Bioinformatics, 2024, pp. 1–10

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

BarcodeBERT: Transformers for Biodiversity Analyses

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

[1]University of Waterloo, [2]University of Guelph, [3]Simon Fraser University, [4]Vector Institute, [5]Alberta Machine Intelligence Institute (Amii) and [6]Aalborg University and Pioneer Centre for AI
[*]Joint first author [†]Joint second author [‡]Joint senior author [♯]Corresponding authors: gwtaylor@uguelph.ca, lila@uwaterloo.ca
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

In the global challenge of understanding and characterizing biodiversity, short species-specific genomic sequences known as DNA barcodes play a critical role, enabling fine-grained comparisons among organisms within the same kingdom of life. Although machine learning algorithms specifically designed for the analysis of DNA barcodes are becoming more popular, most existing methodologies rely on generic supervised training algorithms. We introduce BarcodeBERT, a family of models tailored to biodiversity analysis and trained exclusively on data from a reference library of 1.5 M invertebrate DNA barcodes. We compared the performance of BarcodeBERT on taxonomic identification tasks against a spectrum of machine learning approaches including supervised training of classical neural architectures and fine-tuning of general DNA foundation models. Our self-supervised pretraining strategies on domain-specific data outperform fine-tuned foundation models, especially in identification tasks involving lower taxa such as genera and species. We also compared BarcodeBERT with BLAST, one of the most widely used bioinformatics tools for sequence searching, and found that our method matched BLAST's performance in species-level classification while being 55 times faster. Our analysis of masking and tokenization strategies also provides practical guidance for building customized DNA language models, emphasizing the importance of aligning model training strategies with dataset characteristics and domain knowledge. The code repository is available at `https://github.com/bioscan-ml/BarcodeBERT`.

Key words: Biodiversity informatics, taxonomic classification, DNA barcode, machine learning, transformers, DNA language models

Introduction

The task of estimating and understanding biodiversity on our planet remains a monumental challenge, as traditional methods of taxonomic analysis often struggle to keep pace with the rate of discovery and identification of new species. In this context, the search for highly expressive, short standardized genomic regions containing meaningful taxonomic information (DNA barcodes) has become prominent in biodiversity research over the past two decades [18, 24, 19, 31]. Specifically, a 658-base-pair-long fragment of the Cytochrome c Oxidase Subunit I (COI) gene [23] has emerged as the de facto DNA barcode for kingdom *Animalia* [11] and has proven effective in addressing inherent taxonomic challenges. Particularly, barcodes can be used for fast and accurate queries to categorize novel specimens into existing taxa. Furthermore, in the absence of clear species boundaries, they can be used to systematically separate specimens into groups of closely related organisms. These clusters, known as operational taxonomic units (OTUs) correspond to groups of similar specimens and can be labelled using e.g. a Barcode Index Number (BIN) [27]. As it is defined systematically, such a BIN system overcomes ambiguities in traditional species labelling and thus accelerates biodiversity research. Among the numerous taxonomic groups to which DNA barcoding is applicable, invertebrates, particularly arthropods, stand out as an incredibly diverse and taxonomically complex group [5], making them the focus of many methodological studies [2, 3, 15]. The diversity and taxonomic richness of this group require specialized algorithmic approaches that can capture the taxonomic structure of the data. Consequently, biodiversity researchers are increasingly turning to machine learning methods, including convolutional neural networks (CNNs) [2] and transformer models [15], to scale taxonomic classification of arthropods and accelerate species discovery.

Transformer-based models, pretrained at scale with self-supervised learning (SSL), also referred to as "foundation models," have found applications across diverse domains thanks to their effectiveness in learning from large unlabelled datasets [6, 32]. These models are often task-agnostic and can perform well on a variety of downstream tasks after fine-tuning.

Despite their success in other domains, their application for taxonomic identification using DNA barcodes has not yet been extensively explored. Moreover, most DNA-based foundation models primarily target human chromosomal DNA sequences [34, 9, 20], making them suboptimal for barcode data due to domain shift between the two data types. In particular, DNA barcodes for animals stem from a specific region in mitochondrial DNA, and patterns learned from other genomic regions may be irrelevant for taxonomic classification.

We here aim to unlock the potential of transformer-based architectures for taxonomic identification of arthropod barcodes, providing insights that extend beyond broad, foundation-style approaches. We address the previously mentioned issues (i.e. the taxonomic complexity of arthropods, and the lack of specialized transformer models trained on DNA barcodes) by adopting a semi-supervised learning approach, followed by fine-tuning on high-quality labelled barcode data, demonstrating the value of targeted model development for specialized applications. We propose BarcodeBERT, a self-supervised method that leverages a reference library of 1.5M invertebrate barcodes [10] and a masked language model (MLM) training strategy to effectively compute meaningful embeddings of the data, facilitating successful species-level classification of insect DNA barcodes in general scenarios. In addition to the classification of known species, our pretrained models can be used to generate embeddings for sequences from unseen taxa, enabling non-parametric classification at higher levels of the taxonomic hierarchy.

To summarize our contributions, we first investigate the impact of pretraining using a large and diverse DNA barcode dataset (1 million sequences, from more than 17,000 species, across 6,700 genera) on generalization to other downstream tasks. Second, we compare BarcodeBERT against several baselines such as pretrained DNA foundation models (DNABERT [20], DNABERT-2 [34], DNABERT-S [35], the Nucleotide Transformer NT [9], and HyenaDNA [25]), a CNN baseline following the architecture introduced by [2], and the widely used alignment-based method BLAST [1]. Third, our study provides actionable insights regarding tokenization strategies, optimal masking ratios, and the importance of application-specific pretraining for DNA language models.

Overall, BarcodeBERT outperforms all other foundation models in supervised species classification, matching BLAST's accuracy while being 55 times faster and more scalable. Moreover, a linear classifier trained on BarcodeBERT embeddings has ∼6% higher species classification accuracy than the top-performing foundation model in this task. Lastly, the same embeddings can also be used for accurate genus classification using similarity searches, outperforming the top-performing foundation model by ∼30%.

Related Work

The exponential growth of genomic datasets with the advent of high-throughput sequencing has both demanded and enabled a surge in classification tools for DNA sequences. Such tools are essential for large-scale biodiversity studies, where algorithmic approaches can expedite the taxonomic categorization of novel specimens. One intuitive approach is to embed sequences into a vector space where geometric distances approximate taxonomic similarities [8]. This allows for rapid comparisons between newly sequenced and labelled DNA, enabling accurate taxonomic assignments.

Many machine learning approaches, particularly in the area of representation learning, have demonstrated considerable potential in biodiversity analyses as they can embed raw DNA data into an expressive lower dimensional space. Transformer-based models [33], known for their ability to capture complex patterns within sequential data, have shown exceptional performance in various representation learning tasks across domains, either with or without supervision [6, 32, 7]. These models are especially effective in learning from vast unlabelled datasets, making them ideal candidates for the analysis of genomic data, where obtaining high-quality annotations remains challenging.

There has been a growing number of self-supervised learning-based DNA language models proposed recently, most of which are based on the transformer architecture and trained using the masked language model (MLM) objective. The first foundational model in this space, DNABERT, utilizes a BERT-based transformer architecture along with k-mer tokenization for genome sequence prediction tasks. Following DNABERT, other models have emerged, including the Nucleotide Transformer [9], GENA-LM [12], and HyenaDNA [25]. While each model varies in architectural details, tokenization methods, and training data, their reliance on SSL and the MLM objective for pretraining remains a constant. HyenaDNA is a unique entry in this space as it uses a state-space model (SSM) based on the Hyena architecture [26] and trains it for next-token prediction (a causal MLM).

The landscape of machine learning models specifically tailored for DNA barcodes is less developed. A recent study [3] proposes a Bayesian framework based on CNNs 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 ∼32,000 insect DNA barcodes. This method uses supervised learning to compute meaningful embeddings that can be used as side information in a two-layer Bayesian zero-shot learning framework. Transformer methods have been introduced for the classification of fungal Internal Transcribed Spacer sequences without any self-supervision [28].

Although there has been a growing number of SSL-based DNA language models proposed in the recent literature, our findings indicate that models pretrained on a diverse set of non-barcode DNA sequences underperform on downstream barcode tasks. This suggests that general DNA foundation models may struggle with the domain-specific characteristics of barcode data. In this study, we leverage barcode-specific training to improve both species-level classification accuracy and generalization to other taxonomic ranks. By grounding our approach in targeted data and architectural choices, we seek to advance the utility of machine learning in biodiversity research, moving beyond general off-the-shelf models trained to classify specimens into known taxa. Distinctly, our specialized models are not only capable of classifying known species but also can be used for taxonomic classification for species that are not present in the training set.

Methods

In this section, we outline the key elements of our methodology. We begin with a detailed account of our datasets and data processing pipeline. We then describe the architectures and hyperparameters used in the development of BarcodeBERT.

Dataset

We use a reference library for Canadian invertebrates [10] for training and testing purposes. To benchmark our results against prior work, we also use the INSECT dataset introduced by Badirli et al. [2], a small multimodal dataset designed for zero-shot classification of images from unseen species using DNA as auxiliary information.

The reference library dataset contains 1.5 M DNA samples that were directly queried from the Barcode of Life Data system (BOLD) [27]. The dataset was further pre-processed and subdivided as follows.

Data pre-processing

To ensure data integrity and consistency, we performed a series of pre-processing steps over this dataset. First, empty entries were removed. Then, following standard practices [9], IUPAC ambiguity codes (non-`ACGT` symbols), including alignment gaps, were uniformly replaced with the symbol `N`. Duplicate sequences 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` characters were excluded.

Data partitioning

After pre-processing, 965,289 barcode sequences from 17,464 invertebrate species, across 6,712 genera were obtained. The dataset was divided into three distinct partitions for different training and evaluation purposes: (*i*) *Seen*: This partition is intended for supervised learning pipelines, particularly to evaluate the model's ability to classify specimens from well-represented taxa. Comprised only of samples labelled to species-level, it includes 67,267 barcodes from 1,653 arthropod species, across 500 different genera, with each species represented by at least 20 and at most 50 barcodes. The partition is further split into training (70%), testing (20%), and validation (10%) subsets. (*ii*) *Unseen*: This test partition was sampled to evaluate the models in real-world conditions where specimens from underrepresented species are frequently obtained. It only contains barcodes from "rare" species with fewer than 20 barcodes in the full reference dataset. Specifically, this partition contains 4,278 barcodes from 1,826 arthropod species, none of which are present in any other partition. Moreover, this partition contains all 500 genera labels present in the *Seen* partition, with up to 20 barcodes sampled per genus. The label distribution shifts are shown in Figure 1, with the *Seen* partition reflecting the overall dataset's distribution and the *Unseen* partition exhibiting a greater diversity of rare genera. (*iii*) *Pretrain*: This partition contains the remaining 893,744 barcode sequences from 14,794 invertebrate species across 6,679 genera. Note that only 35% of the sequences in this partition contain full taxonomic annotations up to the species level. The reader is referred to Appendix A for more details on dataset composition.

Proposed method: BarcodeBERT

Inspired by Bidirectional Encoder Representations from Transformers (BERT)-like models, which convert sequence inputs into meaningful embedding vectors, BarcodeBERT is designed to encode DNA barcodes into informative embedding vectors for fast and effective comparisons. This architecture's main building block is the transformer layer, with multi-head attention units playing a crucial role in capturing positional dependencies within each input sequence. Our model features four transformer layers, each with four attention heads, enabling a robust representation of the DNA barcode data while maintaining a manageable number of hyperparameters. Figure 2 shows the details of BarcodeBERT architecture.

Before being fed as input to the model, each barcode is split into a sequence of tokens. After evaluating two of the most common tokenization strategies for DNA sequences, Byte Pair Encoding (BPE) [34, 28] and k-mer tokenization [20, 9], we selected non-overlapping k-mer tokenization for BarcodeBERT (see the Ablation Studies section for more details). The token vocabulary includes all possible k-mer combinations derived from the nucleotide alphabet {`A`,`C`,`G`,`T`}, supplemented by two special tokens: `[MASK]` and `[UNK]`. The `[MASK]` token is utilized for masking k-mers during the pretraining phase, and k-mers containing any symbol that is not present in the nucleotide alphabet are assigned the `[UNK]` token. This results in a vocabulary size of $4^k + 2$.

A limitation of this tokenization strategy is its sensitivity to frame shifts. For example, the k-mer representation of the sequence `GATCGA` differs entirely from that of `CGATCGA`, even though the sequences differ by only a one-nucleotide shift. To address this issue and make our model robust to frame shifts that may occur in practice, we introduce a data augmentation step by randomly offsetting the sequence by a value ($0 \leq \text{offset} < k$) during pretraining to improve generalization. Before tokenization, DNA barcodes are either padded or truncated at 660 nucleotides to ensure coverage of the barcode region in the COI gene. Finally, the tokenized sequences are fed to the model and encoded into a sequence of 768-dimensional vectors.

Following self-supervised training, our model produces a whole barcode-level embedding vector by applying global average pooling over the sequence of d-dimensional output vectors, ignoring padding and any special tokens. During inference, the pipeline mirrors the training setup without the random offset: DNA barcodes are tokenized into non-overlapping k-mers and passed through the model, generating embeddings that capture meaningful taxonomic information across the entire sequence. BarcodeBERT is implemented using PyTorch and the Hugging Face Transformers library. 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 optimize the model parameters by gradient descent using the AdamW [21] optimizer with weight decay set to 1×10^{-5} and a OneCycle schedule with

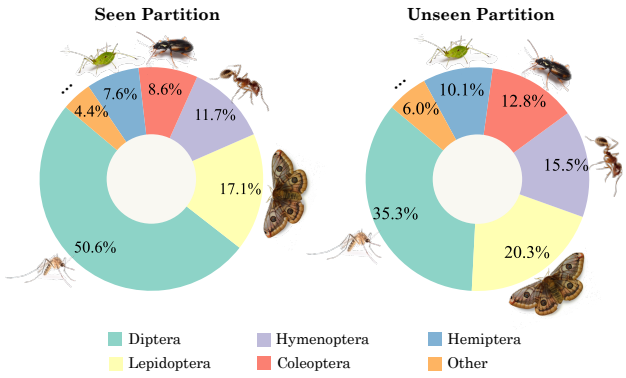

Fig. 1. Distribution of orders in the *Seen* (*left*) and *Unseen* (*right*) partitions of the dataset. Icons: CC BY-SA, Wikimedia; Pro Content license, Canva.

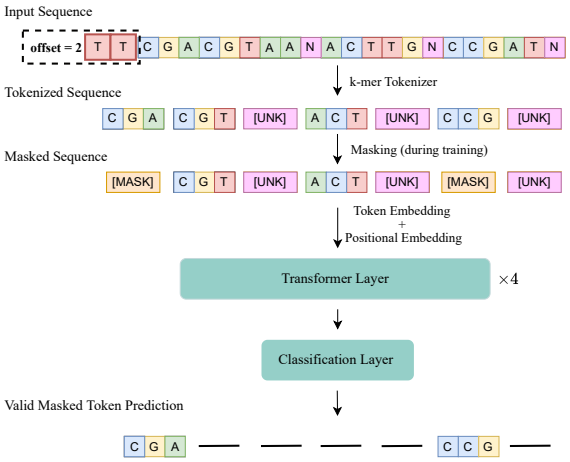

Fig. 2. Architecture of BarcodeBERT, a transformer-based model employing a self-supervised learning strategy. The model is trained on non-overlapping k-mers from DNA barcode sequences. Any k-mer containing a character that is not in the nucleotide vocabulary is replaced by the `[UNK]` token. Pretraining involves masking out certain input parts using the `[MASK]` token and predicting these masked elements using a linear classification head. During training, the model selects a random offset ($0 \leq \text{offset} < k$) for each sequence and begins tokenization from that position. This helps create more robust embeddings and increases resilience to potential mutations.

maximum learning rate of 1×10^{-4}. Additionally, we performed experiments across different k-mer lengths ($2 \leq k \leq 8$) to observe the impact of k-mer length on embedding quality.

Experiments

In this section, we present our evaluation framework and evaluate the performance of BarcodeBERT against the baseline models across several tasks. Additionally, we present a series of ablation studies to justify our design choices and analyze the impact of key hyperparameters on the model's performance.

Experimental setup

To explore the applicability of our model for DNA barcode-based biodiversity analyses, we employ different SSL evaluation strategies [4] and contrast its performance against the baselines. First, we evaluate our models in a "closed-world" setting where the goal is to classify DNA sequences into known taxa.

Fine-tuning. Pretrained models are fine-tuned on the training subset of the *Seen* partition and evaluated on the test subset. This task assesses the ability of models to perform species-level classification with full access to labelled training data.

Linear probing. To evaluate the quality of pretrained embeddings, the backbone of the models is frozen, and a linear classifier is trained on the training subset of the *Seen* partition. The final classifier is evaluated on the test subset, providing insights into the effectiveness of the embeddings without extensive task-specific training.

1-NN probing. This task evaluates model generalization to new species within known genera[1]. Using cosine similarity, we perform 1-NN probing at the genus level with the training subset of the *Seen* partition as the reference set and the *Unseen* partition as the query set.

Second, our models are evaluated in an "open-world" setting where the goal is to group sequences, including those from unknown species, based on shared features.

BIN reconstruction. We merge the test subset of the *Seen* partition with the *Unseen* partition and evaluate the model's ability to reconstruct Barcode Index Numbers (BINs) using embeddings generated without fine-tuning in a zero-shot clustering (ZSC) task [22]. This task assesses how well the embeddings capture the hierarchical structure of taxonomic relationships, including rare or unclassified species.

Finally, we evaluate the utility of learned DNA embeddings as auxiliary information in multi-modal learning.

Bayesian zero-shot learning. We selected a species-level image classification task using the INSECT dataset. DNA embeddings generated by the models are paired with pre-extracted image features to classify species in a zero-shot setup. We evaluate both embeddings from pretrained and fine-tuned models on the species classification task from the INSECT dataset. Following [2], the Bayesian zero-shot learning (BSZL) framework uses image features as priors and DNA embeddings as side information. For unseen species, the K-nearest seen species in the DNA embedding space are used to define local priors, allowing the Bayesian model to generate posterior predictive distributions for unseen categories. To ensure a fair comparison with prior work, image features are pre-extracted using ResNet-101 [17]. Hyperparameter tuning for the Bayesian model is performed using the same grid search space as in [2].

Results

In this section, we describe our results on two categories of evaluation tasks: DNA-specific evaluation tasks, designed to assess model performance in both open- and closed-world taxonomic settings; and zero-shot image classification using DNA embeddings.

DNA-specific categorization tasks

Our evaluation, presented in Table 1, compares several models across species-level and genus-level DNA-specific categorization tasks (fine-tuning, linear probing, 1-NN probing, BIN reconstruction through ZSC). For species-level classification, we performed a BLAST search for reference and obtained a 99.7% classification accuracy after the selection of the best hit. The performance of all fine-tuned deep learning-based models is comparable with this baseline, and all transformer models outperform the CNN model as well. DNABERT-2, DNABERT-S and BarcodeBERT achieved nearly identical accuracies over 99.7%. Notably, only BarcodeBERT continues to closely match BLAST's performance using a linear classifier, highlighting its strength in encoding meaningful features from raw data. In genus-level 1-NN probing, BarcodeBERT achieves the highest accuracy (78.5%) among the deep learning-based models, demonstrating a superior ability to generalize across taxonomic levels. BLAST, however, performs best in this task with 83.9%. This result indicates that, without fine-tuning, BarcodeBERT captures coarser taxonomic distinctions but is limited in representing the full hierarchical taxonomic structure as illustrated in Figure 3. The ZSC task provides additional insights into the model's understanding of the hierarchical taxonomic structure. High performance in ZSC alone indicates a learned representation's ability to finely distinguish between

[1] While 1-NN probing could be considered an open-world task since species are unseen, it is included in the closed-world setting because the genera are part of the seen taxonomy.

Table 1. Classification accuracy of DNA barcode models under different SSL evaluation strategies and different efficiency metrics. The baselines are divided into three groups: alignment-based techniques, BLAST; a deep learning-based non-SSL CNN baseline; and off-the-shelf DNA foundation models pretrained on non-barcode data. These are compared against BarcodeBERT, which is specifically pretrained on DNA barcode-based datasets. For BarcodeBERT we used the best configuration of $k = 4$, with 4 attention heads, and 4 layers (4-4-4). Some models supported variable stride length; for these, the numbers in parentheses are the optimal k-mer values that yield the best results. We also show the throughput-per-second (TPS) of the encoders, and the total duration of the classification tasks. Numbers in **boldface** indicate the best result across each task, and underlined indicates second place.

			Species-level acc (%) of seen species			Genus-level 1-NN probe of unseen species		BIN reconstruction accuracy (%)
Model	#Param.	TPS (seq/s)	Finetuned	Linear probe	Dur (s)	Acc (%)	Dur (s)	ZSC probe
BLAST	N/A	N/A	**99.7***		1495	**83.9**	602	N/A
CNN encoder	1.8 M	934	98.2	51.8	13	47.0	55	26.8
DNABERT	88.1 M	50	(k=6) 99.5	(k=4) 47.1	248	(k=6) 48.1	1021	**79.3**
DNABERT-2	118.9 M	134	**99.7**	87.2	101	23.5	381	38.1
DNABERT-S	117.1 M	134	**99.7**	93.1	101	30.6	381	62.7
HyenaDNA-tiny	1.6 M	**1167**	99.2	93.5	**11**	37.5	**44**	25.8
Nucleotide Transformer	55.9 M	95	99.5	65.1	140	40.1	536	22.4
BarcodeBERT (4-4-4)	29.1 M	484	**99.7**	**99.0**	27	78.5	108	73.2

*BLAST is a deterministic algorithm without any learning component (see Appendix C for details). Consequently, species classification accuracy does not correspond to fine-tuning or linear probing, and it is only included in the table for reference.

closely related clusters (BINs) without necessarily capturing the higher-level taxonomy. In contrast, strong 1-NN performance at higher taxonomic levels but lower ZSC accuracy suggests that the model understands the overall topology of the hierarchical taxonomic structure, even if it lacks the granularity needed for precise clustering. DNABERT and BarcodeBERT exhibit this distinction, with BarcodeBERT achieving a more balanced performance across tasks, making it the more versatile model for comprehensive DNA barcode analysis.

Two efficiency measurements are included: throughput, defined exclusively for deep-learning-based models as the number of sequences processed per second, and total runtime for classification pipelines to ensure a fair comparison with alignment-based baselines. In terms of throughput, HyenaDNA(*tiny*) showcases the capabilities of state space models, demonstrating high throughput with fewer parameters. However, its classification performance is lower compared to BarcodeBERT and DNABERT-2. In total run time, our results indicate that subquadratic methods like the CNN baseline and HyenaDNA perform genus-level similarity searches (1-NN probe) 13× faster than BLAST, while BarcodeBERT is 5.5× faster than BLAST and outperforms other transformer models at this task. For species-level classification pipelines that do not include the computation of the training embeddings, transformer-based models demonstrate clear advantages over traditional baselines in terms of running time. Notably, BarcodeBERT, with a moderate parameter count, matches BLAST's high classification accuracy (99.7%) with a 55× faster running time, thus providing a well-rounded option for large-scale DNA barcode applications. All efficiency experiments were conducted using an Intel(R) Xeon(R) CPU @ 2.20GHz processor and a Quadro RTX 6000 GPU.

Zero-shot image classification using DNA embeddings

We use the Bayesian zero-shot learning task to evaluate the quality of the DNA feature embeddings, by assessing their effectiveness when used as side information for classifying images to species on the INSECT dataset. We consider the embeddings directly from the pretrained models and also after fine-tuning. The accuracy for seen and unseen test species and the harmonic mean are presented in Table 2. Even without fine-tuning, BarcodeBERT substantially outperforms DNABERT and DNABERT-2 on unseen species, regardless of whether they had been fine-tuned previously. BarcodeBERT achieves similar performance to the reported baseline CNN results [2] and improves on the harmonic mean score by 1.2% and unseen accuracy by 1.9%, respectively. Our results demonstrate that in the zero-shot learning task of predicting insect species, employing BERT-like models that have also been trained on insect DNA barcodes as DNA encoders can improve performance over CNNs and general DNA foundation models.

Ablation Studies

Here we review the impact of the different components involved during pretraining. We use the terminology *context tokens* for tokens that are left unchanged to provide context to the model during pretraining and the terminology *substitution tokens* for tokens that will be changed as part of the masked language modelling task. We consider different strategies

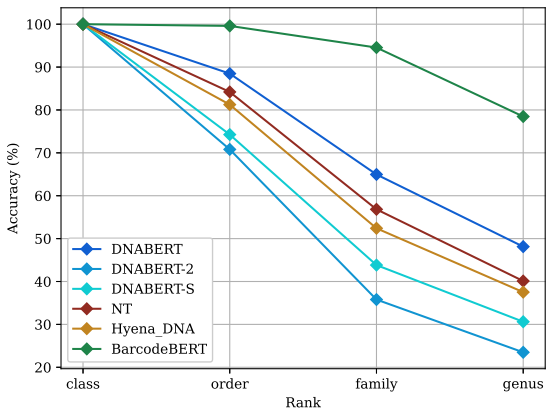

Fig. 3. Comparison of different DNA foundation models on the task of 1-NN probing at different taxonomic levels. The query set contains DNA barcodes from species not present in the key set, and none of the models have undergone fine-tuning.

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 for 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). Numbers in **boldface** indicate the best result across each task, and underlined second best.

	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)	Invertebrates	–	31.6	**20.0**	24.5
BarcodeBERT (ours)	Invertebrates	Insect	38.8	15.3	22.0

to calculate the loss of each group separately. The loss associated with predicting contextual tokens is referred to as the "context component" of the loss[2], while the loss related to predicting substitution tokens is the "substitution component". By assigning different weights to these two loss components, we sought to observe how these adjustments would affect both training and evaluation. In particular, we define the following terms: r_s as the substitution token ratio, $r_{\texttt{[MASK]}}$ as the proportion of substitution tokens assigned to the `[MASK]` token, $r_{\texttt{[RAND]}} = 1 - r_{\texttt{[MASK]}}$ as the proportion of the substitution tokens assigned a random valid token (all tokens except the special tokens), and w_s as the penalty weight given to the substitution component of the loss. Note that $1 - w_s$ is always the weight of the context component of the loss.

Substitution token rate

To examine how varying the substitution token ratio (r_s) affects performance, we tested several ratios, keeping the model architecture (4 attention heads, 4 layers), tokenization ($k = 4$), substitution penalty weight ($w_s = 1$), and masking strategy ($r_{\texttt{[MASK]}} = 1$) constant. Table 3 shows that species-level classification performance remains consistently high across substitution rates, peaking at 99.67% accuracy with 45% and 50% substitution tokens. Linear probe results align closely, reaching the highest accuracy of 99.02% at the 50% substitution rate.

For genus-level 1-NN probing of unseen species, the 50% substitution rate yields the best accuracy at 78.47%, suggesting that this rate provides a balance that strengthens the model's ability to generalize to new taxa. Lower substitution rates show slightly reduced generalization, while a 60% rate begins to degrade performance, indicating that 50% is the optimal value for r_s.

Weight of the substitution component of the loss

In this study, building on the fact that predicting context tokens is inherently easier than predicting substitution tokens for LLMs, we investigated how adjusting the penalty weights between these two tasks affects the performance of the model. For this purpose, we experimented with different penalty weights assigned to the substitution component of the loss (w_s). Table 4 provides the accuracy for genus-level 1-NN probing of unseen species across different values for w_s in combination with four k-mer sizes (2, 4, 6, and 8) and BPE tokenizer obtained from DNABERT-2. Alternative BPE tokenizers specifically fit to our data are investigated later in this section. We kept the architecture (4 layers, 4 attention heads), substitution token rate ($r_s = 50\%$), and masking strategy ($r_{\texttt{[MASK]}} = 1$) constant. As highlighted in Table 4, the optimal performance across all k-mer sizes was achieved with a w_s of 1.0, where the highest accuracy, 78.47%, was observed with $k = 4$. Our experiments indicate that focusing the

[2] Although predictions in the foundation model literature are typically restricted to substitution tokens, we extend this to include context token predictions, maintaining the terminology to explore their potential utility in input reconstruction similar to non-masked autoencoding objectives.

Table 3. Classification accuracy over the different substitution token ratios r_s, while keeping constant all of the model architecture (4-4), the value of $k = 4$ during tokenization, the proportion of the substitution tokens assigned to `[MASK]` ($r_{\texttt{[MASK]}} = 1$) and the penalty weight for the substitution component of the loss ($w_s = 1$). Numbers in **boldface** indicate the highest accuracies and the → *highlighted* ← value shows the selected optimal parameter.

Substitution token ratio (%)	Species-level acc (%) of seen species		Genus-level acc (%) of unseen species
	Finetuned	Linear probe	1-NN probe
15	98.95	98.95	75.15
30	98.79	98.79	74.24
45	**99.67**	98.54	74.42
→ *50* ←	**99.67**	**99.02**	**78.47**
60	99.62	98.45	77.56

Table 4. Genus-level accuracy for 1-NN probing of unseen species with varying penalty weight assigned to the substitution component of the loss (w_s). The model architecture remains fixed (4-4), with substitution token ratio ($r_s = 50\%$) and masking strategy ($r_{\texttt{[MASK]}} = 1$). Two tokenizers were tested: a k-mer tokenizer with k-mer sizes of 2, 4, 6, and 8, and a BPE tokenizer used in DNABERT-2 with a fixed vocabulary size of 4096. Note that the weight of the context component of the loss always equals $1 - w_s$. The number in **boldface** indicates the overall best accuracy, underlined the best per tokenizer, and → *highlighted* ← the selected optimal parameter.

	Genus-level acc (%) of unseen species with 1-NN probe				
Loss weight (w_s)	$k=2$	$k=4$	$k=6$	$k=8$	BPE
0.2	64.18	76.06	75.15	71.15	70.57
0.5	66.47	74.98	76.62	71.22	70.34
0.8	68.84	76.71	74.66	73.33	69.40
0.9	69.51	77.16	76.06	72.23	67.48
→ *1.0* ←	76.92	**78.47**	75.74	75.62	69.85

loss penalty only on the harder task of predicting substitution tokens, while not penalizing the easier task of predicting context tokens, yields the best accuracy. This aligns with observations in other foundation models, such as BERT and DNABERT.

Masking strategies

In this experiment, we adapted our masking strategy to mimic BERT's original methodology. BERT addresses the gap between pretraining and fine-tuning by employing a masking strategy whereby 80% of substitution tokens are replaced by `[MASK]` tokens, 10% are replaced with random tokens, and 10% remain unchanged. This approach ensures more robust embeddings during testing, where masked tokens are absent. We incorporated this methodology into BarcodeBERT and explored various ratios for token replacement and three different values for w_s. In the first case, based on the results in the previous section, where $w_s = 1$ had the best performance, we kept $w_s = 1$ and adjusted $r_{\texttt{[MASK]}}$. In the second case, we set w_s to 0.95 and in the third case, w_s was set to 0.90, to closely replicate BERT's original strategy that keeps 10% of the tokens unchanged. Note that in all experiments $r_{\texttt{[rand]}}$ was set to $1 - r_{\texttt{[MASK]}}$. In this study, we used the best configuration of 4 layers and 4 attention heads, $k = 4$, and $r_s = 50\%$. Figure 4 presents the accuracy of these experiments for genus-level 1-NN probing on unseen species. The results show that for $w_s = 1.0$, the best accuracy is 78.47% with $r_{\texttt{[MASK]}} = 1.0$, for $w_s = 0.95$ the best accuracy is 76.85% with $r_{\texttt{[MASK]}} = 0.9$, and for $w_s = 0.9$, $r_{\texttt{[MASK]}} = 0.5$ gives the best accuracy of 78.14%, which improves the accuracy by 1% compared to the case where $w_s = 0.9$ and $r_{\texttt{[MASK]}} = 1.0$. Our results demonstrate that adopting BERT's masking strategy did not enhance the performance of BarcodeBERT, indicating that maintaining $r_{\texttt{[MASK]}} = 1$ is the optimal configuration.

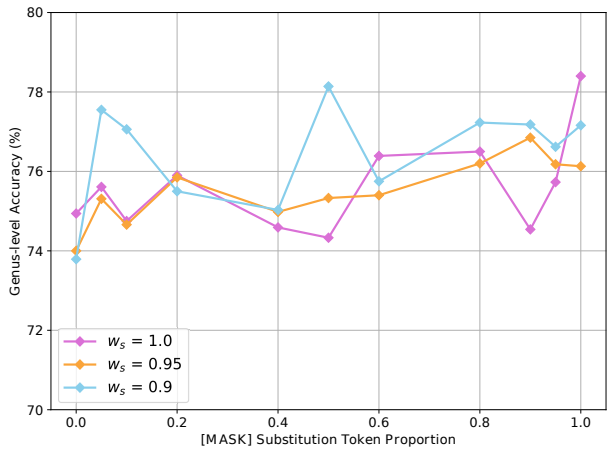

Fig. 4. Genus-level accuracy for 1-NN probing of unseen species across different values of $r_{\texttt{[MASK]}}$, $r_{\texttt{[RNAD]}}$, and w_s. Experiments were conducted using the optimal configuration: 4 layers, 4 attention heads, k= 4, with substitution token ratio ($r_s = 50\%$).

Tokenization strategies

Besides using a k-mer-based tokenizer that creates a vocabulary of fixed size, we also evaluated the Byte Pair Encoding (BPE) tokenizer, which utilizes a fixed-sized vocabulary of variable-length tokens based on the co-occurrence frequency of the characters [30]. The BPE tokenizer is based on a data compression algorithm of the same name [13], which was later adapted to overcome the limitations of fixed vocabularies by tokenizing text at the subword level instead of the word level [30]. It also inherently compresses the sequences, reducing the amount of information required to represent them [14]. Furthermore, BPE has some potential advantages over the k-mer tokenizer when applied to DNA sequences. Unlike overlapping k-mers, BPE tokenization avoids the problem of masked token information being leaked by adjacent tokens, making it more suitable for masked language modelling [34].

BPE tokenization was utilized by both DNABERT-2 and DNABERT-S. In DNABERT-2, the BPE tokenizer was trained on 2.75 billion nucleotide bases from the nuclear human genome and 32.49 billion nucleotide bases from 135 other species across various kingdoms, with a vocabulary size of 4,096 [34]. We employed the same BPE tokenizer from DNABERT-2 within our BarcodeBERT architecture and evaluated its effectiveness across different training scenarios. Additionally, we trained our own BPE tokenizers with a range of vocabulary sizes using the DNA barcode dataset.

Loss weight. Recall from Table 4 that we evaluated the impact of loss weights on the BarcodeBERT model using both the k-mer tokenizer and the BPE tokenizer from DNABERT-2. Our results indicate that increasing the loss weight from 0.2 to 1.0 does not result in a significant improvement in the accuracy of genus-level 1-NN probing on unseen species for BarcodeBERT using the BPE tokenizer. Since a loss weight of 1.0 yielded the best results for the k-mer tokenizer, we used this loss weight in all subsequent experiments. Additionally, for these experiments, we followed the configuration outlined in Table 4, maintaining a consistent substitution token ratio (r_s) of 50% and assigning the entire substitution token proportion to `[MASK]` by setting $r_{\texttt{[MASK]}}$ to 1.

Model size. We evaluated different model sizes in use with the DNABERT-2 BPE and BarcodeBERT BPE tokenizers, reporting the accuracy of genus-level 1-NN probing on unseen species in Table 5. We consider three configurations: 4 layers with 4 heads, 6 layers with 6 heads, and 12 layers with 12 heads. Across different configurations, we did not observe any significant change in accuracy for DNABERT-2 BPE. However, for BarcodeBERT BPE, we see that increasing the model size reduces the accuracy across different vocabulary sizes, possibly due to overfitting.

Vocabulary size. We examine how varying the vocabulary size (v) affects the performance of BarcodeBERT BPE. To address differences in k-mer frequencies between our dataset and those used in DNABERT-2, we trained several BPE tokenizers from scratch with different vocabulary sizes on the DNA barcode dataset. Unlike k-mer tokenizers, in the BPE tokenizer, the length of the tokenized sequence is not determined by the nucleotide sequence length and could vary depending on the composition of the input sequence. Therefore, we need to pad (using `[UNK]` token) or truncate the tokenized sequences to a maximum length before passing them to the BarcodeBERT model. Since BPE with smaller v tends to generate longer tokenized sequences on average (see Supplementary Table S6), we set the maximum sequence length to 128 for larger vocabularies ($v = 4096$ and $v = 1024$) and 256 for smaller vocabularies ($v = 256$ and $v = 128$). In Table 5, we report the accuracy of genus-level 1-NN probing on unseen species for different BPE tokenizers. Our results show that

Table 5. Genus-level accuracy for unseen species with different tokenizers, various model sizes, fixed weight for the substitution component of the loss function ($w_s = 1$), substitution token ratio ($r_s = 50\%$) and substitution token proportion assigned to `[MASK]`, ($r_{\texttt{[MASK]}} = 1$). Two types of tokenizers were tested: a k-mer tokenizer with k-mer sizes of 2, 4, 6, and 8, and five different versions of BPE tokenizers. The BPE tokenizer obtained from DNABERT-2 with a fixed vocabulary size of 4096, and BPE tokenizers trained on our dataset with vocabulary sizes (v) of 4096, 1024, 256, and 128. Numbers in **boldface** indicate the best result across each architecture.

	k-mer tokenizer				DNABERT-2 BPE	BarcodeBERT BPE			
Model size	$k=2$	$k=4$	$k=6$	$k=8$	$v=4096$	$v=4096$	$v=1024$	$v=256$	$v=128$
4 layers, 4 heads	76.92	**78.47**	75.74	75.62	69.85	66.88	68.58	66.57	63.42
6 layers, 6 heads	71.46	**76.95**	76.04	76.60	70.17	67.30	66.95	63.49	60.61
12 layers, 12 heads	74.71	70.17	70.80	**75.81**	68.68	67.79	62.39	56.94	54.09

reducing the vocabulary size from 4,096 to 128 leads to a consistent decrease in accuracy across all model sizes.

Comparing k-mer with BPE. Our results suggest that k-mer tokenizers outperform BPE tokenizers in all model configurations, likely due to the following reasons. First, DNA barcode sequences are too short to benefit from the compression advantages of BPE. Second, BPE is sensitive to minor variations such as single-character substitutions, which can cause a cascade of changes to resulting tokens in the rest of the sequence; this makes it unsuitable for DNA datasets that inherently contain single-nucleotide mutations and ambiguous positions. In other words, even for similar sequences with small Hamming distances between them, BPE tokenization can produce completely different tokenized representations, whereas k-mer tokenization remains more consistent since a single-nucleotide substitution only changes a single token (see Supplementary Figure S6). Third, although BPE has the advantage that it handles small sequence alignment shifts better than k-mer tokenizers, we can overcome this limitation in k-mer tokenizers by using data augmentations with random offsets during pretraining (see Supplementary Table S7).

Discussion

Our results demonstrate that pretraining masked language models on DNA barcode data, as exemplified by BarcodeBERT, is highly effective for arthropod species identification. Our pretrained model performs well on various downstream tasks common in biodiversity analyses, such as taxonomic classification, clustering, and similarity searches. BarcodeBERT excels in these tasks because it efficiently uses hardware acceleration, enabling it to scale effectively for large datasets, while being faster (55x) than alignment-based approaches in species level classification. By systematically evaluating tokenization and masking strategies, we also provide actionable insights for the pretraining of DNA-specific foundation models.

Despite its strengths, BarcodeBERT has some limitations. For instance, its training data may have taxonomic and geographical biases. In particular, the model is trained exclusively on invertebrate species from Canada, potentially limiting its applicability in global studies. In this context, the BOLD dataset, now comprising more than 16 million DNA barcodes from a wide geographical distribution [27], represents a wealth of untapped data that could address these biases. Future work should incorporate more diverse datasets to develop robust, globally scalable self-supervised models for taxonomic classification. The methodology and findings we present should be broadly applicable to barcode regions for other kingdoms, such as the ITS region for fungi [29], however the method is yet to be validated beyond the COI barcode region used for Animalia.

Lastly, while genomic sequences longer than barcodes could offer deeper insights for specialized phylogenetic analyses, the quadratic time complexity of transformer models limits their applicability to such sequences. Future work should include more parameter-efficient architectures such as Structured State Space Models, which scale sub-quadratically with sequence length [16]. These architectures may be pretrained on specific barcode datasets to provide a more efficient alternative for applications involving longer sequences.

Conclusions

BarcodeBERT leverages 1 million DNA barcodes with partial taxonomic annotations to outperform state-of-the-art foundation models in genus-level and species-level classification tasks. Notably, BarcodeBERT matches the high accuracy of the alignment-based classification tool BLAST in species classification, while being 55 times faster and more scalable. In addition, our extensive analysis of pretraining strategies provides actionable insights for building customized DNA language models for large-scale taxonomic classification.

Overall, BarcodeBERT's performance demonstrates how transformer-based architectures can be successfully customized to overcome the challenges of genomic biodiversity data, for effective DNA barcode identification and classification. Lastly, not being limited to a specific dataset or barcode region, our model is highly amenable to future applications, to global datasets or barcode regions in other kingdoms of life.

Competing interests

No competing interest is declared.

Author contributions statement

PMA curated the dataset, removed invalid entries, created preliminary data split files with the assistance of GWT and DS, and partitioned the data into its final form, with assistance from MS. PMA, SCL, MS, and NS worked on BarcodeBERT's code implementation. PMA conducted all the DNA baseline experiments. MS conducted all the masking and loss penalties ablation studies. NS conducted all tokenization ablation studies. ZMG and ATW conducted the multimodal retrieval learning experiments, with the assistance of PMA. PMA, MS, NS, LK, and IZ authored the manuscript text and figures. SCL, GWT, AXC, DS, LK, and JBH provided guidance on experimental design and edited the manuscript. All authors reviewed the manuscript.

Acknowledgments

We acknowledge the support of the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFT-2020-00073]. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through the Canadian Institute for Advanced Research (CIFAR), and companies sponsoring the Vector Institute (`http://www.vectorinstitute.ai/#partners`). GWT acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs program, and the Canada CIFAR AI Chairs program. LK acknowledges support from NSERC Discovery Grant RGPIN-2023-03663. DS acknowledges support from the Canada First Research Excellence Fund to the University of Guelph's "Food From Thought" research program [Project 000054]. The funders had no role in the preparation of the manuscript.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of molecular biology*, 215(3):403–410, 1990.
2. 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.
3. S. Badirli, C. J. Picard, G. Mohler, F. Richert, Z. Akata et al. Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, 14(6):1515–1530, 2023.
4. R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar et al. A cookbook of self-supervised learning, 2023. arXiv: 2304.12210.
5. Y. Basset, L. Cizek, P. Cuénoud, R. K. Didham, F. Guilhaumon et al. Arthropod diversity in a tropical forest. *Science*, 338(6113):1481–1484, 2012.
6. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
7. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
8. G. Corso, R. Ying, M. Pándy, P. Veličković, J. Leskovec et al. Neural distance embeddings for biological sequences. *arXiv preprint arXiv:2109.09740v2*, 2021.
9. H. Dalla-Torre, L. Gonzalez, J. M. Revilla, N. L. Carranza, A. H. Grzywaczewski et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.
10. J. R. deWaard, S. Ratnasingham, E. V. Zakharov, A. V. Borisenko, D. Steinke et al. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*, 6(1):308, Dec 2019.
11. A. Dopheide, L. K. Tooman, S. Grosser, B. Agabiti, B. Rhode et al. Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications*, 29(4):e01877, 2019.
12. V. Fishman, Y. Kuratov, M. Petrov, A. Shmelev, D. Shepelin et al. GENA-LM: A family of open-source foundational models for long DNA sequences. *bioRxiv*, 2023.
13. P. Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.
14. M. Gallé. Investigating the effectiveness of BPE: The power of shorter sequences. In *Conference on Empirical Methods in Natural Language Processing*, 2019.
15. Z. Gong, A. T. Wang, J. B. Haurum, S. C. Lowe, G. W. Taylor et al. BIOSCAN-CLIP: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv preprint arXiv:2405.17537*, 2024.
16. A. Gu, K. Goel, and C. Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.
17. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
18. 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.
19. P. D. N. Hebert, S. Ratnasingham, E. V. Zakharov, A. C. Telfer, V. Levesque-Beaudin et al. Counting animal species with DNA barcodes: Canadian insects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 371(1702), Sept. 2016.
20. 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.
21. I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.
22. S. C. Lowe, J. B. Haurum, S. Oore, T. B. Moeslund, and G. W. Taylor. An empirical study into clustering of unseen datasets with self-supervised encoders. *arXiv preprint arXiv:2406.02465*, 2024.
23. 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.
24. S. E. Miller. Dna barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences*, 104(12):4775–4776, 2007.
25. E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.
26. M. Poli, P. Molchanov, M. Pellat, G. Izacard, J. Liu et al. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.
27. 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.
28. L. Romeijn, A. Bernatavicius, and D. Vu. MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources*, 24(8):e14006, 2024. e14006 MER-24-0165.R1.
29. C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, 109(16):6241–6246, 2012.

30. R. Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
31. A. Srivathsan, L. Lee, K. Katoh, E. Hartop, S. N. Kutty et al. Ontbarcoder and minion barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC biology*, 19(1):217, September 2021.
32. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles et al. 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.
33. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017.
34. Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genome, 2023. arXiv: 2306.15006.
35. Z. Zhou, W. Wu, H. Ho, J. Wang, L. Shi et al. DNABERT-S: Learning species-aware DNA embedding with genome foundation models, 2024.