BarcodeMamba+: Advancing State-Space Models for Fungal Biodiversity Research

Tiancheng Gao[1,2] **Scott C. Lowe**[2] **Brendan Furneaux**[3] **Angel X. Chang**[4,5]
Graham W. Taylor[1,2*]

[1]University of Guelph [2]Vector Institute [3]University of Jyväskylä
[4]Simon Fraser University [5]Amii

Abstract

Accurate taxonomic classification from DNA barcodes is a cornerstone of global biodiversity monitoring, yet fungi present extreme challenges due to sparse labelling and long-tailed taxa distributions. Conventional supervised learning methods often falter in this domain, struggling to generalize to unseen species and to capture the hierarchical nature of the data. To address these limitations, we introduce BarcodeMamba+, a foundation model for fungal barcode classification built on a powerful and efficient state-space model architecture. We employ a pretrain and fine-tune paradigm, which utilizes partially labelled data and we demonstrate this is substantially more effective than traditional fully-supervised methods in this data-sparse environment. During fine-tuning, we systematically integrate and evaluate a suite of enhancements—including hierarchical label smoothing, a weighted loss function, and a multi-head output layer from MycoAI—to specifically tackle the challenges of fungal taxonomy. Our experiments show that each of these components yields significant performance gains. On a challenging fungal classification benchmark with distinct taxonomic distribution shifts from the broad training set, our final model outperforms a range of existing methods across all taxonomic levels. Our work provides a powerful new tool for genomics-based biodiversity research and establishes an effective and scalable training paradigm for this challenging domain. Our code is publicly available at https://github.com/bioscan-ml/BarcodeMamba.

1 Introduction

DNA barcodes, short standardized DNA sequences used for specimen recognition and species identification, enable large-scale, automated biodiversity monitoring (Hebert et al., 2003). Fungal biodiversity presents an extreme challenge for barcode classification. Visual and morphological features help identify other taxa, but fungal species identification is often confounded by minimalistic features, necessitating an almost complete reliance on DNA sequences (Bickford et al., 2007). Currently, up to 93% of collected fungal samples remain unannotated at the species level (Romeijn et al., 2024).

This annotation sparsity has exposed fundamental limitations in existing computational approaches. Traditional algorithmic methods like BLAST (Altschul et al., 1990), RDP classifier (Wang et al., 2007), and dnabarcoder (Vu et al., 2022) are standard tools for sequence identification but face prohibitive inference times on large datasets and poor generalization to novel taxa. Learning-based

*Author for correspondence: gwtaylor@uoguelph.ca

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

methods using specialized convolutional neural networks (CNNs) and transformer architectures show promise with fully supervised training (Badirli et al., 2021; Romeijn et al., 2024) but require densely labelled data, making them vulnerable to the class imbalance and label sparsity that characterize fungal datasets.

Foundation models tackle sparse training labels through a pretrain + fine-tune paradigm. The vast amounts of unlabelled data can be harnessed during pretraining to learn rich, generalizable representations, before adapting to specific tasks with fine-tuning. This approach is effective for biodiversity applications where unlabelled data vastly outnumbers annotated specimens, as demonstrated by the transformer-based BarcodeBERT (Arias et al., 2023) and BarcodeMAE (Safari et al., 2025).

State-space models (SSMs) (Gu et al., 2022) and particularly the Mamba architecture (Gu & Dao, 2024; Dao & Gu, 2024), have emerged as compelling alternatives to Transformers for sequence modelling. SSMs offer competitive performance with significantly lower computational overhead, making them attractive for large-scale biodiversity applications where datasets contain millions of sequences. Our previous work (Gao & Taylor, 2024) introduced BarcodeMamba, demonstrating the effectiveness of SSMs for insect barcode (COI) classification. This suggests a strong potential application of SSMs to fungal data, which faces similar challenges.

We introduce **BarcodeMamba+**, which adapts BarcodeMamba for hierarchical fungal ITS barcode classification. Our experiments demonstrate BarcodeMamba+ outperforms established methods across taxonomic ranks on standard fungal classification benchmarks. Our contributions are:

1. The development and comprehensive evaluation of BarcodeMamba+, an SSM-based foundation model for fungal barcode classification.
2. Demonstration that pretrain + fine-tune approaches outperform fully-supervised methods in this annotation-sparse, taxonomically diverse domain.
3. Systematic analysis of hierarchical smoothing, inverse square root weighted loss (hereafter shortened to weighted loss), and multi-head outputs for adapting foundation models to hierarchical taxonomic classification.
4. Analysis of model scaling effects on taxonomic classification performance.

2 Methods

2.1 Dataset

Our experiments use the MycoAI (Romeijn et al., 2024) splits of the UNITE+INSD data (Abarenkov et al., 2020), a comprehensive fungal internal transcribed spacer (ITS) barcode repository.

Training and Validation Sets. The training set is comprised of 5.23 M sequences, representing 14.7 k distinct species across a taxonomic hierarchy of 18 phyla, 70 classes, 231 orders, 791 families, and 3,695 genera. Only 7% of the samples are annotated to species-level. This creates a complex multi-label, hierarchical classification challenge. The validation set contains 10.5 k sequences, randomly sampled from Abarenkov et al. (2020).

Test Sets. We use the three MycoAI test sets, representing distinct taxonomic distribution shifts from the broad training set. Appendix D analyzes the species-level and identical-barcode overlap between the training set and each test set. Test examples belonging to classes that were unobserved during training were omitted from our evaluation.

- **Test Set 1: Yeast** (Vu et al., 2016). Contains 4.4 k ITS sequences from yeast species, evaluating the model's generalization to a specific and taxonomically concentrated clade.
- **Test Set 2: Filamentous Fungi** (Vu et al., 2019). A set of 11.6 k sequences from filamentous fungi, a broad but distinct collection of taxa not necessarily well-represented in the training set.
- **Test Set 3: MycoAI Benchmark** (Romeijn et al., 2024). The largest test set with 367 k samples, serving as a comprehensive benchmark for overall performance and robustness.

Table 1: Performance of BarcodeMamba+ and baselines on the three test sets for taxonomic ranks family, genus, and species. We report accuracy (micro) (%), model size (parameters), and inference time per sample (ms). ↑: higher is better; ↓: lower is better. Bold: **best**; underlined: second best.

	Yeast Acc. (%)↑			Filamentous Acc. (%)↑			MycoAI Acc. (%)↑				
Model	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Size ↓	Time ↓
BLAST	86.6	92.9	75.4	81.4	71.5	33.4	94.7	93.1	55.0	N/A	208.6 ms
MycoAI-CNN (Vu)	90.5	86.4	60.0	84.1	69.8	28.2	93.9	87.8	57.1	11.6 M	11.8 ms
MycoAI-BERT (base)	88.9	75.7	33.5	85.1	60.8	16.6	93.2	80.3	39.3	18.4 M	4.5 ms
CNN Encoder	94.1	88.3	67.6	84.5	69.1	31.4	97.5	93.6	72.6	12.1 M	5.8 ms
BarcodeBERT	95.4	88.6	59.1	87.8	70.2	27.7	97.8	92.0	58.9	44.6 M	8.8 ms
BarcodeMamba+	98.7	95.3	80.6	**92.6**	81.1	46.5	99.0	96.5	81.7	12.1 M	8.0 ms
BarcodeMamba+ (large)	**98.8**	**95.9**	**83.6**	92.5	**81.6**	**50.4**	**99.3**	**97.7**	**88.9**	49.2 M	14.7 ms

Data Preprocessing. We used the preprocessed MycoAI dataset (Romeijn, 2024), with four filtering steps: (1) removal of duplicate sequence-label pairs, (2) exclusion of sequences with length more than four standard deviations from the mean (558.0 bp ± 126.2 bp), (3) removal of sequences with over 5% ambiguous bases, and (4) elimination of taxonomic classes with fewer than three representative samples. The remaining sequences in the training split are annotated to varying depths within the seven-level taxonomic hierarchy (kingdom, phylum, class, order, family, genus, and species). The dataset was partitioned into training, validation, and test splits after all the above preprocessing steps.

2.2 Model Architectures

2.2.1 Baselines

We compare BarcodeMamba+ against baselines from three categories. BLAST (Altschul et al., 1990) serves as a representative non-learning algorithmic method. For fully-supervised deep learning models, we compare against a CNN Encoder (Badirli et al., 2021), and both MycoAI-CNN and MycoAI-BERT (Romeijn et al., 2024). BarcodeBERT (Arias et al., 2023) provides a competitive foundation model baseline, pretrained on COI barcodes. The two MycoAI models incorporate the same enhancements for hierarchical modelling which we evaluate in Section 3.2.2. Complete architectural details and experimental configurations for all baselines are provided in Appendix A.1.

2.2.2 BarcodeMamba+

Our BarcodeMamba+ model adapts the BarcodeMamba SSM architecture for hierarchical fungal ITS barcode classification. We use a BPE tokenizer following Romeijn et al. (2024)'s recommendation for fungal data. Complete implementation details are provided in Appendix A.4.

Training Paradigm. We employ a two-stage approach:

- **Pretraining:** The tokenizer and model learn fungal ITS sequence patterns from unlabelled UNITE+INSD data through next-token prediction, without taxonomic labels.
- **Fine-tuning:** We add a classification head and fine-tune on labelled data, incorporating the enhancements from Appendix A.2 to address hierarchical labels and class imbalance.

3 Experiments

3.1 Comparison study

We trained the models (as described in Appendix B), then evaluated the performance on the three test datasets. The results (Table 1) demonstrate BarcodeMamba+ outperforms all baseline models across all taxonomic levels and metrics, while maintaining inference efficiency. On the largest MycoAI test set (Benchmark), our model achieves a species-level accuracy of 81.7%, surpassing the next-best performing baseline, CNN Encoder, by 9.1 percentage points (72.6%). This gap is even more pronounced on the challenging Filamentous Fungi test set, where our model's species-level

Table 2: Ablation comparing three tokenizers and two training paradigms: supervised from scratch (✗) and fine-tuned following pretraining (✓). Results show accuracy (micro), precision (macro), and recall (macro) on three test sets at family, genus, and species level. Bold: **best** result for a given taxonomic rank and test set; underlined: second best.

			Accuracy (%)↑			Precision (%)↑			Recall (%)↑		
Test set	Tokenizer	Pretrain	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.
Yeast	Char	✓	98.4	94.6	77.4	90.5	88.7	72.9	87.0	79.5	39.7
		✗	97.9	94.2	76.7	86.9	87.1	72.0	87.8	77.9	39.6
	k-mer	✓	98.6	94.8	77.8	**93.2**	90.5	72.6	**90.5**	79.8	40.9
		✗	97.8	93.9	73.0	84.3	83.1	64.9	83.7	73.4	32.3
	BPE	✓	**98.7**	**95.3**	**80.6**	93.1	**92.3**	**77.0**	90.2	**82.2**	**46.0**
		✗	97.9	93.7	78.6	79.4	84.6	72.0	87.0	79.0	42.6
Filamentous	Char	✓	91.7	79.5	42.2	80.5	68.2	44.2	75.3	56.2	26.1
		✗	91.4	79.7	42.0	79.8	67.2	43.5	74.5	55.4	25.9
	k-mer	✓	91.4	78.9	42.3	80.9	65.8	42.2	74.6	53.8	25.9
		✗	89.3	74.6	36.4	73.4	58.2	34.3	67.3	45.9	20.8
	BPE	✓	**92.6**	**81.1**	**46.5**	**81.8**	**71.2**	**48.9**	**77.3**	**60.7**	**31.3**
		✗	90.3	78.6	43.2	75.9	66.7	43.6	73.5	57.0	27.3
Myco	Char	✓	98.8	96.2	79.0	95.6	91.1	84.8	95.8	89.9	54.5
		✗	98.7	95.9	78.2	94.2	90.6	85.2	95.6	89.5	55.0
	k-mer	✓	99.0	**96.9**	81.1	**96.4**	**93.2**	86.4	**97.7**	**93.2**	56.5
		✗	99.0	96.5	77.0	96.1	92.6	82.2	96.6	89.1	45.6
	BPE	✓	**99.0**	96.5	**81.7**	95.2	91.3	**88.3**	97.0	93.0	**65.8**
		✗	98.8	95.8	78.8	93.5	89.0	85.7	96.1	89.9	57.5

accuracy (46.5%) is over 15 points higher than that of CNN Encoder (31.4%). This highlights our architecture's enhanced robustness to the distributional shifts present between the training domain and the Filamentous Fungi test set.

Our model achieves this performance with a compact model size of 12.1 M parameters, comparable to MycoAI-CNN (Vu) (11.6 M) and MycoAI-BERT (base) (18.4 M) and significantly smaller than BarcodeBERT (44.6 M). Compared to the non–learning-based baseline BLAST, our model achieves vastly higher accuracy on fine-grained classification (e.g., 81.7% vs. 55.0% on MycoAI species) and demonstrates over 25× faster inference (8.0 ms vs. 208.6 ms), rendering it far more suitable for large-scale biodiversity applications. After scaling up to 49.2 M parameters, BarcodeMamba+ improves performance on every task. It boosts the species-level accuracy on MycoAI by another 7.2 points, from an already high 81.7% to an exceptional 88.9%. Similarly, on Filamentous Fungi, Species accuracy increases from 46.5% to 50.4%. This result confirms that our model architecture scales effectively, and its capacity to leverage increased parametrization translates directly into improved accuracy, especially for classifying the long tail of rare species.

3.2 Ablation Studies

We conduct two ablation studies, with the first study focusing on the effectiveness of pretraining and different tokenizers, and the second on the impact of the three enhancements for hierarchical data.

3.2.1 Ablation A: Pretrain + Finetune vs. Supervised Learning on UNITE Dataset

We compare pretrain + fine-tune against fully supervised training while evaluating three tokenization methods. All models use hierarchical label smoothing, multi-head outputs, and weighted loss. Results, shown in Table 2, demonstrate the benefits of pretraining and find BPE is the best tokenizer.

3.2.2 Ablation B: Label smoothing, Multi-head, and Weighted loss

Various methods for handling hierarchical labels present opportunities to improve model training, which we incorporated into our model training paradigm. To investigate the impact of each of these on the performance of our model, we ablated these configurations, with results shown in Table 3 and statistical tests shown in Appendix C. We find that hierarchical label smoothing consistently

Table 3: Ablation of supervised learning enhancements: label smoothing (None, Standard, Hierarchical), weighted loss (WL), and multi-head outputs (MH). ✓: enabled, ✗: ablated (standard alternative). Results show accuracy, precision, and recall at the species level across test sets. Bold: **best** result; underlined: second best.

Components			Accuracy (%)↑			Precision (%)↑			Recall (%)↑		
Smoothing	WL	MH	Yeast	Filam.	Myco	Yeast	Filam.	Myco	Yeast	Filam.	Myco
None	✗	✗	67.4	37.8	72.6	63.3	41.7	75.7	27.3	22.0	38.3
	✗	✓	68.1	37.2	73.3	57.9	38.5	78.4	26.6	20.7	34.8
	✓	✗	72.0	41.3	75.9	63.4	41.4	77.8	31.9	25.1	46.9
	✓	✓	73.1	40.0	76.0	60.3	40.1	81.4	32.4	23.9	45.0
Standard	✗	✗	64.6	35.2	69.3	58.7	40.3	70.2	22.1	19.8	29.9
	✗	✓	61.8	35.4	69.8	54.7	36.7	72.9	21.6	18.6	27.2
	✓	✗	70.5	40.7	75.2	63.3	41.2	76.7	31.1	24.6	44.3
	✓	✓	72.1	39.8	75.6	61.7	40.6	80.0	31.1	23.2	42.3
Hierarchical	✗	✗	71.8	41.3	77.5	66.6	43.0	82.7	33.0	25.4	51.8
	✗	✓	73.3	40.3	76.5	68.9	42.5	83.9	33.8	24.3	48.6
	✓	✗	76.3	**42.6**	78.0	71.5	**43.9**	83.1	**39.9**	**26.8**	**55.5**
	✓	✓	**76.8**	42.0	**78.2**	**72.0**	43.8	**85.3**	39.2	26.1	55.1

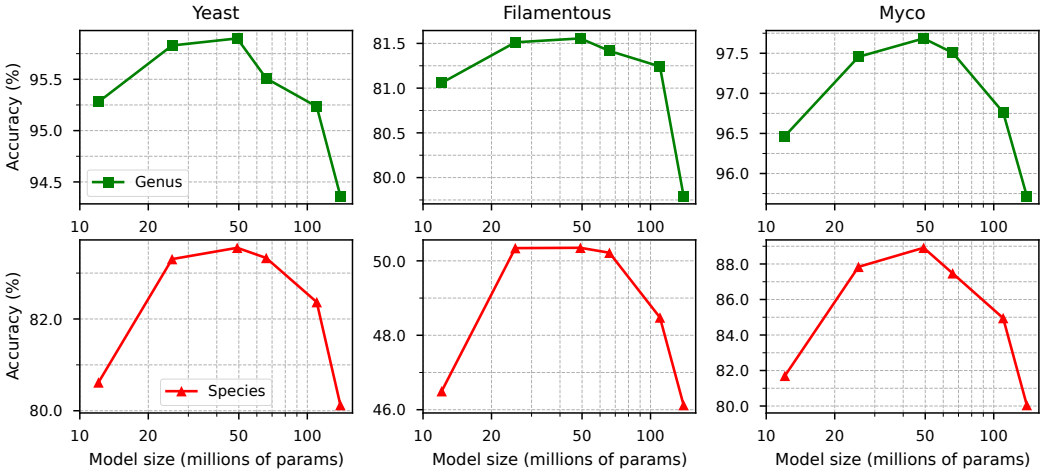

Figure 1: Scaling of BarcodeMamba+ to different model sizes. For each test set, we report the accuracy for classifying at genus (upper panels) and species (lower panels) ranks.

improves performance across all metrics (avg. +3.3% acc.), whereas standard smoothing does not provide a significant benefit. Using a weighted loss also provides a consistent improvement for the imbalanced data (avg. +4.1% acc.). However, multi-head outputs provided inconsistent gains over using a single, species-level, head (avg. −0.04% acc.) when fine-tuning on data labelled to species.

3.3 Scaling study

Using our default configuration (pretrain+fine-tune with BPE tokenizer, hierarchical label smoothing, weighted loss, and multi-head output), we conduct a scaling study (Figure 1). Accuracy on fine-grained ranks (genus and species) is highly sensitive to model capacity. Performance peaks at ~50 M parameters, consistent with MycoAI findings (Romeijn et al., 2024, Fig. 8), then degrades at 140 M parameters for species-level tasks, suggesting overfitting on fine-grained classification.

4 Conclusion

We addressed fungal DNA barcode classification, a domain with extreme label sparsity and long-tailed distributions. BarcodeMamba+ demonstrates that SSM-based foundation models using

pretrain+fine-tune paradigms substantially outperform fully-supervised approaches. Our systematic evaluation shows BPE tokenization, hierarchical label smoothing, and weighted loss are effective, especially enhancing recall for rare classes. Our scaling study shows benefits of SSM-based architectures over Transformer-based alternatives while revealing inherent limits on useful model capacity for this task.

This work enables broader biodiversity research. The enhanced model structure can extend to other genetic markers like COI for insects (Elbrecht et al., 2019; Steinke et al., 2024), and rbcL for plants (CBOL Plant Working Group, 2009; Hollingsworth et al., 2011). We also see opportunities to integrate genomic data with imaging and environmental modalities (c.f. Gong et al., 2025; Gu et al., 2025), aligning with growing recognition that comprehensive biodiversity understanding requires diverse data types. By developing scalable AI tools for under-resourced domains like mycology, we can accelerate the pace of species discovery and taxonomic annotation in Earth's most biodiverse yet least understood kingdoms.

Acknowledgments and Disclosure of Funding

BIOSCAN is supported in part by funding from the Government of Canada's New Frontiers in Research Fund (NFRF). 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. AXC and GWT acknowledge support from the Natural Sciences and Engineering Research Council (NSERC), the Canada Research Chairs program, and the Canada CIFAR AI Chairs program.

References

Kessy Abarenkov, Allan Zirk, Timo Piirmann, Raivo Pöhönen, Filipp Ivanov, R. Henrik Nilsson, and Urmas Kõljalg. UNITE QIIME release for fungi, Jan 2020. URL https://cir.nii.ac.jp/crid/1880583643043641216.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi:10.1016/S0022-2836(05)80360-2.

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.

Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M Dundar. Fine-grained zero-shot learning with DNA as side information. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19352–19362. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a18630ab1c3b9f14454cf70dc7114834-Paper.pdf.

David Bickford, David J. Lohman, Navjot S. Sodhi, Peter K.L. Ng, Rudolf Meier, Kevin Winker, Krista K. Ingram, and Indraneil Das. Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*, 22(3):148–155, 2007. ISSN 0169-5347. doi:10.1016/j.tree.2006.11.004.

CBOL Plant Working Group. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31):12794–12797, 2009. doi:10.1073/pnas.0905845106.

Tri Dao and Albert Gu. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR, 2024. doi:10.48550/arXiv.2405.21060.

Jeremy R. deWaard, Sujeevan Ratnasingham, Evgeny V. Zakharov, Alex V. Borisenko, Dirk Steinke, Angela C. Telfer, Kate H. J. Perez, Jayme E. Sones, Monica R. Young, Valerie Levesque-Beaudin, Crystal N. Sobel, Arusyak Abrahamyan, Kyrylo Bessonov, Gergin Blagoev, Stephanie L. deWaard, Chris Ho, Natalia V. Ivanova, Kara K. S. Layton, Liuqiong Lu, Ramya Manjunath, Jaclyn T. A. McKeown, Megan A. Milton, Renee Miskie, Norm Monkhouse, Suresh Naik, Nadya Nikolova, Mikko Pentinsaari, Sean W. J. Prosser, Adriana E. Radulovici, Claudia Steinke, Connor P. Warne, and Paul D. N. Hebert. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*, 6(1):308, Dec 2019. ISSN 2052-4463. doi:10.1038/s41597-019-0320-2.

Vasco Elbrecht, Thomas W.A. Braukmann, Natalia V. Ivanova, Sean W.J. Prosser, Mehrdad Hajibabaei, Michael Wright, Evgeny V. Zakharov, Paul D.N. Hebert, and Dirk Steinke. Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7:e7745, October 2019. ISSN 2167-8359. doi:10.7717/peerj.7745.

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.

ZeMing Gong, Austin Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X Chang. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d5HUnyByAI.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. doi:10.48550/arXiv.2312.00752. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint arXiv:2505.23883*, 2025. doi:10.48550/arxiv.2505.23883.

Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321, 2003. doi:10.1098/rspb.2002.2218.

Peter M. Hollingsworth, Sean W. Graham, and Damon P. Little. Choosing and using a plant DNA barcode. *PLOS One*, 6(5):1–13, May 2011. doi:10.1371/journal.pone.0019254.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf.

Luuk Romeijn. MycoAI data. *Zenodo*, April 2024. doi:10.5281/zenodo.10946477.

Luuk Romeijn, Andrius Bernatavicius, and Duong Vu. MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources*, 24(8):e14006, 2024. doi:10.1111/1755-0998.14006.

Monireh Safari, Pablo Millan Arias, Scott C. Lowe, Lila Kari, Angel X. Chang, and Graham W. Taylor. Enhancing DNA foundation models to address masking inefficiencies. *arXiv preprint arXiv:2502.18405*, 2025. doi:10.48550/arXiv.2502.18405.

Dirk Steinke, Sujeevan Ratnasingham, Jireh Agda, Hamzah Ait Boutou, Isaiah C. H. Box, Mary Boyle, Dean Chan, Corey Feng, Scott C. Lowe, Jaclyn T. A. McKeown, Joschka McLeod, Alan Sanchez, Ian Smith, Spencer Walker, Catherine Y.-Y. Wei, and Paul D. N. Hebert. Towards a taxonomy machine: A training set of 5.6 million arthropod images. *Data*, 9(11), 2024. ISSN 2306-5729. doi:10.3390/data9110122.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016. doi:10.1109/CVPR.2016.308.

Duong Vu, Marizeth Groenewald, S Szöke, Gianluigi Cardinali, Ursula Eberhardt, Benjamin Stielow, Michel De Vries, GJM Verkleij, Pedro W Crous, Teun Boekhout, and Vincent Robert. DNA barcoding analysis of more than 9 000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in mycology*, 85(1):91–105, 2016. ISSN 0166-0616. doi:10.1016/j.simyco.2016.11.007.

Duong Vu, Marizeth Groenewald, Michel De Vries, Thomas Gehrmann, Benjamin Stielow, Ursula Eberhardt, Abdullah Al-Hatmi, Johannes Zacharias Groenewald, Gianluigi Cardinali, Jos Houbraken, et al. Large-scale generation and analysis of filamentous fungal DNA barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation. *Studies in mycology*, 92(1):135–154, 2019. doi:10.1016/j.simyco.2018.05.001.

Duong Vu, R Henrik Nilsson, and Gerard JM Verkley. Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources*, 22(7):2793–2809, 2022. doi:10.1111/1755-0998.13651.

Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007. doi:10.1128/AEM.00062-07.

Appendices

A Model Architecture Details

A.1 Baseline Models

Non-learning-based Baseline. We use BLASTN (Altschul et al., 1990) as a representative non-learning-based method. First, a searchable nucleotide database was constructed from the training set sequences (`trainset.fasta`) using the `makeblastdb` command. Sequences from each test set were then aligned against this database using the `blastn` algorithm. The search was parallelized across 16 CPU threads (`-num_threads=16`) for computational efficiency. Results were generated in tabular format (`-outfmt=6`), providing a list of all significant alignments for each query. In line with exploratory sequence similarity searches, we did not set explicit thresholds for e-value, query coverage, or sequence identity, allowing for the capture of a broad range of potential matches for downstream analysis.

Fully Supervised Baselines. We compare against deep learning architectures trained end-to-end without a self-supervised pre-training phase.

- **CNN Encoder:** This model, introduced by Badirli et al. (2021), is known for its computational efficiency and accuracy. The architecture consists of three 2D-convolutional layers with kernel sizes of 3×3, channel dimensions of 64, 32, and 16, respectively, each followed by batch normalization and a ReLU activation function, and interleaved with max-pooling layers of size 3×1. The final feature maps are flattened and passed through a fully-connected layer.
- **MycoAI-CNN and MycoAI-BERT:** These are the state-of-the-art fully supervised models from Romeijn et al. (2024). The MycoAI-BERT model is a Transformer-based architecture with 8 encoder layers, 8 attention heads, a hidden dimension of 512, and a feed-forward dimension of 1024. The MycoAI-CNN model is a simple CNN with two convolutional layers (5 and 10 channels, respectively) using a kernel size of 5, followed by max-pooling with pool size 2 and a fully-connected layer of size 256. Both are enhanced by the techniques discussed in Appendix A.2. For tokenization, the strongest performing variants were used: BPE for BERT and k-mer-spectral for CNN.

A.2 Supervised Learning Enhancements for Fine-tuning

Following Romeijn et al. (2024), we conduct ablations evaluating three techniques during the fine-tuning stage for both our model and the supervised baselines.

- **Hierarchical Label Smoothing (HLS):** Standard label smoothing penalizes confident predictions (Szegedy et al., 2016). HLS, introduced by Romeijn et al. (2024), adapts this concept to taxonomy by reducing the penalty for misclassifications that are taxonomically close to the true label (e.g., predicting the correct genus but wrong species). This encourages the model to learn the taxonomic hierarchy.
- **Classification Head:** We compare two output strategies by Romeijn et al. (2024). The first is a **multi-head** architecture where separate linear layers predict each of the seven taxonomic ranks simultaneously, allowing the model to learn shared representations. The second is a **single-head** baseline that predicts only at the species level, with higher-rank probabilities inferred from the species predictions using a pre-defined taxonomic matrix.
- **Weighted Loss:** To counteract the severe class imbalance in the dataset, we adopt the weighted cross-entropy loss from Romeijn et al. (2024). The loss for each sample is weighted by the inverse square root of its class frequency, encouraging the model to pay more attention to rare taxa.

A.3 Foundation Model Baseline

To benchmark our SSM-based approach against the current state-of-the-art in biodiversity foundation models, we include BarcodeBERT (Arias et al., 2023) as our primary comparison point. This transformer-based model was pre-trained on a large-scale invertebrate COI barcode dataset (de-Waard et al., 2019) using a masked language modeling objective. It has established strong performance on benchmarks such as the BIOSCAN-5M dataset (Gharaee et al., 2024) and is considered

Table 4: Optimal settings after hyperparameter search for the comparison study. The reported learning rate is during supervised learning/fine-tuning.

	Label Smoothing	Multi-head	Loss weighting	Learning Rate	Training Strategy
BLAST	N/A	N/A	N/A	N/A	Index&Query
MycoAI-CNN (Vu)	Hierarchical	✓	✓	1e-4	Fully Supervised
MycoAI-BERT (base)	Hierarchical	✓	✓	1e-4	Fully Supervised
CNN Encoder	None	✓	✓	8e-4	Fully Supervised
BarcodeBERT	None	✓	✓	1e-4	Fine-tuned
BarcodeMamba+	Hierarchical	✓	✓	8e-5	Pretrained, Fine-tuned
BarcodeMamba+ (large)	Hierarchical	✓	✓	8e-5	Pretrained, Fine-tuned

an effective architecture for insect biodiversity studies. For our experiments, we use the officially released pre-trained weights and fine-tune the model on our fungal ITS dataset.

A.4 BarcodeMamba+

BarcodeMamba+ is a foundation model adapted for the challenges of fungal ITS barcode classification. The model utilizes the BarcodeMamba architecture as its backbone (Gao & Taylor, 2024), a powerful SSM previously developed for general DNA sequence analysis.

Backbone Architecture. The BarcodeMamba backbone consists of a stack of n identical blocks. Each block processes the input sequence through three main components: a layer normalization step, a multi-layer perceptron, and a Mamba-2 mixing layer. The Mamba-2 layer is the core of the SSM, efficiently capturing long-range dependencies in the DNA sequence by mapping a d-dimensional input representation through a p-dimensional head. The final hidden states from the backbone serve as rich sequence representations.

Tokenizer. To convert raw DNA sequences into input embeddings for the backbone, we evaluated several tokenization strategies. While character-level (Nguyen et al., 2023) and k-mer-based tokenizers (Arias et al., 2023) have shown success on insect barcode datasets (Gao & Taylor, 2024), we integrated a Byte-Pair Encoding (BPE) tokenizer as recommended by Romeijn et al. (2024). BPE balances between the single-nucleotide resolution of character-level tokens and the pattern-capturing ability of k-mers, while also being vocabulary-efficient and robust to k-mer frameshift issues.

B Implementation Details

For the BLAST baseline, indexing the training set required 4.6 hours, and classification was performed using a best-hit approach. For the MycoAI-CNN and MycoAI-BERT models, we followed their official implementation[2], using the Adam optimizer with 1e-4 weight decay and training for 24 and 16 epochs respectively. All other models, including the CNN Encoder, BarcodeBERT, and our BarcodeMamba+, were trained using a cross-entropy loss and the AdamW optimizer (weight decay $= 0.1$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). We used a universal training strategy with an early stopping patience of 3 epochs on the validation loss and a 12-hour time limit. For our BarcodeMamba+, the fully supervised version was trained for 7 epochs with a learning rate (LR) of 8e-4. In the pre-train/fine-tune paradigm, the model was pre-trained for 15 epochs (LR=8e-4) and subsequently fine-tuned for 12 epochs with a decayed learning rate of 8e-5. The BarcodeBERT pretrained model was obtained from HuggingFace[3]. The model was fine-tuned with a learning rate of 1e-4 (both as recommended value as reported and the hyperparameter search). Fine-tuning was conducted for 1 epoch. The fully-supervised training process for the CNN Encoder was conducted over 3 epochs. Table 4 summarizes the enhancement settings used for all models.

[2] https://github.com/MycoAI/MycoAI

[3] https://huggingface.co/bioscan-ml/BarcodeBERT

Table 5: Statistical significance (p-values) of different model components on species-level metrics across three test sets. Both label smoothing types are compared against a baseline of no label smoothing. Significant results ($p < 0.05$) on a paired t-test are highlighted in bold.

Component	Test Set	Accuracy	Precision	Recall
Weighted loss	Yeast	**0.002**	**0.013**	**< 0.001**
	Filamentous	**0.004**	0.052	**0.003**
	MycoAI	**0.014**	**0.030**	**0.003**
Multi-head	Yeast	0.550	0.173	0.762
	Filamentous	**0.018**	0.053	**< 0.001**
	MycoAI	0.626	**< 0.001**	**0.004**
Hierarchical label smoothing	Yeast	**< 0.001**	**0.021**	**< 0.001**
	Filamentous	**0.017**	**0.020**	**0.010**
	MycoAI	**0.019**	**0.004**	**0.003**
Standard label smoothing	Yeast	0.093	0.315	0.081
	Filamentous	0.091	0.266	0.056
	MycoAI	0.095	0.070	**0.042**

C Statistical Tests

To establish whether ablated model components had a significant effect on the model performance, we conducted a series of paired t-tests for each component. We assumed each component would have an independent effect on the performance, and compared the accuracy, precision, or recall with one component present versus ablated, across all other component configurations. The results are shown in Table 5.

D Dataset Analysis

In Table 6, both Test Set 1 (Yeast) and Test Set 3 (MycoAI Benchmark) have high identical barcode overlap with the training data, at 86.73% and 100.00% respectively. This shows that they primarily measure model performance on in-distribution sequences. Conversely, Test Set 2 (Filamentous Fungi) shows minimal overlap at only 6.48%, establishing it as the most rigorous and reliable benchmark in this study for evaluating generalization capabilities on unseen species and barcodes.

Table 6: Analysis of dataset overlap between the training set and the three test sets. Overlap percentages are calculated relative to the unique species or barcodes in each test set, respectively.

	Species Overlap			**Identical Barcode Overlap**		
Test Set	**Total (Test)**	**Overlap (n)**	**Overlap (%)**	**Total (Test)**	**Overlap (n)**	**Overlap (%)**
Test Set 1: Yeast	1,157	616	53.24%	4,235	3,673	86.73%
Test Set 2: F. Fungi	5,537	2,650	47.86%	10,721	695	6.48%
Test Set 3: MycoAI	14,742	14,742	100.00%	363,420	363,420	100.00%