Enhancing DNA Foundation Models to Address Masking Inefficiencies

Monireh Safari[*,1], Pablo Millan Arias[*,1], Scott C. Lowe[4], Lila Kari[1], Angel X. Chang[3,5], and Graham W. Taylor[2,4]
*Joint first author

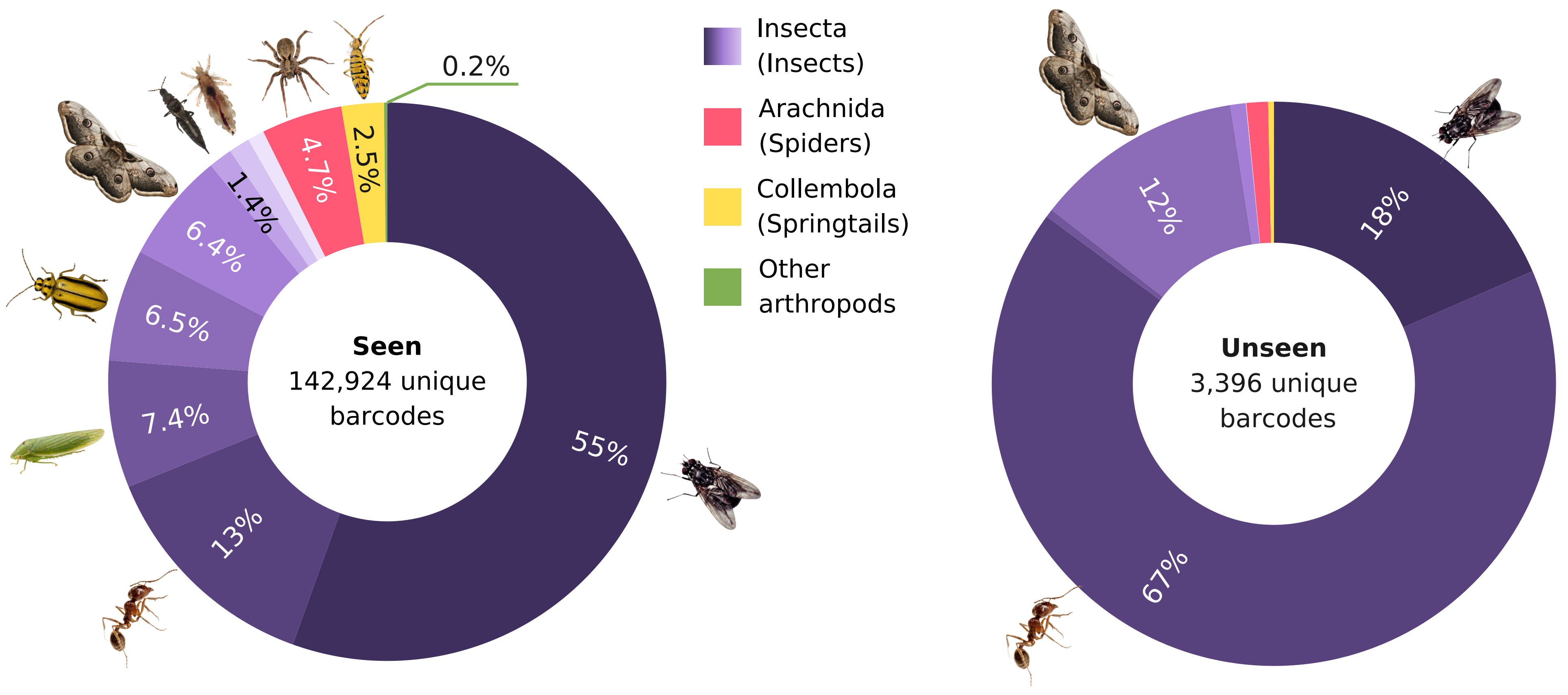
1 UNIVERSITY OF WATERLOO
2 UNIVERSITY OF GUELPH
3 SFU SIMON FRASER UNIVERSITY
4 VECTOR INSTITUTE | INSTITUT VECTEUR
5 amii

Background

- **DNA foundation models** are typically pretrained using **Masked Language Modeling (MLM)** and have shown strong performance on tasks like **specimen classification** to **taxonomic labels**
- The **[MASK] token** appears during **pretraining** for the **MLM task** but is **absent** at **inference**, causing a **distribution shift**. This leads to unused **[MASK] embeddings**, degrading **representation quality** and **downstream** performance
- In this work, we explore the **Masked Autoencoder for MLM (MAE-LM)**[1] to fix the distribution shift in the DNA foundation model. Our results suggest that MAE is effective and improves performance

Dataset

- **DNA Barcode:** 658 bp genetic sequences used for specimen identification
- **BIOSCAN-5M**[2] contains 5.1M records with **2.4M unique DNA barcodes**
- 2.28M barcodes in Pretrain and 145k barcodes in Seen and Unseen subsets:



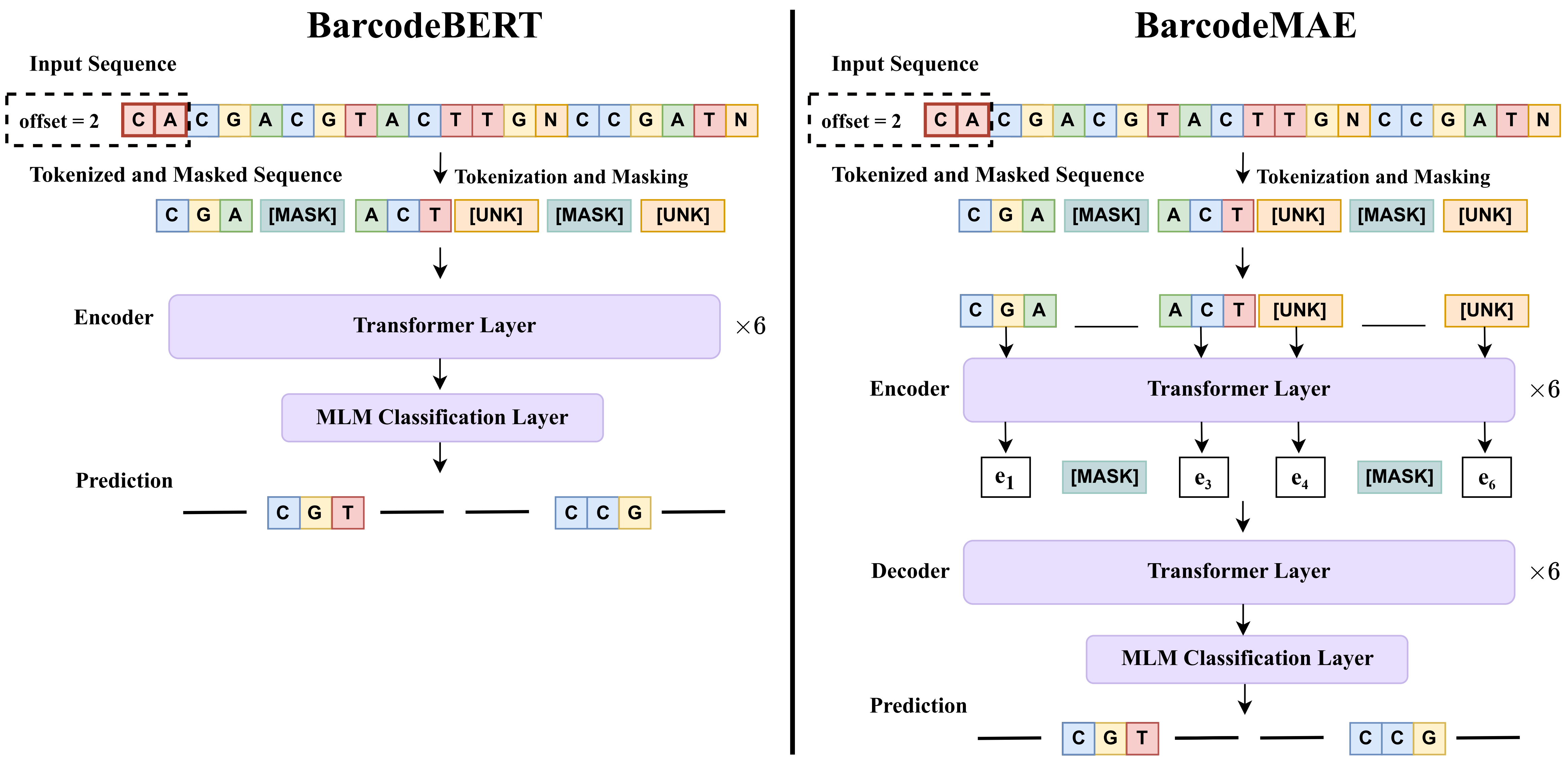
MLM vs MAE-LM in DNA Foundation Models

Encoder-only model: **BarcodeBERT**[3]

- The encoder predicts masked tokens during pretraining
- The absence of [MASK] token in downstream tasks can cause representational deficiency

Our **proposed model**:
Encoder-Decoder model: **BarcodeMAE**

- Encoder **never** sees [MASK] tokens during pretraining
- The decoder predicts masked tokens
- Only the encoder is used for downstream tasks



Experimental Setup and Results

- **1-NN probe:** Tests generalization to new species within known genera using genus-level classification
- **Zero-Shot Clustering (ZSC) probe:** Assesses the ability to identify new species through zero-shot clustering

Architecture	Model	1-NN probe acc (%)	ZSC probe AMI (%)	Harmonic Mean
Encoder-only	DNABERT-2	18.0	77.0	29.2
	DNABERT-S	17.7	**87.7**	29.5
	Nucleotide Transformer	21.7	37.3	27.4
	BarcodeBERT	58.3	79.3	67.2
Encoder-decoder	BarcodeMAE w/MASK	65.4	80.6	72.2
	BarcodeMAE	**69.0**	80.3	**74.2**

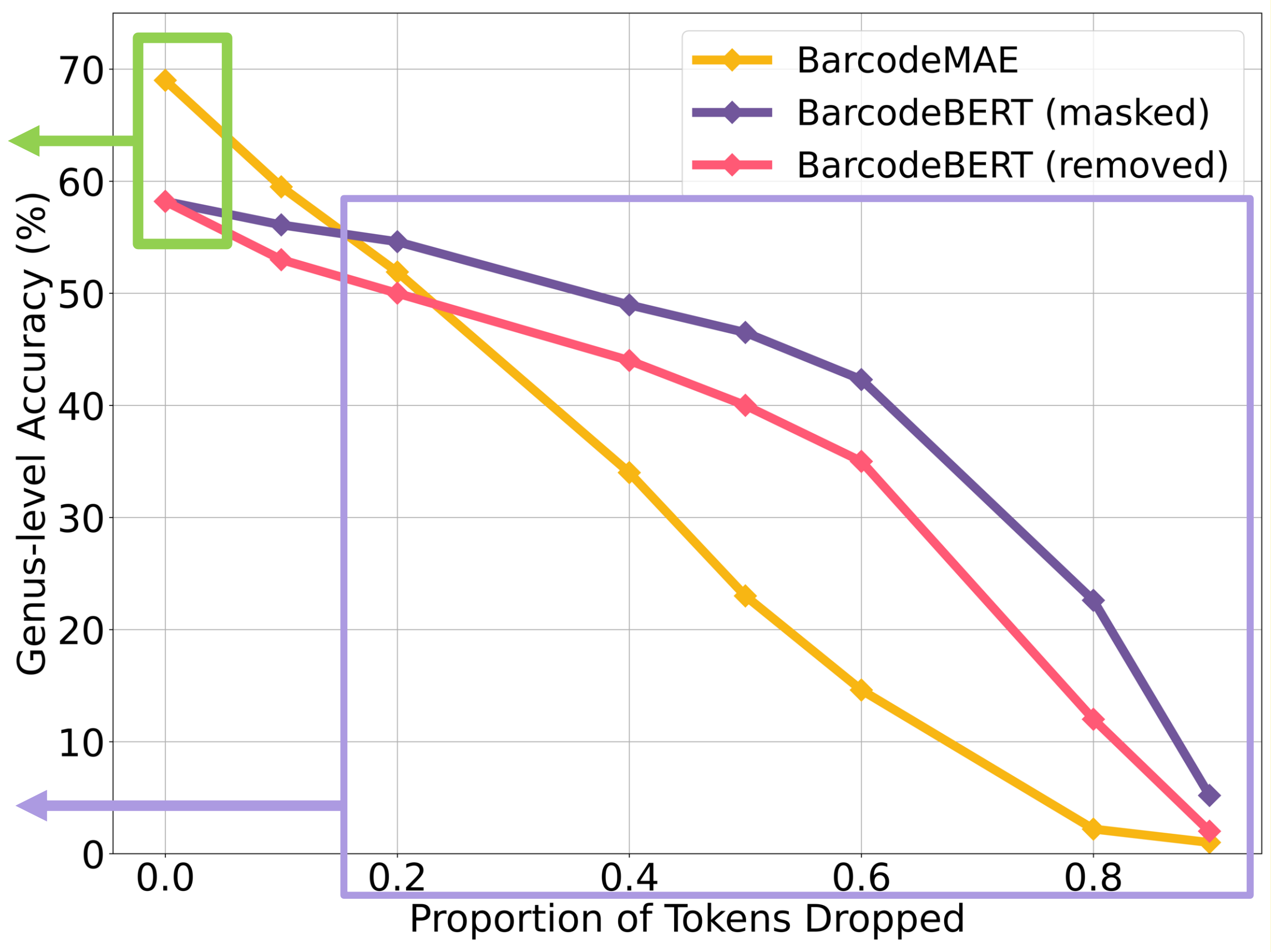
- **BarcodeMAE** outperforms baselines by **10%** in **1-NN probe**
- **BarcodeMAE** achieves **80.3%** AMI **ZSC probe**
- **BarcodeMAE** gets the **highest harmonic mean** across both tasks

Further Findings

- Performance of BarcodeBERT in a **hypothetical** experiment when tokens are **removed** or **masked** during **inference**

The **real-world** setting: The proportion of tokens dropped is **zero**, and BarcodeMAE outperforms BarcodeBERT

The **hypothetical** setting: BarcodeBERT performs well, which corresponds to its training task



- The **performance gap** in masking vs removal setup suggests BarcodeBERT, uses **computation** associated with the **[MASK] token** to better extract information from the remaining sequence

Conclusion

- **BarcodeMAE** performance is **improved** by adopting MAE-LM architecture
- Our results show the BarcodeBERT model develops a **dependency** on the **[MASK] token** during pretraining

GitHub Paper References

BIOSCAN is supported in part by funding from the Government of Canada's New Frontiers in Research Fund (NFRF).

Centre for Biodiversity Genomics

Government of Canada Gouvernement du Canada

Contact:
monireh.safari@uwaterloo.ca
pmillana@uwaterloo.ca