

- *Unknown*: samples without a species label (note: these may truly belong in any of the other three categories).
- *Seen*: all samples whose species label is an established scientific name of a species. Species which did not begin with a lower case letter, contain a period, contain numerals, or contain "malaise" (case insensitive) were determined to be labelled with a catalogued, scientific name for their species, and were placed in the *seen* set.
- *Unseen*: Of the remaining samples, we considered the placeholder species which we were most confident were labelled reliably. These were species outside the seen species, but the genus occurred in the seen set. Species which satisfied this property and had at least 8 samples were placed in the *unseen* set.
- *Heldout*: The remaining species were placed in *heldout*. The majority of these have a placeholder genus name as well as a placeholder species name, but some have a scientific name for their genus name.

This partitioning ensures that the task that is posed by the dataset is well aligned with the task that is faced in the real-world when categorizing insect samples. Example species for each species set are shown in Table 17, and the number of categories for each taxonomic rank are shown in Table 18.

Table 18: Number of (non-empty) categories for each taxa, per species set.

Species set	**Phylum**	**Class**	**Order**	**Family**	**Subfamily**	**Genus**	**Species**
unknown	1	10	52	869	1,235	4,260	0
seen	1	9	42	606	1,147	4,930	11,846
unseen	1	3	11	64	118	244	914
heldout	1	4	22	188	381	1,566	9,862
overall	1	10	55	934	1,542	7,605	22,622

R.2 Splits

To construct partitions appropriate for a closed world training and evaluation scenario, we partitioned the seen data into `train`, `val`, and `test` partitions. Because many of the DNA barcodes have more than one sample (i.e. multiple images per barcode), we partitioned the data at the barcode level. The data was highly imbalanced, so to ensure the `test` partition had high sample efficiency, we flattened the distribution for the `test` set. For each species with at least 2 barcodes and at least 8 samples, we selected barcodes to place in the `test` set. We tried to place a number of samples in the `test` set which scaled linearly with the number of samples for the species, starting with a minimum of 4, and capped at a maximum of 25 (reached at 92 samples total). The target increased at a rate of $1/4$. We capped the number of barcodes to place in the `test` set at a number that increased linearly with the number of barcodes for the species, starting at 1 and increasing at a rate of $1/3$. This flattened the distribution across species in the `test` set, as shown in Figures 20e, 21e, and 22e.

Table 19: Number of (non-empty) categories for each taxa, per partition.

Partition	Phylum	Class	Order	Family	Subfamily	Genus	Species
pretrain	1	10	52	869	1,235	4,260	0
train	1	9	42	606	1,147	4,930	11,846
val	1	5	27	350	598	1,704	3,378
test	1	6	27	352	594	1,736	3,483
key_unseen	1	3	11	64	118	244	914
val_unseen	1	3	11	62	116	240	903
test_unseen	1	3	11	62	113	234	880
other_heldout	1	4	22	188	381	1,566	9,862
overall	1	10	55	934	1,542	7,605	22,622

Table 20: Number of species in common between each pair of partitions.

	pretrain	train	val	test	key_unseen	val_unseen	test_unseen	other_heldout
pretrain	0	0	0	0	0	0	0	0
train	0	11,846	3,378	3,483	0	0	0	0
val	0	3,378	3,378	2,952	0	0	0	0
test	0	3,483	2,952	3,483	0	0	0	0
key_unseen	0	0	0	0	914	903	880	0
val_unseen	0	0	0	0	903	903	878	0
test_unseen	0	0	0	0	880	878	880	0
other_heldout	0	0	0	0	0	0	0	9,862

To evaluate model performance during model development cycles, we also created a validation partition (`val`) with the same distribution as the `test` set. This was partition was created to contain around 5% of the remaining samples from each of the seen species, by selecting barcodes to place in the `val` partition. To mimic the long tail of the distribution, for each species with fewer than 20 samples and at least 6 samples, and for which one of their barcodes had only a single image, we added one single-image barcode to the `val` partition. This step added 1,766 individual samples from the tail; for comparison, our target of 5% of the samples from the tail would be 1,955 samples.

The remaining barcodes with samples of seen species are placed in the `train` partition. For retrieval paradigms, we use the `train` partition as keys and the `val` and `test` partitions as queries.

For the unseen species, we use the same methodology as for the seen species to create and `val_unseen`, with the exception that the proportion of samples placed in the `val_unseen` partition was increased to 20% to ensure it is large enough to be useful. The remaining samples of unseen species are placed in the `keys_unseen` partition. For retrieval paradigms, we use the `keys_unseen`

Table 21: Fraction of species (%) in common between each pair of partitions, relative to the number of species for the row.

	pretrain	train	val	test	key_unseen	val_unseen	test_unseen	other_heldout
pretrain	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
train	0.0	100.0	28.5	29.4	0.0	0.0	0.0	0.0
val	0.0	100.0	100.0	87.4	0.0	0.0	0.0	0.0
test	0.0	100.0	84.8	100.0	0.0	0.0	0.0	0.0
key_unseen	0.0	0.0	0.0	0.0	100.0	98.8	96.3	0.0
val_unseen	0.0	0.0	0.0	0.0	100.0	100.0	97.2	0.0
test_unseen	0.0	0.0	0.0	0.0	100.0	99.8	100.0	0.0
other_heldout	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Table 22: Number of genera in common between each pair of partitions.

	pretrain	train	val	test	key_unseen	val_unseen	test_unseen	other_heldout
pretrain	4,260	2,372	1,190	1,206	217	214	209	682
train	2,372	4,930	1,704	1,736	244	240	234	519
val	1,190	1,704	1,704	1,517	151	148	145	266
test	1,206	1,736	1,517	1,736	157	154	151	276
key_unseen	217	244	151	157	244	240	234	177
val_unseen	214	240	148	154	240	240	232	175
test_unseen	209	234	145	151	234	232	234	172
other_heldout	682	519	266	276	177	175	172	1,566

partition as keys and the `val_unseen` and `test_unseen` partitions as queries. For open world evaluation, we train on the `test` partition, without presenting any samples from the unseen species during training, and evaluate on `test_unseen`.

The samples of heldout species are placed in the partition `other_heldout`. The utility of these species varies depending on the model paradigm. In particular, we note that as these species are in neither the seen nor unseen species, they can be used to train a novelty detector without the novelty detector being trained on unseen species.

The samples of unknown species are placed entirely in the pretrain partition, which can be used for self-supervised pretraining, or semi-supervised learning.

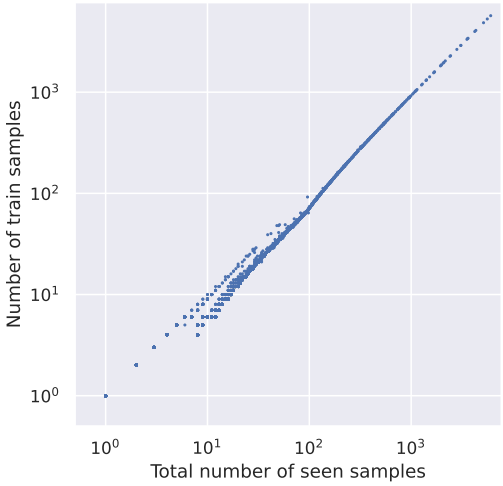
To aid comparison between the coverage of the partitions, we show the number of species in common between each pair of partitions (Table 20), and the percentage of species in common (Table 21). This is a block-diagonal matrix as species labels do not overlap between species sets. The `train` partition has higher diversity than the `val` and `test` partitions, which each cover less than 30% of the seen species. This is due to the long-tail of the distribution — of the 11,846 species, 7,919 species (two thirds) have 6 or fewer samples, and of these 3,756 species only have a single sample. However, these rare species only constituted a small fraction of the `train` samples—only 17,572 samples are members of species with 6 or fewer samples, which is 6% of the `train` partition. Due to our selection process for unseen species, in which only species with enough samples to be confident they are accurate are included, a much higher fraction of the unseen species are included in `val_unseen` and `test_unseen`.

Table 23: Fraction of genera (%) in common between each pair of partitions, relative to the number of genera for the row.

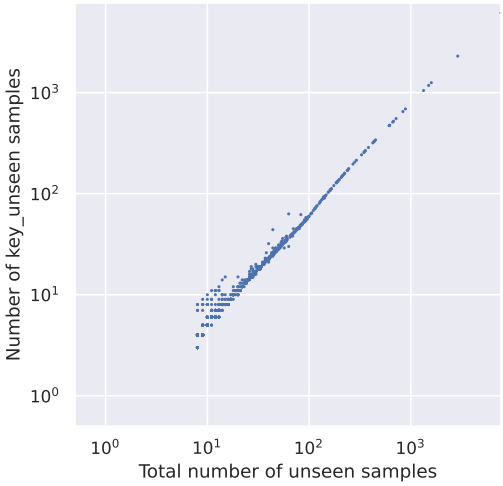
	pretrain	train	val	test	key_unseen	val_unseen	test_unseen	other_heldout
pretrain	100.0	55.7	27.9	28.3	5.1	5.0	4.9	16.0
train	48.1	100.0	34.6	35.2	4.9	4.9	4.7	10.5
val	69.8	100.0	100.0	89.0	8.9	8.7	8.5	15.6
test	69.5	100.0	87.4	100.0	9.0	8.9	8.7	15.9
key_unseen	88.9	100.0	61.9	64.3	100.0	98.4	95.9	72.5
val_unseen	89.2	100.0	61.7	64.2	100.0	100.0	96.7	72.9
test_unseen	89.3	100.0	62.0	64.5	100.0	99.1	100.0	73.5
other_heldout	43.6	33.2	17.0	17.6	11.3	11.2	11.0	100.0

Similarly, we show the number and percentage of genera in common between pairs of partitions (Table 22 and Table 23, respectively). We see that the genera across all seen and unseen species set partitions are contained in the train partition.

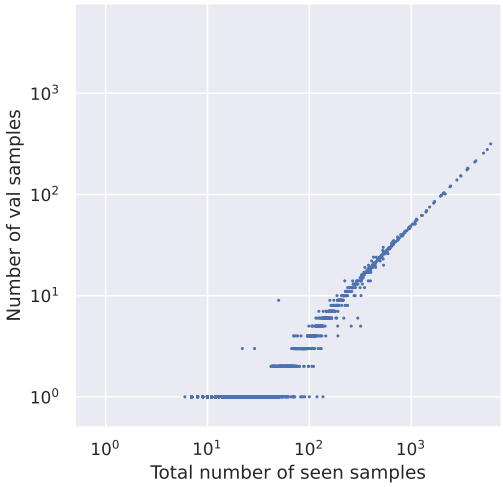
In Figure 23, we show the number of samples per partition. The plot illustrates the vast majority of the samples (91%) are in the pretrain partition, and most samples are only labelled to family level (67%).



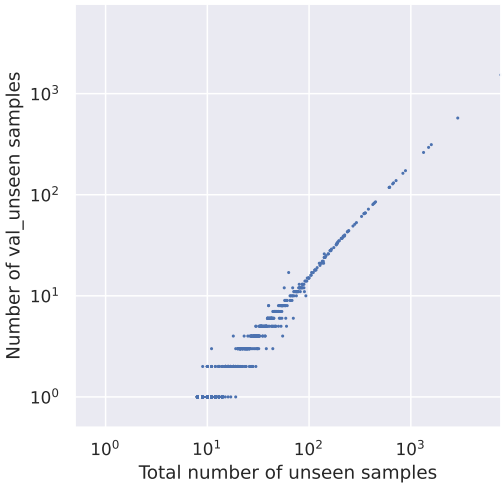
(a) `train` partition.



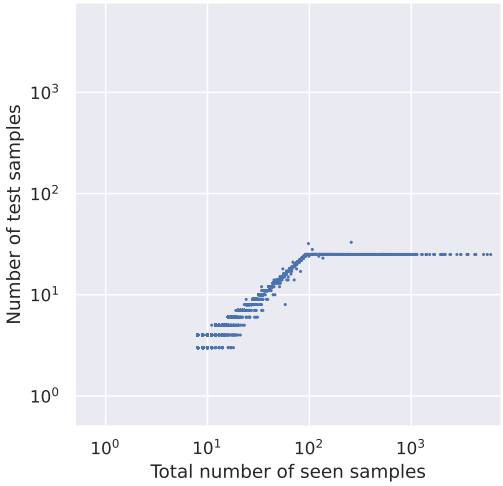
(b) `key_unseen` partition.



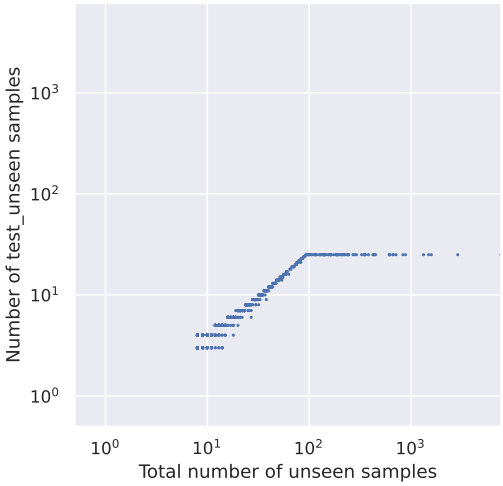
(c) `val` partition.



(d) `val_unseen` partition.

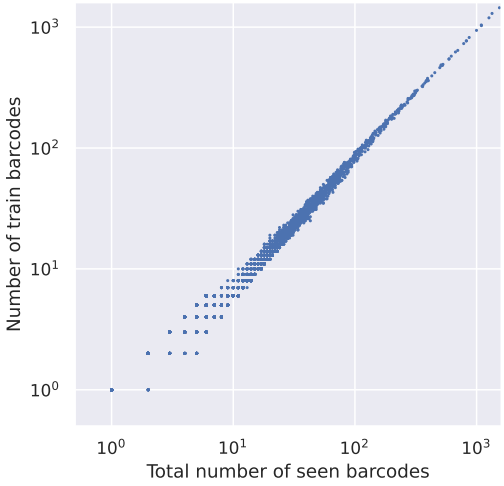


(e) `test` partition.

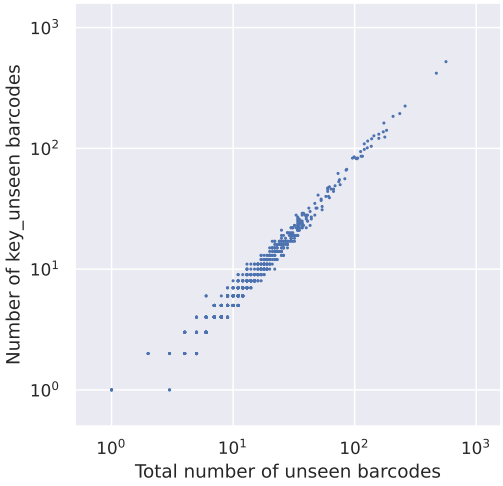


(f) `test_unseen` partition.

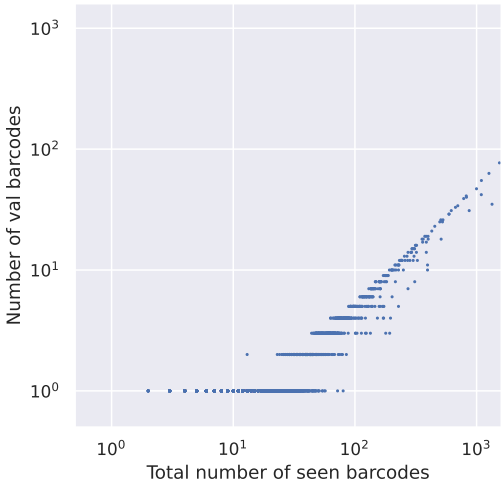
Figure 20: Number of samples in species set and partition, per species.



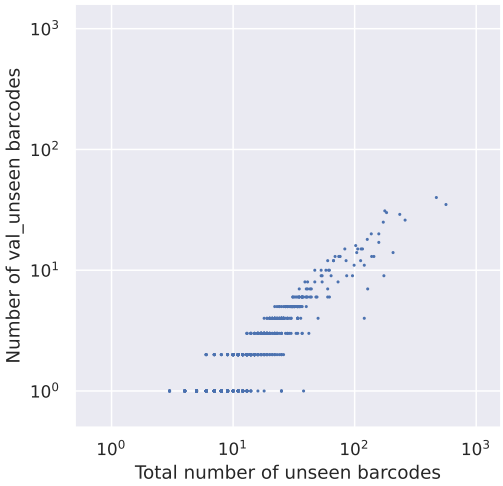
(a) `train` partition.



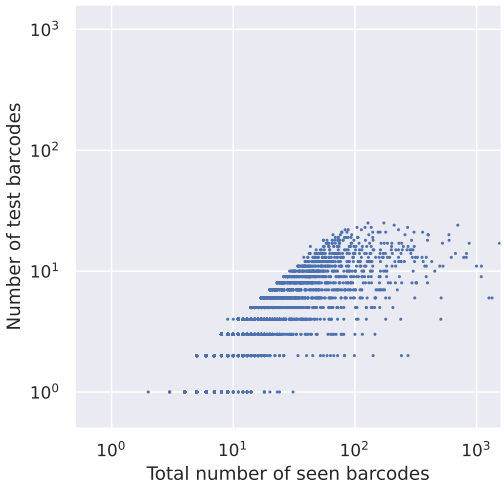
(b) `key_unseen` partition.



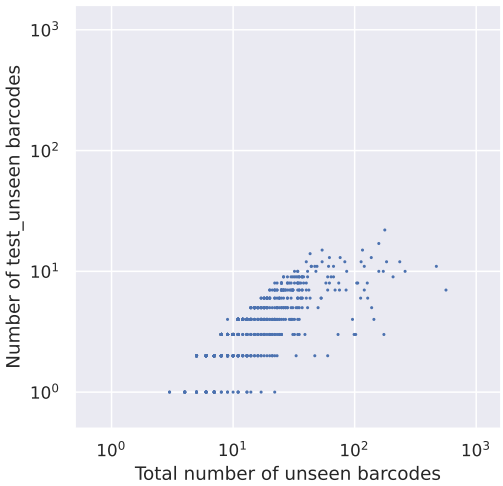
(c) `val` partition.



(d) `val_unseen` partition.

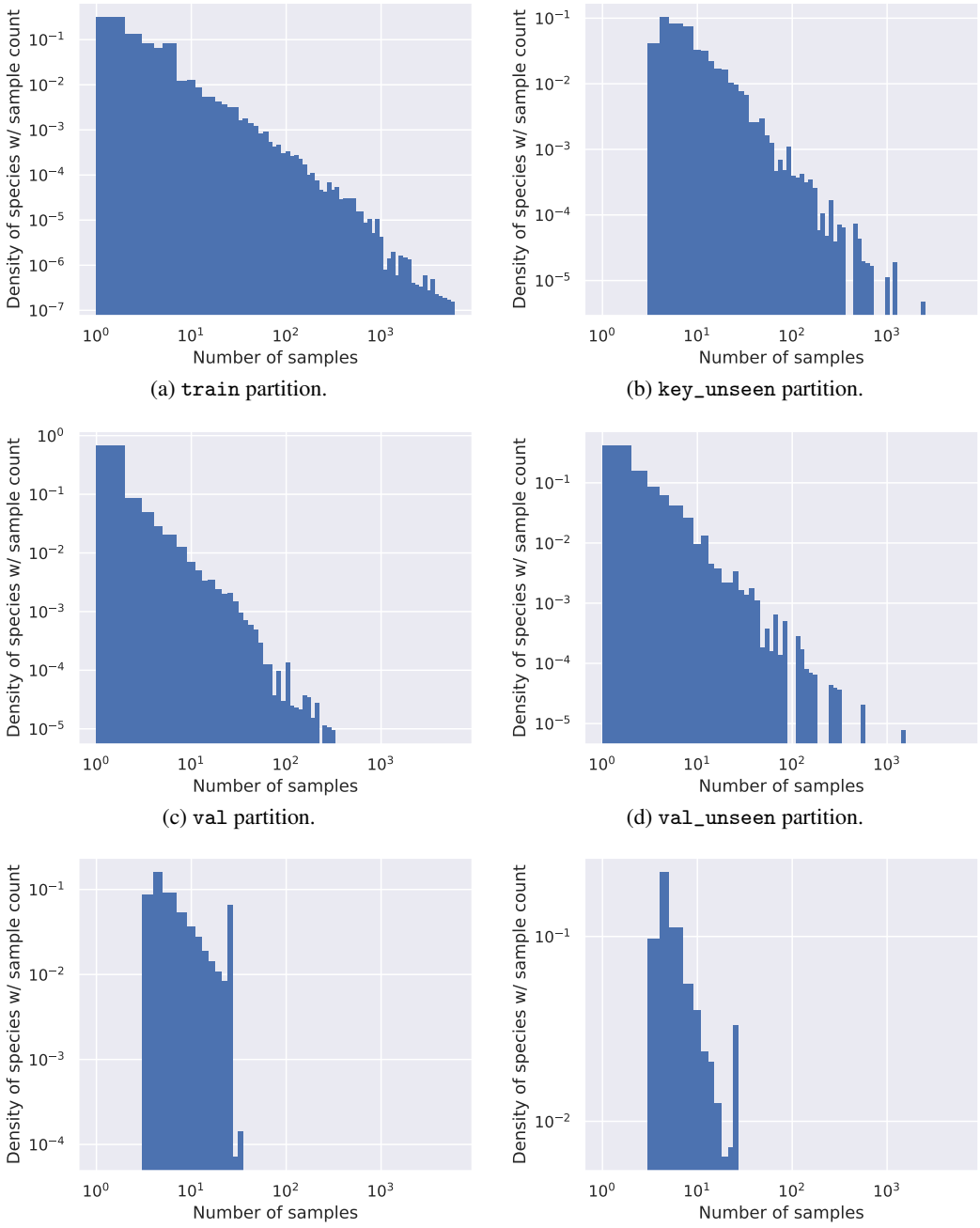


(e) `test` partition.



(f) `test_unseen` partition.

Figure 21: Number of barcodes in species set and partition, per species.



(a) `train` partition.

(b) `key_unseen` partition.

(c) `val` partition.

(d) `val_unseen` partition.

(e) `test` partition.

(f) `test_unseen` partition.

Figure 22: **Distribution of species prevalences across the main data partitions.** Note the log-log axes due to the power law distribution of the data.. The majority of species are infrequent, but some species have many samples. The `train` and `key_unseen` partitions have similar distributions to the overall distribution for *seen* and *unseen* species. The `val` partitions have the same distribution, but shifted left as they they contain a fixed fraction of the samples per species. The `test` partitions are truncated with a minimum and maximum number of samples per species, which flattens the distribution over species for these partitions.

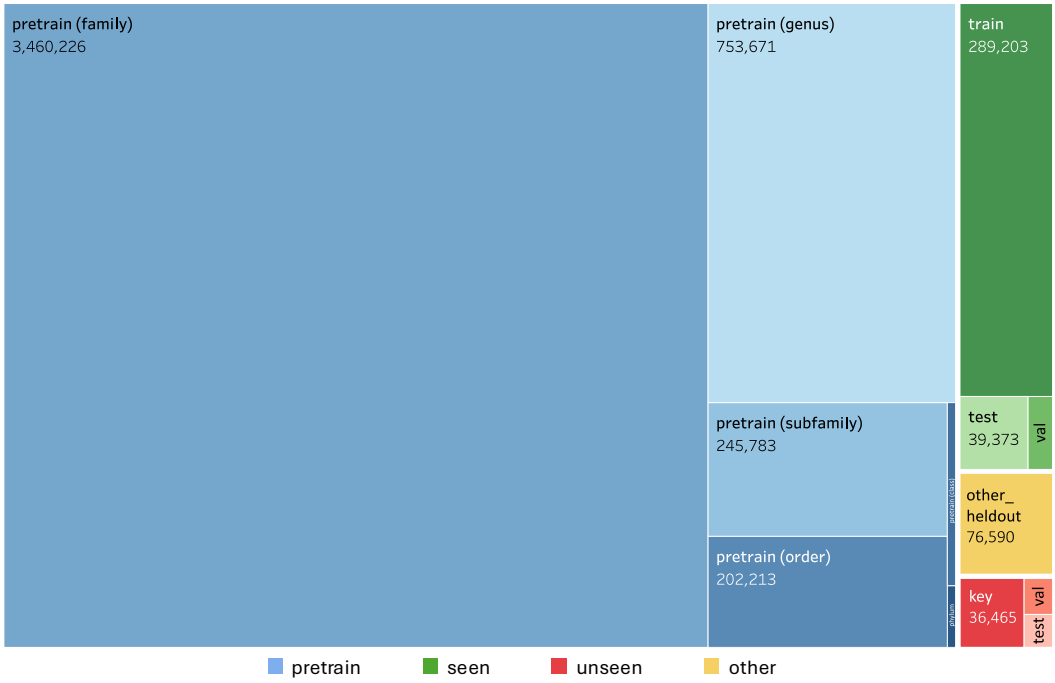
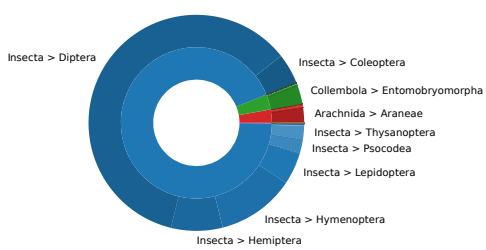
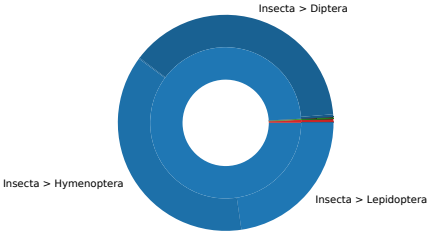


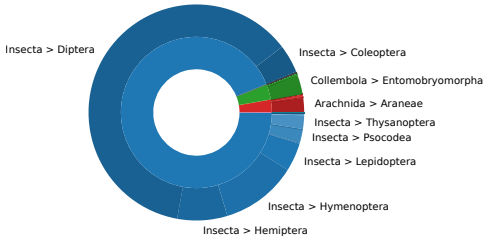
Figure 23: **Treemap diagram showing number of samples per partition.** For the `pretrain` partition (blues), we provide a further breakdown indicating the most fine-grained taxonomic rank that is labelled for the samples. For the remainder of the partitions (all of which are labelled to species level) we show the number of samples in the partition. Samples for seen species are shown in shades of green, and unseen in shades of red.



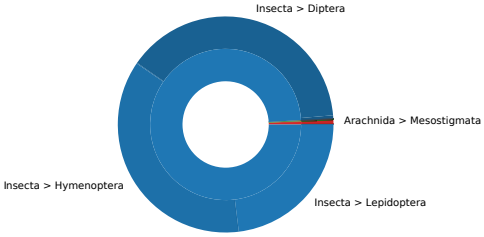
(a) `train` partition.



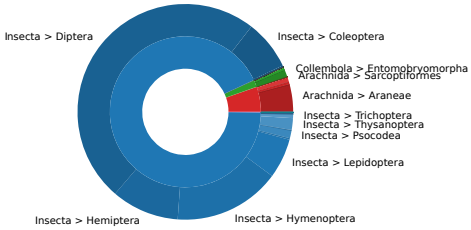
(b) `key_unseen` partition.



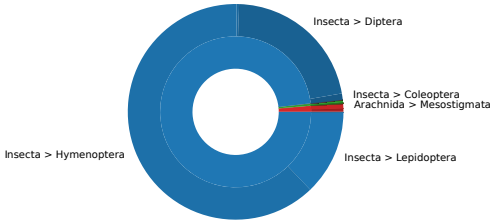
(c) `val` partition.



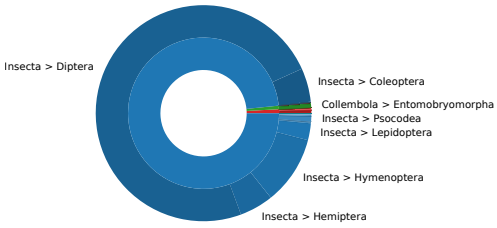
(d) `val_unseen` partition.



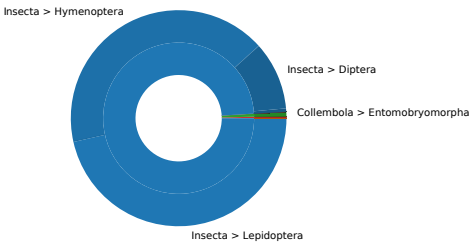
(e) `test` partition.



(f) `test_unseen` partition.



(g) `pretrain` partition.



(h) `other_heldout` partition.

Figure 24: **Distribution of class and order taxa for each partition.** In each panel, the distribution of taxa is shown for one data split in the dataset. Classes are each shown in a different hue (inner ring; Arachnida: red, Collembola: green, Insecta: blue), with orders within a given class shown in the outer ring a different shades of the same hue. Taxonomy hierarchy for orders with more than 0.5% of the data are annotated as `ClassName > OrderName`, where `ClassName` denotes the class and `OrderName` denotes the order.

R.3 Distributional shift

As described above, we partitioned our data into sets to use for open- and closed-world tasks. The division of our data was directed by the labels, with scientific names places in the "seen" species set and placeholder names in the "unseen" species set. This partitioning method means our open-world dataset should be, by construction, well-aligned with the open-world task seen in practice for novel data collection. Novel arthropod species are continually being discovered and identified as species

Table 24: **Distribution of predominant classes and orders across data splits.** For each taxonomic class present in the dataset, and selected orders which have a prevalence of at least 0.5% for at least one split, we show the proportion of samples in each split (%) bearing this taxonomic label. Values for orders which never occur in a split are left empty. Background: linear colour scale from 0% (white) to 75% (blue).

			Seen species			Unseen species			
Class	**Order**	pretrain	train	val	test	key_unseen	val_unseen	test_unseen	other_heldout
Arachnida	Araneae	0.35	2.25	2.15	4.11	0.17	0.16	0.49	0.14
	Mesostigmata	0.08	0.13	0.14	0.36	0.49	0.53	0.81	0.16
	Sarcoptiformes	0.09	0.34	0.33	0.61				
	(Other)	0.31	0.13	0.12	0.27				0.01
Branchiopoda	(Total)	0.00	0.01	0.01	0.04				
Chilopoda	(Total)	0.00	0.00						0.01
Collembola	Entomobryomorpha	0.57	2.80	2.91	1.03	0.06	0.07	0.14	0.65
	(Other)	0.19	0.43	0.45	0.49	0.16	0.17	0.36	0.02
Copepoda	(Total)	0.00	0.00						
Diplopoda	(Total)	0.00	0.00		0.01				
Diplura	(Total)	0.00							
Insecta	Coleoptera	5.02	4.47	4.20	7.44	0.39	0.43	0.94	0.48
	Diptera	73.64	60.56	61.75	49.21	38.44	38.96	21.74	10.19
	Hemiptera	5.06	7.75	7.51	10.15	0.18	0.12	0.36	0.05
	Hymenoptera	10.23	11.64	11.42	16.00	37.46	36.50	62.32	41.84
	Lepidoptera	2.51	4.75	4.26	5.96	22.65	23.04	12.79	46.39
	Psocodea	0.84	2.05	2.09	1.22				0.01
	Thysanoptera	0.20	2.02	2.04	1.77				0.00
	Trichoptera	0.17	0.24	0.24	0.58	0.01	0.01	0.05	0.01
	(Other)	0.38	0.31	0.30	0.66				0.06
Malacostraca	(Total)	0.00	0.09	0.08	0.10				
Ostracoda	(Total)	0.00	0.00						

that are new to science, and if we assume there is a uniform efficiency for naming across taxa the distribution is of placeholder names is likely to match the distribution of new species discovery. However, this distribution does not necessarily match that of taxa prevalence, due to several factors such as non-uniform speciation rates across arthropods.

We investigated the difference in the distribution at order and class level for the dataset partitions, tabulated in Table 24 and illustrated in Figure 24. We observe that the Diptera (e.g. fly) class of Insecta dominates the `pretrain` dataset, but "seen" partitions have a flatter distribution with more prevalence of two non-Insecta orders—Arachnida (spiders, etc.) and Collembola (springtails)—and more instances of non-Diptera Insecta classes. The distribution is even flatter for the `test` partition, due to our capped subsampling methodology when creating the partition.

For "unseen" partitions, we find the data is split nearly equally between three dominant Insecta classes—Diptera (flies, etc.), Hymenoptera (bees, ants, etc.), and Lepidoptera (butterflies, etc.). The `test_unseen` partition contains even more Hymenoptera (around 62%). The `other_heldout` partition has even less Diptera, and is instead dominated by Lepidoptera and Hymenoptera.

Users of the BIOSCAN-5M dataset should thus be sure to consider the effect of this distributional shift on their results if they wish to make direct comparisons between the `test` performance and the `test_unseen` performance—results for these partitions are not intended to be directly comparable to each other.