ENHANCING DNA FOUNDATION MODELS TO ADDRESS MASKING INEFFICIENCIES

Monireh Safari*
University of Waterloo, Canada

Pablo Millan Arias*
University of Waterloo, Canada

Scott C. Lowe
Vector Institute, Canada

Lila Kari†
University of Waterloo, Canada
lila@uwaterloo.ca

Angel X. Chang
Simon Fraser University
Alberta Machine Intelligence Institute (Amii), Canada

Graham W. Taylor†
University of Guelph
Vector Institute, Canada
gwtaylor@uguelph.ca

ABSTRACT

Masked language modelling (MLM) as a pretraining objective has been widely adopted in genomic sequence modelling. While pretrained models can successfully serve as encoders for various downstream tasks, the distribution shift between pretraining and inference detrimentally impacts performance, as the pretraining task is to map [MASK] tokens to predictions, yet the [MASK] is absent during downstream applications. This means the encoder does not prioritize its encodings of non-[MASK] tokens, and expends parameters and compute on work only relevant to the MLM task, despite this being irrelevant at deployment time. In this work, we propose a modified encoder-decoder architecture based on the masked autoencoder framework, designed to address this inefficiency within a BERT-based transformer. We empirically show that the resulting mismatch is particularly detrimental in genomic pipelines where models are often used for feature extraction without fine-tuning. We evaluate our approach on the BIOSCAN-5M dataset, comprising over 2 million unique DNA barcodes. We achieve substantial performance gains in both closed-world and open-world classification tasks when compared against causal models and bidirectional architectures pretrained with MLM tasks. The code repository is available at https://github.com/bioscan-ml/BarcodeMAE.

1 INTRODUCTION

DNA foundation models have emerged as effective tools for analyzing genomic sequences, utilizing a wide variety of architectures, including transformers (Ji et al., 2021; Millan Arias et al., 2023; Zhou et al., 2024), state space models (SSMs) (Poli et al., 2023; Gao & Taylor, 2024), and convolutional neural networks (CNNs) (Benegas et al., 2023). These models leverage different pretraining strategies, from causal to bidirectional learning, enabling strong performance across diverse genomic tasks. Among these pretraining strategies, masked language modelling (MLM) has become widely adopted, enabling models to learn effective sequence representations for downstream tasks like specimen identification to taxon and species discovery. However, the effectiveness of MLM is highly dependent on how masking is implemented, as different strategies can affect the model's performance.

In DNA sequence modelling, foundation models typically adopt BERT's three-part masking strategy (Devlin et al., 2019), where 80% of selected tokens are replaced with [MASK], 10% remain unchanged, and 10% are randomly substituted. Models such as the Nucleotide Transformer (Dalla-Torre et al., 2024) and BarcodeBERT (Millan Arias et al., 2023) followed this approach, while DNABERT

*Joint first author
†Corresponding author

(Ji et al., 2021) and DNABERT-2 (Zhou et al., 2023) adopted a simpler strategy, replacing 100% of selected positions with `[MASK]` tokens. Despite its popularity, MLM introduces a notable limitation: a distribution shift between pretraining and inference due to the absence of `[MASK]` tokens during downstream tasks. This mismatch leads to representational inefficiencies, as models prioritize the quality of encodings and predictions corresponding to `[MASK]` tokens but lack a direct target for non-`[MASK]` token inputs. Consequently, they allocate parameters and compute to tokens never encountered during inference, potentially limiting their ability to capture biologically relevant patterns. While this limitation and its impact on model performance have been studied in natural language processing (NLP) settings (Meng et al., 2024; Clark et al., 2020), its effects on DNA sequence foundation models remain unexplored.

In this study, we propose BarcodeMAE which uses a modified encoder-decoder architecture based on the masked autoencoder for MLM (MAE-LM; Meng et al., 2024). BarcodeMAE is designed to address the MLM inefficiency with BERT-style transformer models for biodiversity analysis using DNA barcodes. This approach eliminates `[MASK]` tokens during encoding, thereby mitigating the distribution shift between pretraining and inference. Computation and parameters needed to predict values for `[MASK]` tokens is isolated to the decoder block, which is discarded after pretraining and not called at inference time. We empirically show that this mismatch is particularly detrimental in genomic pipelines where models are used for feature extraction without fine-tuning. To evaluate our model, we conduct self-supervised pretraining on the BIOSCAN-5M dataset (Gharaee et al., 2024), which comprises over 2 million unique DNA barcodes. Our model outperforms existing foundation models in genus-level classification, surpassing a comparable encoder-only architecture by over 10 percentage points. Although it does not achieve the highest performance in BIN reconstruction, BarcodeMAE demonstrates superior average performance across evaluation tasks.

2 METHOD

In this section, we first describe the BIOSCAN-5M dataset and its partitioning scheme. Next, we introduce BarcodeMAE, our proposed model that adapts the masked auto-encoder architecture to address the representational limitations of masking approaches for DNA foundation models.

2.1 DATA

Our analysis utilizes the BIOSCAN-5M dataset[1], a comprehensive collection of 2.4 M unique DNA barcodes organized into three distinct partitions: *(i) Pretrain*: Contains 2.28 M unique DNA barcodes from unclassified specimens, used for self-supervised pretraining. *(ii) Seen*: Encompasses DNA barcodes with validated scientific species names, split into training (118 k barcodes), validation (6.6 k barcodes), and test (18.4 k barcodes) subsets for closed-world evaluation tasks. *(iii) Unseen*: Contains novel species with reliable placeholder taxonomic labels, distributed across reference key (12.2 k barcodes), validation (2.4 k barcodes), and test (3.4 k barcodes) subsets for open-world evaluation tasks. For each sample in *unseen*, its species does not appear in *seen*, but its genus does appear. This structure enables the evaluation of both closed-world classification and open-world species identification capabilities.

2.2 BARCODEMAE: A MASKED AUTO-ENCODING MODEL FOR DNA BARCODE SEQUENCES

In this section, we present BarcodeMAE, the encoder-decoder architecture for DNA sequence modelling. We describe the core architectural design and masking strategy that addresses the distribution shift between pretraining and inference, followed by the implementation specifications: tokenization, positional embeddings, and training procedures.

2.2.1 ENCODER-DECODER ARCHITECTURE WITH MODIFIED MASKING

To address the representational inefficiency in DNA sequence modelling, we adapt the MAE-LM approach (Meng et al., 2024) for genomic applications. In training using masked language modelling

[1]The BIOSCAN-5M dataset contains 5.15 M arthropod records, each with associated an image and DNA barcode sequence. Although the images are different for each record, the same barcode can occur across multiple records, hence there are fewer than 5 M unique barcodes.

objectives, part of the encoder's capacity must be allocated to processing [MASK] tokens, which potentially limits the model's overall representational capacity to encode real tokens. The MAE-LM architecture effectively mitigates this limitation by using a bidirectional encoder and shallow bidirectional decoder, where the masked tokens are only presented to the decoder.

The encoder operates on nucleotide sequences with masked-out tokens removed entirely. Given a DNA sequence $\mathbf{x} = [x_1, \ldots, [\texttt{MASK}]_i, \ldots, x_n]$ and the set of masked positions $\mathcal{M}$, the encoder processes only nucleotide tokens. The encoder's input sequence $\mathbf{H}^0$ is composed of token embeddings $\mathbf{e}_{x_i}$ and positional embeddings $\mathbf{p}_i$ for non-masked positions:

$$\mathbf{H}^0 = \{h_i^0\}_{i \notin \mathcal{M}}, \quad h_i^0 = \mathbf{e}_{x_i} + \mathbf{p}_i \tag{1}$$

The decoder then processes sequences containing both masked and unmasked positions, explicitly incorporating the [MASK] token in its input. The decoder's input sequence $\hat{\mathbf{H}}^0$ is constructed as:

$$\hat{\mathbf{H}}^0 = \{\hat{h}_i^0\}_{1 \leq i \leq n}, \quad \hat{h}_i^0 = \begin{cases} \mathbf{e}_{[\text{MASK}]} + \mathbf{p}_i & i \in \mathcal{M} \\ h_L^i + \mathbf{p}_i & i \notin \mathcal{M} \end{cases} \tag{2}$$

where h_L^i represents the final layer output from the encoder for non-masked positions and $e_{[\text{MASK}]}$ is the token embedding for the [MASK] token.

This approach prevents the encoder from learning specific embeddings for the [MASK] token, ensuring the decoder's representational capacity is not devoted to encoding this special token. Consequently, the encoder's representations remain unaffected by the [MASK] token and will use the full representational capacity to learn meaningful patterns from the nucleotide sequences. During downstream tasks, only the encoder is utilized, effectively isolating any potential limitations or inefficiencies related to the [MASK] tokens.

2.2.2 Model implementation

BarcodeMAE uses a transformer architecture to implement the MAE-LM framework for DNA barcodes. It is trained using masked language modelling objectives. The architecture consists of a symmetrical design: an encoder and decoder, each comprising 6 transformer layers with 6 attention heads. Both components maintain a consistent hidden dimension of 768 units to ensure uniform representation capacity throughout the network. To obtain an embedding of the entire DNA barcode, the model employs global average pooling across the sequence of 768-dimensional output vectors, excluding padding and special tokens. Figure 1 illustrates the architectural differences between BarcodeBERT, an encoder-only foundation model, and BarcodeMAE, an encoder-decoder model.

For DNA sequence processing, we use non-overlapping k-mer tokenization with a vocabulary size of $4^k + 2$, including the [UNK] and [MASK] special tokens. To handle frame-shift sensitivity, we incorporate the data augmentation strategy proposed in BarcodeBERT (Millan Arias et al., 2023), where sequences are randomly offset before tokenization. Based on previous studies (Millan Arias et al., 2023; Dalla-Torre et al., 2024) showing optimal performance with k values of 4 or 6, we evaluate our model using both of these k-mer lengths.

In this model, the encoder processes DNA sequences without [MASK] tokens, requiring a modified positional encoding scheme. Our implementation preserves sequence order by skipping masked position indices during encoding. This design maintains the relative positions of unmasked tokens from the original sequence, enabling spatial relationship modelling in DNA sequences.

We implement our model using PyTorch and the Hugging Face Transformers library. Our model is trained using masked token prediction with a 50% token masking strategy. To optimize the cross-entropy loss of masked tokens, we use AdamW (Loshchilov & Hutter, 2017) with a weight decay coefficient of 1×10^{-5} and a OneCycle scheduler with a maximum learning rate of 1×10^{-4}.

3 Experiments

In this section, we describe both closed-world and open-world evaluation tasks designed to assess different aspects of the model's performance. Additionally, we present comparative results against current state-of-the-art baselines and conclude with an ablation study examining the impact of k-mer length and the number of layers.

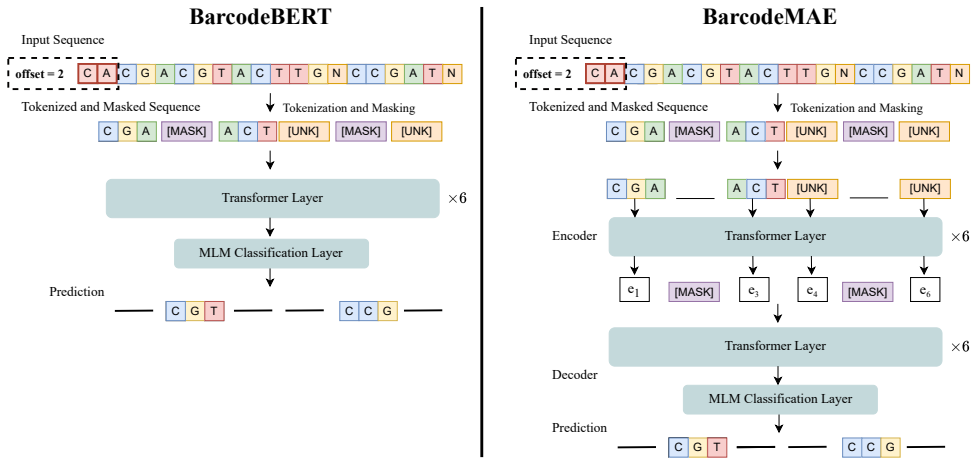

Figure 1: *Comparison of pretraining processes for BarcodeBERT (left) and BarcodeMAE (right).* BarcodeBERT uses an encoder-only transformer architecture with direct masking. BarcodeMAE processes DNA barcode sequences through a transformer encoder-decoder architecture. The masking strategy differs from other foundation models by excluding the `[MASK]` token from the encoder input, requiring the decoder to predict masked sequences. After pretraining, the decoder is discarded and only the encoder is used for downstream tasks.

3.1 Evaluation framework

We evaluate our model through two self-supervised learning (SSL) tasks: a closed-world task assessing generalization to new species within known genera, and an open-world task evaluating the model's ability to handle unseen taxonomic groups.

Closed-World Task: 1-NN Probing. To evaluate model generalization to new species within known genera, we perform genus-level 1-NN classification using cosine similarity. We use the training subset of the *Seen* partition as the reference set and the *Unseen* partition as the query set. This task, while involving unseen species, operates within the closed-world setting as it evaluates performance on known genera from the training taxonomy.

Open-World Task: BIN Reconstruction. To assess the model's ability to identify novel species and capture taxonomic relationships, we implement a Barcode Index Number (BIN) reconstruction task. We merge the test subset from the *Seen* partition with the test subset of *Unseen* partition and employ zero-shot clustering on embeddings generated without fine-tuning (Lowe et al., 2024). This evaluation is particularly crucial for understanding the model's capability to group sequences from rare or previously unclassified species based on shared biological features.

3.2 Results

We compared BarcodeMAE against a comprehensive set of baselines, four encoder-only transformer-based models, DNABERT-2 (Zhou et al., 2023), DNABERT-S (Zhou et al., 2024), Nucleotide Transformer (Dalla-Torre et al., 2024), all trained on non-barcode data, and BarcodeBERT (Millan Arias et al., 2023), trained on DNA barcodes. Since BarcodeBERT is a 4-layer encoder-only model and BarcodeMAE uses 6 encoder layers we pretrained a 6-layer model on the BIOSCAN-5M to ensure the fair comparison. We also implemented a baseline that uses an encoder-decoder architecture whilst maintaining the standard masking (BarcodeMAE w/MASK). This serves as a controlled baseline to isolate the impact of architectural choices from masking strategies. Note that, even though BarcodeMAE is conceptually an encoder-decoder model, for both BarcodeMAE and BarcodeMAE w/MASK, we discard the decoder component at inference time, using only the pretrained encoder for downstream tasks.

As shown in Table 1, BarcodeMAE achieves state-of-the-art performance in genus-level classification with 69.0% accuracy, significantly outperforming the previous best baseline, BarcodeBERT, by over 10%. This strong performance in the 1-NN probing task suggests that BarcodeMAE develops more

effective representations of the taxonomic hierarchy, particularly in closed-world scenarios where the genera are known but the species are unseen. Notably, even our BarcodeMAE w/MASK baseline model outperforms existing approaches, demonstrating that decoupling the encoder and decoder alone contributes to improving representation learning in DNA barcode sequences, independent of masking strategy optimizations.

For BIN reconstruction using ZSC, DNABERT-S achieves the highest AMI score of 87.7%, potentially due to its diverse pretraining dataset that aligns well with the clustering objective (Zhou et al., 2024). Notably, BarcodeMAE reaches comparable performance with an AMI of 80.3%, outperforming models like DNABERT-2 and BarcodeBERT. To assess the performance across closed and open-world tasks, we calculated the harmonic mean between genus-level accuracy and BIN reconstruction AMI. BarcodeMAE achieves the highest harmonic mean of 74.2%, outperforming all other baselines. This balanced metric highlights BarcodeMAE's robust performance across both genus-level classification and BIN reconstruction tasks.

Table 1: *Performance comparison of DNA foundation models on BIOSCAN-5M.* We evaluate on two key tasks: genus-level accuracy for 1-NN probing of unseen species and BIN reconstruction AMI using ZSC. The harmonic mean between these metrics provides a balanced assessment of each model's performance across both tasks. The models are divided into two groups: encoder-only transformer-based DNA foundation models, and our proposed model, BarcodeMAE. The **best** results are indicated in bold, and second best underlined. BarcodeMAE achieves the highest harmonic mean of 74.2%, demonstrating superior balanced performance across closed and open-world tasks.

			Genus-level acc (%) of unseen species	BIN clustering AMI (%)	Harmonic Mean
Architecture	**SSL Pretraining**	**Model**	**1-NN probe**	**ZSC probe**	
Encoder-only	Multi-species DNA	DNABERT-2	18.0	77.0	29.2
	Multi-species DNA	DNABERT-S	17.7	**87.7**	29.5
	Multi-species DNA	Nucleotide Transformer	21.7	37.3	27.4
	BIOSCAN-5M	BarcodeBERT	58.3	79.3	67.2
Encoder-decoder	BIOSCAN-5M	BarcodeMAE w/MASK	65.4	80.6	72.2
	BIOSCAN-5M	BarcodeMAE	**69.0**	80.3	**74.2**

To further validate the effectiveness of the model on underrepresented taxa, Figure 2 visualizes the embeddings for 20 randomly sampled genera with fewer than 50 sequences in the dataset. The embeddings from BarcodeMAE and the second best-performing model, BarcodeBERT, are projected to two dimensions using t-SNE (van der Maaten & Hinton, 2008). The visualization shows that BarcodeMAE produces more cohesive and well-separated clusters compared to BarcodeBERT, indicating its ability to learn more discriminative embeddings even for genera with limited samples.

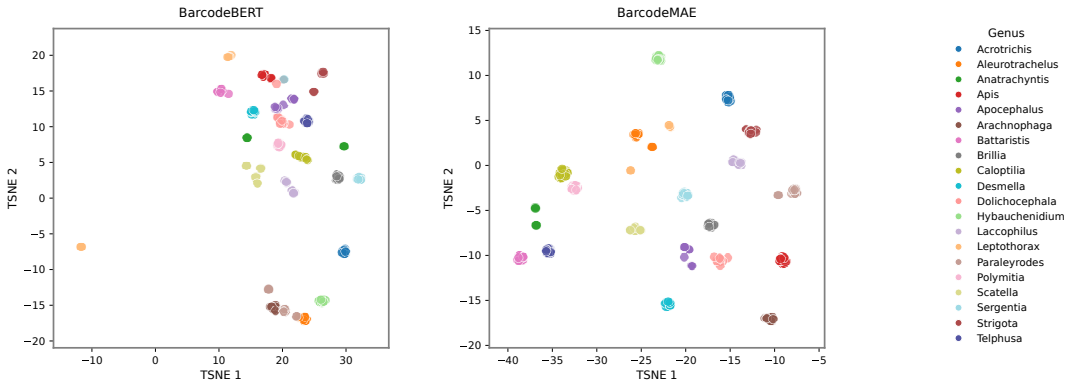

Figure 2: *t-SNE visualization of DNA barcode embeddings* from BarcodeBERT (left) and BarcodeMAE (right) for 20 randomly selected underrepresented genera. Each point represents a DNA barcode sequence, and colours indicate different genera. BarcodeMAE shows more distinct and well-separated clusters, suggesting better discrimination between genera compared to BarcodeBERT.

3.3 ABLATION STUDY

We conducted an ablation study to analyze the impact of different architectural configurations on BarcodeMAE model performance, focusing on two key parameters: k-mer size and the number of layers and attention heads of the encoder-decoder. As shown in Table 2, we systematically varied the number of layers in both the encoder and decoder. The notation "enc:L-H dec:M-J" indicates an encoder with L layers and H attention heads and a decoder with M layers and J attention heads. For each architecture, we evaluated both a k-mer size of $k = 4$ and $k = 6$.

Table 2: *Impact of k-mer size and model architecture on genus-level classification accuracy.* Architecture notation "enc:L-H dec:M-J" indicates L layers and H attention heads in the encoder, and M layers and J heads in the decoder. Results are shown for $k = 4$ and $k = 6$, with **best** performance per k-mer size in bold, and <u>second best</u> performance underlined.

k-mer size	Genus-level acc (%) of unseen species with 1-NN probe				
	enc:4-4 dec:2-2	enc:4-4 dec:4-4	enc:6-6 dec:2-2	enc:6-6 dec:4-4	enc:6-6 dec:6-6
4	64.1	**68.4**	65.0	<u>66.1</u>	65.3
6	60.5	64.9	64.0	<u>67.1</u>	**69.0**

Our experiments demonstrate that the best performance is achieved with balanced encoder and decoder architectures (enc:6-6 dec:6-6), achieving 69.0% accuracy for $k = 6$. This contradicts traditional NLP approaches where shallower decoders are preferred (Meng et al., 2024). The improved performance with deeper decoders indicates that DNA sequence modelling requires more complex feature reconstruction capabilities. This finding provides evidence that effective DNA language models need architectures specifically designed for genomic data rather than direct adaptations from NLP.

3.4 EMPIRICAL EVIDENCE OF REPRESENTATIONAL DEFICIENCY IN DNA FOUNDATION MODELS

In this section, we investigate the effects of `[MASK]` token embeddings across both the encoder-only foundation model, BarcodeBERT, and our proposed encoder-decoder model, BarcodeMAE. To understand how the presence of `[MASK]` tokens impact taxonomic information during inference, we conducted two experiments using the genus-level 1-NN classification task from Section 3.2. First, we replaced different portions of input sequences with `[MASK]` tokens, varying the masking ratio from 0.1 to 0.9, and evaluated the performance of the pretrained BarcodeBERT model (for which the encoder saw `[MASK]` tokens during training). Second, we performed a comparative analysis where instead of substituting the dropped tokens with `[MASK]`, we instead, removed them entirely. This version was performed for both BarcodeMAE and BarcodeBERT. This allows us to study the impact of removing portions of the sequence on both models, and the effect of the `[MASK]` token versus token deletion on model performance.

As shown in Figure 3, we find that BarcodeMAE outperforms BarcodeBERT in the expected downstream use-case where the whole input sequence is shown to the model. The performance of BarcodeMAE decreases approximately linearly as input tokens are removed from the sequence, and begins to fall as soon as any tokens are removed. Meanwhile, the BarcodeBERT model, in both masked and removed variants, demonstrates only a shallow decline in performance as tokens are dropped until reaching its training masking ratio of 50%, after which the accuracy decreases much more rapidly. These results demonstrate that BarcodeBERT is better able to operate on partially complete information, but can not integrate together all information in the sequence. Given that the two training tasks are the same, it is surprising that BarcodeMAE does not match the performance of BarcodeBERT for partial sequences, and this suggests there may be potential for further performance gains.

Additionally, we find that BarcodeBERT performs better when dropped tokens are replaced with the `[MASK]` token instead of being removed completely from the input. The performance gap emerges immediately (+4% at 10% dropped) and increases to reach approximately +10% at 80%

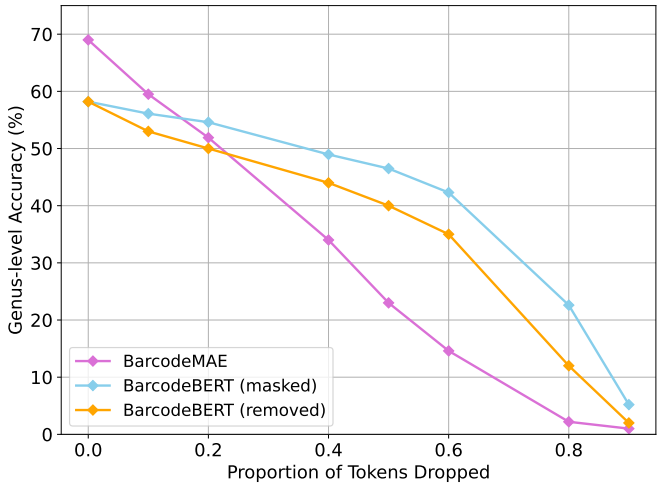

Figure 3: *Impact of masking and token deletion on genus-level classification accuracy.* While BarcodeBERT shows stability at higher drop rates, the practical inference scenario occurs at $x = 0$ with no masking, where BarcodeMAE demonstrates superior performance. The robustness to masking or removing tokens shown by BarcodeBERT does not correspond to an improved real-world performance since these conditions are not encountered during inference.

dropped. Since the `[MASK]` tokens do not contain any information about the specimen's genus, the fact that the BarcodeBERT model performs better when they are present indicates it learned to use the computation associated with `[MASK]` tokens to better extract information from rest of the sequence.

These results empirically demonstrate the distribution shift challenge inherent in masked language modelling, as the model develops dependency on `[MASK]` tokens during training that are absent during inference. The contrasting behaviour of BarcodeMAE, which learns representations solely from observed nucleotides, suggests its architecture may better align with inference-time conditions, where the `[MASK]` token is not present.

4 Conclusion

We introduced BarcodeMAE, an encoder-decoder architecture that mitigates the fundamental limitations of masked language modelling in DNA barcode sequence analysis. By eliminating `[MASK]` tokens during encoding, BarcodeMAE reduces the distribution shift between pretraining and inference, significantly enhancing performance over existing DNA foundation models. Notably, it achieves over a 10% improvement in genus-level classification accuracy on the BIOSCAN-5M dataset compared to the previous state-of-the-art, BarcodeBERT.

While BarcodeMAE does not have the best performance in BIN reconstruction, it achieves the highest harmonic mean across the evaluation tasks, demonstrating a robust performance between closed-world and open-world settings. Our ablation studies reveal that, unlike NLP models that favour shallow decoders, DNA sequence modelling benefits from balanced encoder-decoder architectures, underscoring the need for domain-specific architectural designs.

These findings highlight the critical impact of `[MASK]` token distribution shifts on foundation model effectiveness, particularly in genomic applications where models are used for feature extraction without fine-tuning. The superior performance of BarcodeMAE across diverse evaluation scenarios validates its architectural approach to addressing masking inefficiencies in genomic foundation models.

5 ACKNOWLEDGMENTS

We acknowledge the support of the Government of Canada's New Frontiers in Research Fund [NFRFT-2020-00073]. This research was supported, in part, by the Province of Ontario and the Government of Canada through the Canadian Institute for Advanced Research (CIFAR), and companies sponsoring[2] the Vector Institute. GWT is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs program, and the Canada CIFAR AI Chairs program. LK is supported by NSERC Discovery Grant RGPIN-2023-03663. DS is supported by the Canada First Research Excellence Fund through the University of Guelph's "Food From Thought" program [Project 000054].

REFERENCES

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120 (44):e2311219120, 2023. doi: 10.1073/pnas.2311219120.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. doi: 10.48550/2003.10555.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, Nov 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Tiancheng Gao and Graham W. Taylor. BarcodeMamba: State space models for biodiversity analysis. *arXiv preprint arXiv:2412.11084*, 2024. doi: 10.48550/arxiv.2412.11084.

Zahra Gharaee, Scott C. Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Lila Kari, Dirk Steinke, Graham W. Taylor, Paul Fieguth, and Angel X. Chang. BIOSCAN-5M: A multimodal dataset for insect biodiversity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36285–36313. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fdbb472813041c9ecef04c20c2b1e5a-Paper-Datasets_and_Benchmarks_Track.pdf.

Paul D N Hebert, Sujeevan Ratnasingham, Evgeny V Zakharov, Angela C Telfer, Valerie Levesque-Beaudin, Megan A Milton, Stephanie Pedersen, Paul Jannetta, and Jeremy R deWaard. Counting animal species with DNA barcodes: Canadian insects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 371(1702), 2016. doi: 10.1098/rstb.2015.0333.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

Scott C. Lowe, Joakim Bruslund Haurum, Sageev Oore, Thomas B. Moeslund, and Graham W. Taylor. An empirical study into clustering of unseen datasets with self-supervised encoders. *arXiv preprint arXiv:2406.02465*, 2024. doi: 10.48550/arxiv.2406.02465.

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

Yu Meng, Jitin Krishnan, Sinong Wang, Qifan Wang, Yuning Mao, Han Fang, Marjan Ghazvininejad, Jiawei Han, and Luke Zettlemoyer. Representation deficiency in masked language modeling. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=b3l0piOrGU.

Pablo Millan Arias, Niousha Sadjadi, Monireh Safari, ZeMing Gong, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Dirk Steinke, Lila Kari, Angel X. Chang, Scott C. Lowe, and Graham W. Taylor. BarcodeBERT: Transformers for biodiversity analysis. *arXiv preprint arXiv:2311.02401*, 2023. doi: 10.48550/arxiv.2311.02401.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. Hyena hierarchy: Towards larger convolutional language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28043–28078. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023. doi: 10.48550/arxiv.2306.15006.

Zhihan Zhou, Winmin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. 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 COMPARISON WITH CAUSAL MODELS

To compare our model with recently developed causal models for DNA sequence analysis, we conducted additional experiments comparing BarcodeMAE with several state-of-the-art models, such as HyenaDNA-tiny and Caduceus-PS-1k which are trained on non-barcode data and BarcodeMamba which is trained on DNA barcode data. For a fair comparison, since BarcodeMamba was trained on the CANADA-1.5M dataset (Hebert et al., 2016), we trained BarcodeMAE on the same dataset and evaluated all models on BIOSCAN-5M.

As shown in Table 3, While causal models show strong performance in BIN clustering, with HyenaDNA-tiny achieving 85.0% AMI, they underperform in the genus-level classification of unseen species. BarcodeMamba, specifically trained on DNA barcodes, achieves the highest balanced performance among state space models with 36.3% genus-level accuracy and 82.7% BIN clustering AMI, resulting in a harmonic mean of 50.5%.

The encoder-only architecture, BarcodeBERT, demonstrates enhanced genus-level classification through 1-NN probing compared to causal models, achieving 40.9% accuracy. BarcodeMAE surpasses all competing models with 51.2% genus-level classification accuracy and a harmonic mean of 63.2%, indicating superior balanced performance across metrics. One interesting finding of these results is that BarcodeMAE surpasses models pre-trained on BIOSCAN-5M by 3% in the ZSC bin reconstruction task, despite being trained on the smaller CANADA-1.5M dataset.

Table 3: *Performance comparison of DNA foundation models on BIOSCAN-5M.* We evaluate on two key tasks: genus-level accuracy for 1-NN probing of unseen species and BIN reconstruction AMI using ZSC. The harmonic mean between these metrics provides a balanced assessment of each model's performance across both tasks. The models are divided into three groups: transformer-based DNA foundation models, state space models, and our proposed model, BarcodeMAE. The **best** results are indicated in bold, and second best are underlined.

			Genus-level acc (%) of unseen species	BIN clustering AMI (%)	Harmonic Mean
Architecture	**SSL Pretraining**	**Model**	**1-NN probe**	**ZSC probe**	
State space	Human genome	HyenaDNA-tiny	19.3	**85.0**	31.4
	Human genome	Caduceus-PS-1k	9.1	65.4	15.9
	CANADA-1.5M	BarcodeMamba	36.3	82.7	50.5
Encoder-only	CANADA-1.5M	BarcodeBERT	40.9	73.4	52.5
Encoder-decoder	CANADA-1.5M	BarcodeMAE w/MASK	49.4	83.8	62.2
	CANADA-1.5M	BarcodeMAE	**51.2**	82.4	**63.2**

B BASELINE MODELS

For evaluation, we utilized the respective Pretrained 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
- BarcodeBERT: bioscan-ml/BarcodeBERT
- Caduceous: kuleshov-group/caduceus-ps_seqlen-1k_d_model-256_n_layer-4_lr-8e-3
- BarcodeMamba: bioscan-ml/BarcodeMamba-dim384-layer2-char

B.1 PRETRAINING

BarcodeBERT and BarcodeMAE were pretrained for 35 epochs using the AdamW optimizer Loshchilov & Hutter (2017) with a learning rate of 2×10^{-4}, a batch size of 128, and a OneCycle learning rate scheduler. The pretraining process utilized four NVIDIA V100 GPUs and required approximately 36 hours to complete for each experiment executed. To examine the impact of pretraining, we also trained a model from scratch on the training subset of the *Seen* partition without any pretraining.

B.2 ZERO-SHOT CLUSTERING

We evaluated the models' ability to group sequences without supervision using a modified version of the framework from Lowe et al. (2024). Embeddings were extracted from the pretrained encoders and reduced to 50 dimensions using UMAP (McInnes et al., 2018) to enhance computational efficiency while preserving data structure. These reduced embeddings were clustered with Agglomerative Clustering (cosine distance, Ward's linkage), using the number of true species as the target number of clusters. Clustering performance was assessed with adjusted mutual information (AMI) to measure alignment with ground-truth labels.