BarcodeBERT: Transformers for Biodiversity Analysis

Pablo Millan Arias[1*], **Niousha Sadjadi**[1*], **Monireh Safari**[1*], **ZeMing Gong**[3†], **Austin T. Wang**[3†], **Scott C. Lowe**[4,7], **Joakim Bruslund Haurum**[6], **Iuliia Zarubiieva**[2,4], **Dirk Steinke**[2], **Lila Kari**[1], **Angel X. Chang**[3,5], **Graham W. Taylor**[2,4‡]
[1]University of Waterloo, [2]University of Guelph, [3]Simon Fraser University,
[4]Vector Institute for AI, [5]Alberta Machine Intelligence Institute (Amii),
[6]Aalborg University and Pioneer Centre for AI, [7]Dalhousie University

Abstract

Understanding biodiversity is a global challenge, in which DNA barcodes—short snippets of DNA that cluster by species—play a pivotal role. In particular, invertebrates, a highly diverse and under-explored group, pose unique taxonomic complexities. We explore machine learning approaches, comparing supervised CNNs, fine-tuned foundation models, and a DNA barcode-specific masking strategy across datasets of varying complexity. While simpler datasets and tasks favor supervised CNNs or fine-tuned transformers, challenging species-level identification demands a paradigm shift towards self-supervised pretraining. We propose BarcodeBERT, the first self-supervised method for general biodiversity analysis, leveraging a 1.5 M invertebrate DNA barcode reference library. This work highlights how dataset specifics and coverage impact model selection, and underscores the role of self-supervised pretraining in achieving high-accuracy DNA barcode-based identification at the species and genus level. Indeed, without the fine-tuning step, BarcodeBERT pretrained on a large DNA barcode dataset outperforms DNABERT and DNABERT-2 on multiple downstream classification tasks. The code repository is available at `https://github.com/Kari-Genomics-Lab/BarcodeBERT`

1 Introduction

The task of estimating and understanding the biodiversity of our planet remains a monumental challenge, as traditional methods of taxonomic analysis often struggle to keep pace with the discovery and identification of new species. In this context, a 658 base pair long fragment of the Cytochrome c Oxidase Subunit I (COI) gene [13], commonly called the DNA barcode for animals [11], has emerged as a fundamental tool in biodiversity analysis [9] as it can be used to address the challenges related to the large number of unidentified species and the general complexities of taxonomic identification.

Among the numerous taxonomic groups on our planet to which DNA barcoding is applicable, invertebrates, in particular arthropods, stand out as an incredibly diverse and taxonomically complex group [4], where multiple new species are described every day. As a result, the pursuit of effective algorithmic approaches to decipher and understand their taxonomy has become a primary goal in the utilization of barcodes for broader species- and genus-level identification, with previous methodologies yielding promising results in related tasks [15, 1]. Furthermore, the adoption of machine learning (ML) has gained traction, given the classification-oriented nature of these tasks. Recent studies [2] propose a Bayesian framework based on convolutional neural networks (CNN) which, when combined with visual information, achieves high accuracies in species-level identification of seen species and genus-level inference of novel species in a dataset of $\sim$32,000 insect DNA barcodes.

*Joint first author. †Joint second author. ‡Author for correspondence: `gwtaylor@uoguelph.ca`

4th Workshop on Self-Supervised Learning: Theory and Practice (NeurIPS 2023).

Transformer-based models [17] have revolutionized various ML tasks, even those typically dominated by convolutional neural networks (CNNs) [16, 6]. These models, known for their ability to capture complex patterns in sets and sequences, have found applications across diverse domains thanks to their effectiveness in learning from large unlabelled datasets [5, 16]. Transformers pretrained with self-supervised learning (SSL) at scale, a.k.a. foundation models, are often task-agnostic and expected to perform well after fine-tuning for various downstream tasks. Yet, their application for taxonomic identification using DNA barcodes has not been extensively explored. Foundation models for DNA primarily target human sequences [12, 18, 7], which intuitively makes them unsuitable for barcode data. In response, this paper delves into the intersection of genomics, ML, and biodiversity, aiming to unlock the potential of transformer-based architectures for species-level identification of insects.

We propose BarcodeBERT, a self-supervised method using DNA barcodes for general biodiversity analysis. BarcodeBERT leverages a reference library containing 1.5 M invertebrate barcodes [8] for the training of a masked language model (MLM) that is effective in learning meaningful embeddings of the data and that can be used for successful species-level classification of DNA barcodes of insects in general scenarios. In our evaluations, BarcodeBERT is compared against recent DNA-based foundation models [18] and a CNN baseline, all trained/fine-tuned on a medium-size dataset comprising DNA barcodes from 1,390 species. While all models excelled on the DNA barcode-based species-level identification task, the results for the challenging task of zero-shot learning of images with barcodes as side-information evidenced the superiority, for general biodiversity analysis, of transformer models pretrained on domain-specific datasets.

The main contributions of this paper are: (1) Introducing BarcodeBERT, a pioneering self-supervised method employing DNA barcodes for biodiversity analysis through transformer-based models; (2) An in-depth comparison across several taxonomic classification tasks between BarcodeBERT, an earlier successful CNN baseline, and two fine-tuned foundation models pretrained on human DNA (DNABERT) and multi-species DNA (DNABERT-2); (3) A pipeline demonstrating how transformer-based DNA barcode encoders facilitate zero-shot classification of insect images. Unlike prior work [1] that utilized DNABERT for DNA barcode feature extraction, our strategy emphasizes that direct masked pretraining on barcodes can bridge the performance gap between supervised and semi-supervised training paradigms.

2 Methods

This section presents an account of the data processing pipeline, including the steps taken to curate the dataset from the reference library, and a description of the used architectures and metaparameters. In addition, it describes the evaluation framework and the downstream tasks used for testing.

2.1 Dataset

The primary source of data for this study is the reference library for Canadian invertebrates [8], containing 1.5 M DNA samples, which was collated from the Barcode of Life Database (BOLD) [14].

Data Preprocessing. To ensure data integrity and consistency, we performed a series of preprocessing steps over this dataset. First, empty entries were removed and IUPAC Ambiguity Codes (non-`ACGT` symbols), including alignment gaps, were uniformly replaced with the symbol `N`. Duplicated sequences, even with different identifiers, were removed to avoid redundancy and increase the complexity of the training and pretraining tasks. Sequences with trailing `N`'s were truncated. Finally, sequences falling below 200 base pairs or exhibiting over 50% `N` content were excluded.

Data Split. After preprocessing, 974,057 sequences were obtained. The dataset was divided into three subsets for various evaluation purposes: (*i*) Fine-Tuning subset: This dataset was curated to assess the model's efficacy in classifying known species. It consists of 67,267 barcodes from 1,653 species representing 500 different genera. Each genus is represented by at least 20 barcodes and at most 50 barcodes. This subset was further divided into training (70%), testing (20%), and validation (10%) splits. (*ii*) Unseen evaluation subset: This dataset was created to simulate the real-world scenario of encountering previously unknown species. We sampled a maximum of 20 barcodes from each of the 500 representative genera present in the Fine-Tuning subset, for genus-level identification. This subset comprises 4,278 sequences from 1,826 "rare" species, all of which are absent from the training data and have fewer than 20 barcodes in total in the reference dataset. (*iii*) Unsupervised pretraining subset: The remaining barcodes, including sequences with incomplete taxonomic annotations at

Table 1: Classification accuracy of DNA barcode models under different SSL evaluation strategies. Some models supported variable stride length; for these we show results at several k-mer lengths.

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

various levels, were used. We excluded all sequences belonging to the species present in Unseen. Finally, to benchmark against prior work, we also utilized the INSECT dataset as introduced in [1].

2.2 Network Architectures

CNN baseline. Adapted from [1], it comprises three convolutional layers, each followed by batch normalization and max-pooling. The output of the third convolutional layer is flattened, batch normalized, and connected to a linear layer with 500 units that are then connected to the output layer.

Foundation models. Our comparison includes two pretrained foundation models based on the Bidirectional Encoder Representations from Transformers model (BERT). These are capable of converting sequence inputs into embedding vectors and they can be further trained using self-supervised and/or supervised objectives. Within this transformer-based architecture, multi-head attention units play a vital role in capturing relations among input sequences at various scales, encompassing both small-scale and large-scale interactions. The first model, DNABERT, captures global and transferable genomic understanding by leveraging nucleotide contexts using an overlapping k-mer window for tokenization. The model is highly accurate at predicting splicing and transcriptor factor binding sites. The second, DNABERT-2, pioneers the use of Byte-Pair Encoding (BPE) in this domain and overcomes inefficiencies in genomic tokenization through non-overlapping k-mers.

BarcodeBERT. Inspired by the BERT architecture, BarcodeBERT features 12 attention heads, 12 layers, and a maximum sequence length of 512. After DNA barcodes are segmented into non-overlapping k-mers, BarcodeBERT encodes the sequence of k-mers into a sequence of d-dimensional vectors ($d=768$). Since our primary objective is to generate an embedding vector that encapsulates information across the entire DNA barcode, following a self-supervised training phase, we merge these d-dimensional vectors for each DNA sequence to create a comprehensive vector representation for the entire sequence using global average pooling. During training, we focused exclusively on masked token prediction, masking 50% of the input tokens and performed experiments across different k-mer lengths ($4 \leq k \leq 6$) to observe the impact of k-mer length on embedding quality.

3 Experiments

To explore the transformer architecture applicability for DNA barcode-based biodiversity analyses, we employ various SSL evaluation strategies [3], comparing their performance to a supervised baseline.

3.1 Taxonomic classification of DNA barcodes

Methodology. We first perform task-specific fine-tuning, *i.e.,* we fine-tune the models on the supervised training dataset and assess their performance at species-level classification. Second, we gauge the influence of pretraining on DNA barcodes by using the models as feature extractors. We first implement genus-level 1-NN probing on sequences from unseen species, providing insights into the models' ability to generalize to new taxonomic groups. Additionally, we perform species-level classification using a linear classifier trained on embeddings from the pretrained models. Similar to training, we tokenize DNA barcodes into non-overlapping k-mers and feed the sequence of tokens into the model. We average over the k-mer outputs to generate an overall embedding for the barcode.

Table 2: Evaluation of DNA barcode models in a Bayesian zero-shot learning task on the INSECT dataset. The pretraining and fine-tuning data source is indicated by the respective DNA type and '–' signifies the absence of training of that type. We also indicate the most specific taxon subset. For the baseline CNN encoder, we report the original paper result (left) and reproduced result (right).

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

Results. As detailed in Table 1, fine-tuning revealed no significant performance gap, with DNABERT-2 marginally outperforming all other models. In the genus-level 1-NN probing task, both BarcodeBERT and DNABERT-2 outperformed the baseline, with DNABERT-2 performing less competitively. Linear probing, however, favored our pretrained models and DNABERT-2 over the baseline and DNABERT. It is noteworthy that both BarcodeBERT and DNABERT-2 outperformed DNABERT in two out of three tasks. This likely stems from the non-overlapping tokenization approach and the fact that DNABERT-2 was not exclusively pretrained on human data. Although the baseline model performed well, the transformer-based models demonstrate their potential to contribute significantly to the field of DNA barcode analysis.

3.2 Bayesian zero-shot learning of images with DNA as side information

Methodology. Following [1], we evaluate BarcodeBERT's performance in the context of Bayesian zero-shot learning (BZSL) on the INSECT dataset for species claneen species using the K-nearest seen classes in the DNA feature space, with local priors defined by image features. We evaluate use of the DNA feature embeddings directly from the pretrained BERT models as well as after fine-tuning the models on the species classification task on the INSECT dataset. We utilize image features from the INSECT dataset [1], pre-extracted using ResNet-101 [10], to ensure that our results can be compared effectively to the baseline supervised CNN used in [1]. We tokenize the barcode data using overlapping k-mers for DNABERT, with $k = 6$ and the BPE tokenizer for DNABERT-2. We did not align barcodes as in [1], as we found that it did not significantly affect the results. For each model, we perform a grid search over the same hyperparameter space used by [1] for the Bayesian model. The accuracy for seen and unseen test species, as well as the harmonic mean, are presented in Table 2.

Results. Even without fine-tuning, BarcodeBERT substantially outperforms DNABERT and DNABERT-2 on unseen species, regardless of whether they had been fine-tuned previously or not. BarcodeBERT achieves similar performance to the reported baseline CNN results [1] and improves on the harmonic mean score by 1.2% and unseen accuracy by 1.9%, respectively. We thus find that in the zero-shot learning task of predicting insect species, employing BERT-like models that have also been pretrained on insect DNA barcodes as DNA encoders can improve performance.

4 Conclusions

Our research shows that pretraining masked language models on DNA barcode data, as demonstrated by BarcodeBERT, is both effective and essential for arthropod species identification. This underscores the need to diversify datasets beyond human DNA sequences, to advance the field of biodiversity science. While we have made strides in improving the classification of arthropod species from both DNA sequences and images, our findings point to a wealth of untapped data, *e.g.*, the BOLD dataset, currently comprising 14 million DNA barcodes, continuously augmented by data from previously seen or unseen species. Future work includes further investigation of such DNA barcode data, to develop more robust and scalable self-supervised models for taxonomic classification.

Acknowledgements

We acknowledge the support of the Government of Canada's New Frontiers in Research Fund (NFRF) [NFRFT-2020-00073], the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Discovery Grant [RGPIN-2023-03663] to LK, Canada Research Chairs [950-231707], and a NVIDIA Academic Grant to GWT.

References

[1] S. Badirli, Z. Akata, G. Mohler, C. Picard, and M. M. Dundar. Fine-grained zero-shot learning with DNA as side information. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19352–19362. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a18630ab1c3b9f14454cf70dc7114834-Paper.pdf.

[2] S. Badirli, C. J. Picard, G. Mohler, F. Richert, Z. Akata, and M. Dundar. 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. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14104.

[3] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum. A cookbook of self-supervised learning, 2023. arXiv: 2304.12210.

[4] Y. Basset, L. Cizek, P. Cuénoud, R. K. Didham, F. Guilhaumon, O. Missa, V. Novotny, F. Ødegaard, T. Roslin, J. Schmidl, A. K. Tishechkin, N. N. Winchester, D. W. Roubik, H.-P. Aberlenc, J. Bail, H. Barrios, J. R. Bridle, G. Castaño-Meneses, B. Corbara, G. Curletti, W. D. da Rocha, D. D. Bakker, J. H. C. Delabie, A. Dejean, L. L. Fagan, A. Floren, R. L. Kitching, E. Medianero, S. E. Miller, E. G. de Oliveira, J. Orivel, M. Pollet, M. Rapp, S. P. Ribeiro, Y. Roisin, J. B. Schmidt, L. Sørensen, and M. Leponce. Arthropod diversity in a tropical forest. *Science*, 338(6113):1481–1484, 2012. doi: 10.1126/science.1226727. URL https://www.science.org/doi/abs/10.1126/science.1226727.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

[7] H. Dalla-Torre, L. Gonzalez, J. M. Revilla, N. L. Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679. URL https://www.biorxiv.org/content/early/2023/01/15/2023.01.11.523679.

[8] J. R. deWaard, S. Ratnasingham, E. V. Zakharov, A. V. Borisenko, D. Steinke, A. C. Telfer, K. H. J. Perez, J. E. Sones, M. R. Young, V. Levesque-Beaudin, C. N. Sobel, A. Abrahamyan, K. Bessonov, G. Blagoev, S. L. deWaard, C. Ho, N. V. Ivanova, K. K. S. Layton, L. Lu, R. Manjunath, J. T. A. McKeown, M. A. Milton, R. Miskie, N. Monkhouse, S. Naik, N. Nikolova, M. Pentinsaari, S. W. J. Prosser, A. E. Radulovici, C. Steinke, C. P. Warne, and P. D. N. Hebert. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*, 6(1):308, Dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0320-2.

[9] A. Dopheide, L. K. Tooman, S. Grosser, B. Agabiti, B. Rhode, D. Xie, M. I. Stevens, N. Nelson, T. R. Buckley, A. J. Drummond, and R. D. Newcomb. Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications*, 29(4):e01877, 2019. doi: 10.1002/eap.1877. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/eap.1877.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[11] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard. 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. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2002.2218.

[12] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083.

[13] D. H. Lunt, D.-X. Zhang, J. M. Szymura, and O. M. Hewltt. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. *Insect Molecular Biology*, 5(3):153–165, 1996. doi: 10.1111/j.1365-2583.1996.tb00049.x. URL https://resjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2583.1996.tb00049.x.

[14] S. Ratnasingham and P. D. N. Hebert. BOLD: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3):355–364, 2007. doi: 10.1111/j.1471-8286.2007.01678.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-8286.2007.01678.x.

[15] S. Ratnasingham and P. D. N. Hebert. A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLOS ONE*, 8(7):1–16, 07 2013. doi: 10.1371/journal.pone.0066213.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

[18] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome, 2023. arXiv: 2306.15006.

A Appendix

Details about dataset distribution, training, and optimization are provided in this section.

A.1 Dataset Distribution

We designed a data split that mirrored real-life conditions, with a trade-off between dataset size, challenging scenarios and suitability for evaluation purposes. The Fine-tuning and Unseen subsets contain data from 500 representative genera, sampled according to the original dataset distribution, resulting in 500 genera from the class Insecta. Figure A.1 displays the distribution of the most relevant orders present in each dataset. It is important to note that the inclusion of more "rare" species is evidenced in the increased percentage of "other" orders in the Unseen dataset with respect to the Fine-Tuning dataset. Table A.1 displays the distribution of sequences obtained from the reference library for Canadian invertebrate [8]. These sequences were used for pretraining of our models following the preprocessing step. The dataset contains 14,794 unique species, with the majority classified under the Arthropoda phylum. Note that several sequences with incomplete taxonomy were included in this dataset.

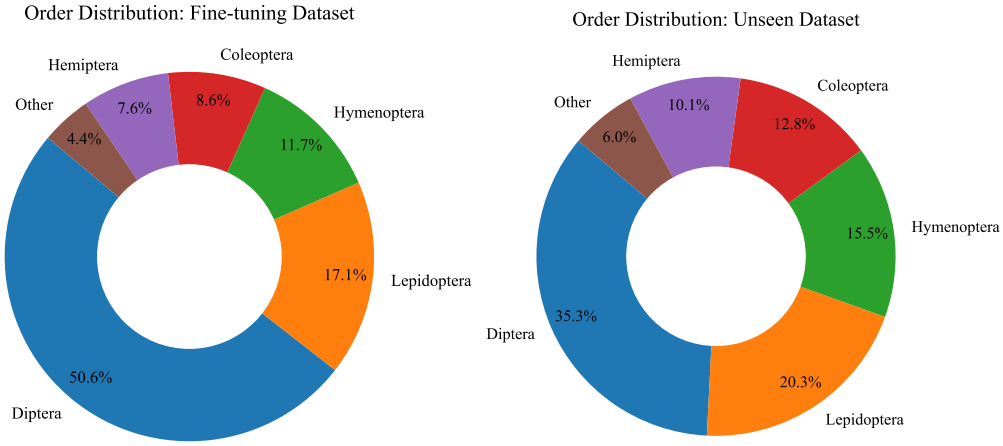

Figure A.1: Distribution of orders in the Fine-tuning (left) and unseen (right) datasets.

Table A.1: The distribution of barcode sequences used in the pre-training phase.

Phylum name	# ID	# BIN	# Class	# Order	# Family	# Genus	# Species	# Sequences
Annelida	2102	516	2	16	48	150	329	2102
Arthropoda	888934	61328	14	67	929	6211	13991	888934
Brachiopoda	20	2	1	2	2	2	2	20
Bryozoa	5	4	3	3	3	2	2	5
Chordata	289	102	5	18	37	67	89	289
Cnidaria	112	46	4	10	24	25	24	112
Echinodermata	276	79	5	17	26	43	74	276
Hemichordata	4	2	1	1	1	2	1	4
Mollusca	1912	372	6	30	97	162	271	1912
Nematoda	24	8	2	5	10	5	2	24
Nemertea	56	22	3	2	5	5	5	56
Platyhelminthes	1	1	0	0	0	0	0	1
Porifera	7	5	1	3	4	4	3	7
Priapulida	1	1	1	1	1	1	1	1
Tardigrada	1	1	1	1	1	0	0	1

A.2 Training and Optimization

Pretraining. As previously mentioned, our method entails the segmentation of each DNA barcode into a series of non-overlapping k-mers. The standard DNA alphabet comprises the nucleotides `A`, `C`, `G`, and `T`. However, note that certain DNA barcodes may incorporate other symbols, such as `N`'s or alignment gaps '`-`' within their sequences, denoting ambiguity. Our vocabulary encompasses all possible combinations of k-length strings derived from the nucleotide alphabet, supplemented by two special tokens: `<MASK>` and `<UNK>`. The `<MASK>` token is utilized for masking k-mers during the training phase, and k-mers containing any symbol that is not present in the nucleotide alphabet, are assigned the `<UNK>` token. Consequently, the total vocabulary size is determined by the expression $4^k + 2$.

We implement BarcodeBERT using the Hugging Face Transformers library and PyTorch. During training, we focused exclusively on masked token prediction, masking 50% of the input tokens and optimizing the network with a cross-entropy loss. We utilize the AdamW optimizer and incorporate a linear scheduler with an initial learning rate of 10^{-4} during the optimization process. Additionally, we performed experiments across different k-mer lengths ($4 \leq k \leq 6$) to observe the impact of k-mer length on embedding quality.

The network is trained for 40 epochs with a batch size of 16. Figure A.2 displays the pretraining loss over the course of these 40 epochs. The loss behavior demonstrates convergence, regardless of the k-mer length.

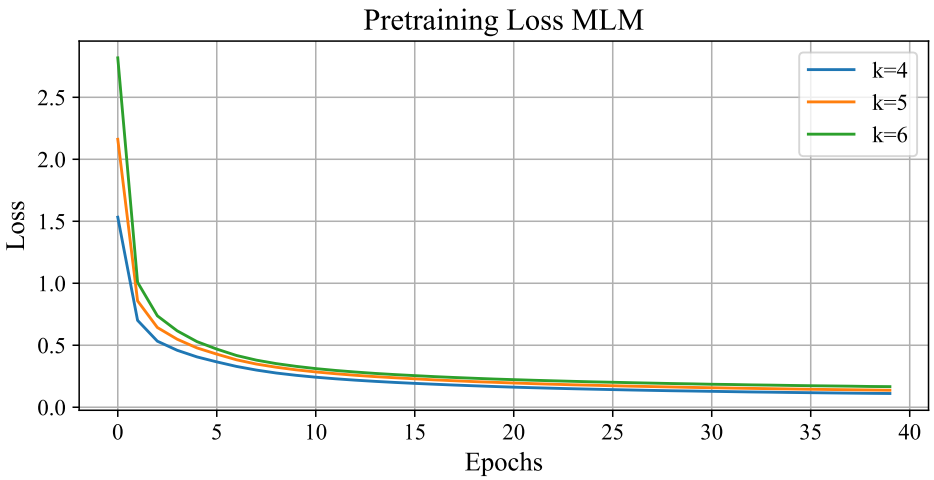

Figure A.2: Mask prediction loss over 40 epochs of training for different k-mer lengths.

Linear-Probing. As outlined in the experimental section, one of the evaluation methodologies involves the utilization of linear probing. For this purpose, a linear classifier is applied to the embeddings generated by pretrained models for species-level classification. The linear classifier is a simple linear perceptron model implemented using the Scikit-learn library in Python. This evaluation procedure is conducted across a range of models, including the base model, which is the Convolutional Neural Network (CNN), as well as two foundational models, DNABERT and DNABERT-2, and BarcodeBERT for different k-mers values ($4 \leq k \leq 6$).

A.3 Zero-shot Learning

The INSECT dataset [1] consists of 21,212 total images across 1,213 unique insect species, with images and associated DNA barcodes from the Barcode of Life dataset [14]. The species are distributed among three orders—Diptera, Coleoptera, and Hymenoptera—with the most species coming from Coleoptera. We use the same splits as in [1] for fine-tuning and evaluation, where 10% of the species are set aside as unseen classes for the test set, and 10% of the remaining species are used as unseen classes for the validation set. The remaining seen species are split 80/20% for training and testing, ensuring in all cases that no instance of a given insect is assigned to multiple splits.

We use the same image features from [1], extracted from a base ResNet-101 model pretrained on ImageNet using images resampled to 256×256 and center-cropped.

Fine-tuning with supervised learning. BarcodeBERT was fine-tuned for 12 epochs with a batch size of 32 and learning rate of 5×10^{-3}, using the SGD optimizer and step learning rate scheduler of 0.5 decay per 3 epochs. Cross entropy was used for the loss function. The classifier consisted of two linear layers with a tanh activation and dropout of 0.2. For tokenization, we used non-overlapping 6-mers, as with the pretrained model. A similar training setup was used for DNABERT but with overlapping 6-mers with as the tokenizer.

DNABERT-2 [18] was fine-tuned for 20 epochs using the released DNABERT-2 fine-tuning code and defaults, with a learning rate of 3×10^{-5} and batch size of 32. AdamW was used as the optimizer, and a warmup learning rate scheduler was used during training. Tokenization was performed using the BPE method described in [18], with a maximum length of 265 tokens to approximately allow it to encode the entire barcode sequence. Surprisingly, we found that fine-tuning DNABERT-2 in all variations of our experiments ended up decreasing performance on the test set rather than improving it, unlike for BarcodeBERT and DNABERT. All fine-tuned models achieved similar accuracy—over 95%—on the supervised task.

Bayesian model tuning. For each model, before and after fine-tuning, PCA was applied to the image features, reducing the dimensionality of the embeddings down to 500 before passing it to the Bayesian model. We applied a metaparameter grid search to find the optimal parameters for the Bayesian model. The metaparameters were as follows: the scaling constant (k_0) for the dispersion of centers of metaclasses around a global mean; the scaling constant (k_1) for the dispersion of actual class means around the corresponding metaclass means; the dimension (m) of the Wishart distribution for sampling covariance matrices of metaclasses; a scalar (s) for the mean of class covariances; and the number of nearest-seen classes (K) to use in defining the PPD. Following [1], our grid search space was over $k_0 = [0.1, 1]$, $k_1 = [10, 25]$, $m = [2500, 12500, 50000, 250000]$, $s = [1, 5, 10]$, and $K = [1, 2, 3]$. The best hyperparameters were selected on a validation partition.