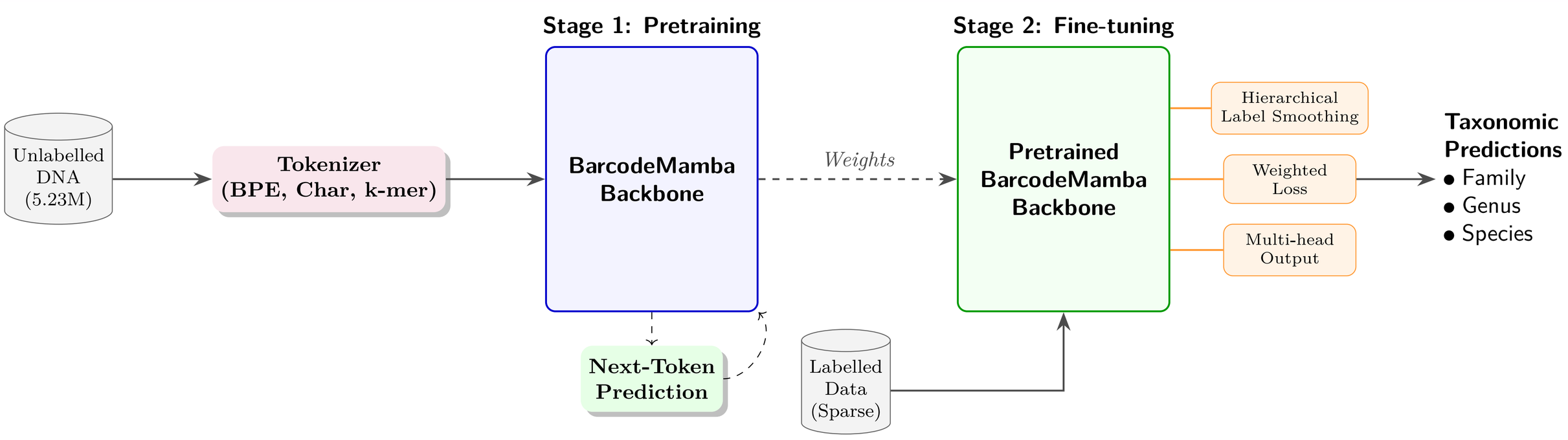
BarcodeMamba+

Advancing State-Space Models for Fungal Biodiversity Research

Tiancheng Gao, Scott C. Lowe
Brendan Furneaux, Angel X. Chang, Graham W. Taylor

VECTOR INSTITUTE
UNIVERSITY of GUELPH
UNIVERSITY OF JYVÄSKYLÄ
SFU

Introduction



BarcodeMamba+ is a novel foundation model to identify fungal species from DNA sequences. Fungi are extremely difficult to classify because current datasets have very few labeled examples for many species. Our model utilizes an efficient **State-Space Model (SSM)** architecture called Mamba-2. We employ a two-stage training strategy: first, the model learns general patterns from millions of unlabeled DNA sequences; second, it is fine-tuned on all the labeled data. We also apply specific enhancements to handle the complex fungal "family tree". Experiments show that BarcodeMamba+ significantly outperforms traditional methods like BLAST and recent deep learning models, offering a robust solution for global biodiversity monitoring.

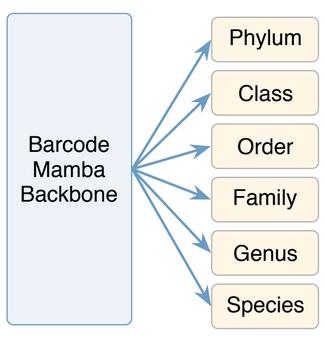
Overview & Methodology

- **Sparse Data Problem:** Fungal identification is hindered by extreme **label scarcity** (93% unlabeled), causing standard models (CNNs) to fail.
- **Our Approach:** BarcodeMamba+ uses an efficient Mamba-2 SSM backbone. By pretraining on unlabeled data before fine-tuning, it learns robust representations, overcoming annotation bottlenecks.

Methodology Design

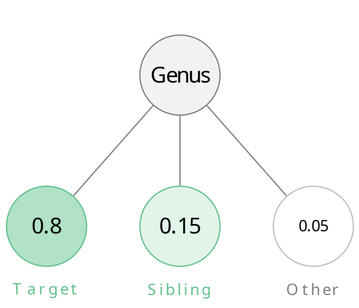
- **Dataset:** UNITE+INSD dataset, 5.23M sequences, 7% labeled at the species level
- **Architecture:** BarcodeMamba for linear scaling with sequence length, ensuring high speed.
- **Pretrain + Fine-tune:** Two-stage training leveraging Next-Token Prediction on unlabeled data.
- **Hierarchical optimizations from MycoAI:**
  - *Label Smoothing:* Distinguishes closely related species.
  - *Weighted Loss:* Focuses on rare, long-tailed taxa.
  - *Multi-head Output:* Simultaneous prediction of all the taxonomic ranks.

1. Multi-head output



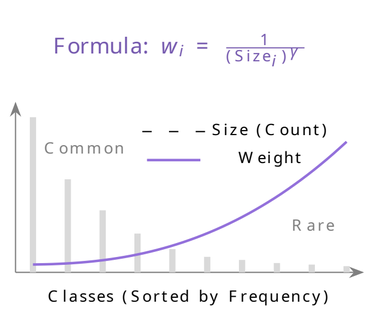
Simultaneouly predicts 6 taxonomic ranks to learn hierarchical features.

2. Label smoothing



Assigns soft probabilities to taxonomic siblings, penalizing "near-misses" less.

3. Weighted Loss



Rare species receive geometrically higher weights to balance learning.

Results

Species classification accuracy (%) on three fungi test sets (Yeast, Filamentous Fungi, and MycoAI).

Model	Yeast	Filam.	MycoAI	Size ↓	Time ↓
BLAST	75.4	33.4	55.0	N/A	208.6 ms
MycoAI-CNN (Vu)	60.0	28.2	57.1	11.6 M	11.8 ms
MycoAI-BERT (base)	33.5	16.6	39.3	18.4 M	4.5 ms
CNN Encoder	67.6	31.4	72.6	12.1 M	5.8 ms
BarcodeBERT	59.1	27.7	58.9	44.6 M	8.8 ms
BarcodeMamba+	80.6	46.5	81.7	12.1 M	8.0 ms
BarcodeMamba+ (large)	**83.6**	**50.4**	**88.9**	49.2 M	14.7 ms

- **Superior Performance:** On the comprehensive MycoAI benchmark, BarcodeMamba+ achieves **81.7%** species-level accuracy, significantly outperforming the CNN Encoder (72.6%) and BarcodeBERT (58.9%).
- **Robust Generalization:** On "Filamentous Fungi"—a dataset with distinct distribution shifts—our model improves accuracy from 31.4% to **46.5%**, demonstrating superior adaptability to unseen species.
- **High Efficiency:** The SSM architecture ensures rapid inference at **8.0 ms per sample**, making it over 25x faster than the traditional BLAST algorithm (208.6 ms).

Acknowledgements

CIFAR
Government of Canada / Gouvernement du Canada

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