**SHORT TITLE FOR RUNNING HEAD: Fungal trophic mode classification**

**Automated Classification of Fungal Trophic Modes Using BioBERT: A Pilot Study**

**Beatrice M. Bock**1,2

1Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011, USA  
2Center for Adaptable Western Landscapes, Northern Arizona University, Flagstaff, AZ 86011, USA

**ABSTRACT**

Fungi exhibit striking diversity in trophic strategies, with some species capable of occupying multiple ecological roles while others are restricted to a single mode. Manual classification of trophic versatility from the scientific literature is time-consuming and subjective. Here, we tested the feasibility of using BioBERT, a biomedical language model, for automated classification of fungal trophic modes. We curated 56 fungal research abstracts, labeled as dual (occupying multiple modes) or solo (restricted to one mode), and fine-tuned BioBERT for binary classification. The model achieved 86% accuracy with balanced precision and recall and showed little bias across categories. This proof-of-concept demonstrates that language models can capture ecological nuance in fungal literature, providing a foundation for scaling to larger datasets and integrating natural language processing into mycological research. Future work will expand dataset size, incorporate multi-label classification, and link predictions to fungal trait databases. Code and trained models are openly available for community use.

**KEY WORDS**: fungal ecology, trophic modes, BioBERT, natural language processing, machine learning

**INTRODUCTION**

Fungi play critical roles in ecosystems as decomposers, pathogens, and symbionts, and many taxa exhibit flexibility in trophic strategy. Some species are capable of shifting between modes depending on context, while others remain restricted to a single role (Berbee et al. 2017; Martin and Tan 2025). This trophic versatility influences ecosystem processes, agricultural outcomes, and the resilience of fungal communities under environmental change.

Classification of fungal trophic modes typically relies on manual literature review (e.g. Brundrett and Tedersoo 2018), which is labor-intensive and difficult to scale as the volume of publications increases. Automated approaches using natural language processing (NLP) could accelerate this work by extracting and categorizing trophic information directly from research abstracts and full texts.

Transformer-based models such as BioBERT (Lee et al. 2020), trained on large biomedical corpora, have shown strong performance in text classification tasks. Their use in fungal ecology, however, remains largely unexplored. In this pilot study, we fine-tuned BioBERT for binary classification of fungal trophic modes from abstracts. Our objectives were to (1) assemble a small but balanced dataset of fungal abstracts labeled by trophic versatility, (2) train and evaluate a BioBERT model, and (3) assess the feasibility of extending this approach to larger datasets.

**MATERIALS AND METHODS**

**Dataset Curation --**

We compiled fungal research abstracts from Web of Science, focusing on taxa with well-documented ecological roles. Abstracts were manually labeled into two categories:

**Dual**: species reported to occupy more than one trophic mode (e.g., facultative pathogens that also decompose organic matter).

**Solo**: species limited to a single trophic mode (e.g., obligate symbionts or strict saprotrophs).

Ambiguous cases lacking explicit statements were excluded. The final dataset contained 56 abstracts (28 dual, 28 solo).

**Preprocessing and Training --**

Abstracts were cleaned and tokenized using BioBERT’s tokenizer with a maximum sequence length of 512 tokens. The dataset was split into training (34 abstracts) and testing (22 abstracts) using stratified sampling.

We fine-tuned BioBERT v1.1 (monologg/biobert\_v1.1\_pubmed) for binary sequence classification using the Hugging Face Transformers library (Wolf et al. 2020). Standard hyperparameters were used: learning rate 5 × 10⁻⁵, batch size 8, 10 epochs, optimizer AdamW (Paszke et al. 2019).

**Evaluation --**

Performance on the test set was assessed using accuracy, precision, recall, and F1-score. A confusion matrix was generated to evaluate classification errors (Fig. 1). Training and validation curves were inspected to confirm convergence (Fig. 2).

**RESULTS**

**Dataset Characteristics --**

The dataset was balanced between dual and solo labels. Abstracts ranged from 150–500 words (mean 360). The training set contained 34 abstracts (18 dual, 16 solo), and the test set contained 22 (10 dual, 12 solo).

**Model Performance --**

BioBERT fine-tuning converged efficiently, with no evidence of overfitting. On the held-out test set, the model achieved strong performance (Table 1):

**Table 1.** Classification performance metrics.

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 86.4% |
| Precision | 86.8% |
| Recall | 85.8% |
| F1-Score | 86.1% |

The confusion matrix (Fig. 1) showed balanced predictions across classes, with 19 correct predictions out of 22 test samples.

**DISCUSSION**

This pilot study shows that BioBERT can classify fungal trophic modes from scientific abstracts with high accuracy, even with a small dataset. The balanced performance across dual and solo categories suggests that the model can detect ecological nuance in text rather than overfitting to one label.

The main limitation is dataset size: 56 abstracts cannot capture the full range of fungal trophic strategies, and manual labeling introduces subjectivity. In addition, binary classification oversimplifies trophic modes, which often span a continuum. Nevertheless, this study provides an early demonstration that transformer-based NLP can be applied to fungal ecology.

Future directions include scaling to larger datasets, incorporating multi-label classification for fungi with context-dependent roles, and applying active learning to reduce annotation costs. Linking predictions to databases such as MycoBank or FUNGuild would enable integration with ecological and trait data.

Automated classification of fungal trophic modes could support meta-analyses, biodiversity monitoring, and predictive modeling in agriculture and conservation. By making fungal ecological data more accessible, this approach has the potential to accelerate research on fungal diversity and ecosystem function.

**DATA AVAILABILITY STATEMENT**

All code, trained models, and the labeled abstract dataset are available at: <https://github.com/beabock/biobert_dualsolo>

**ACKNOWLEDGEMENTS**

Thank you to my co-advisors, Dr. Nancy Johnson and Dr. Kitty Gehring, for their support and guidance on this, as well as on many other projects.

**LITERATURE CITED**

Berbee, M. L., James, T. Y., and Strullu-Derrien, C. 2017. Early diverging fungi: diversity and impact at the dawn of terrestrial life. Annual Review of Microbiology 71:41–60.

Brundrett, M. C., and Tedersoo, L. 2018. Evolutionary history of mycorrhizal symbioses and global host plant diversity. New Phytologist 220:1108–1115.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36:1234–1240.

Martin, F., and Tan, H. 2025. Saprotrophy-to-symbiosis continuum in fungi. Current Biology 35:R469–R475.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 2019. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12:2825–2830.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Davison, J. 2020. Transformers: state-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.

**LEGENDS**

Fig. 1. Confusion matrix of BioBERT predictions on the test set showing true versus predicted trophic mode labels (dual vs. solo).

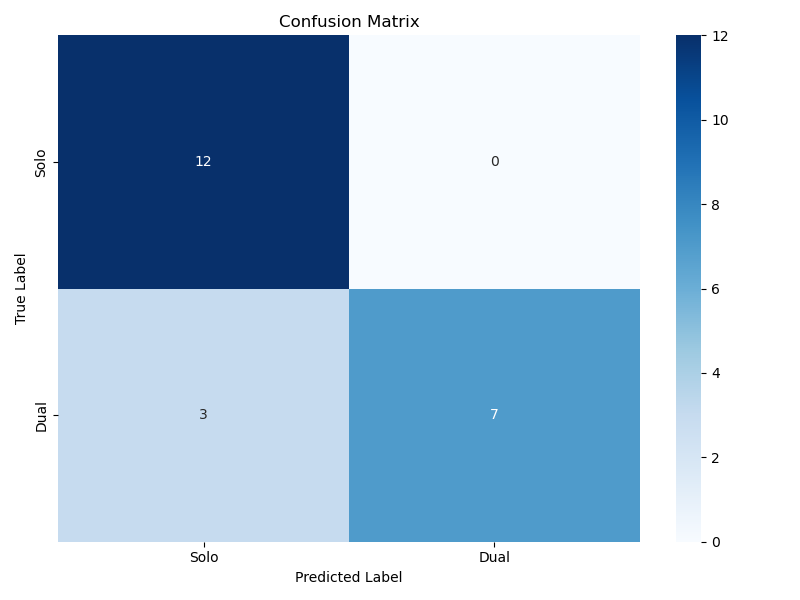
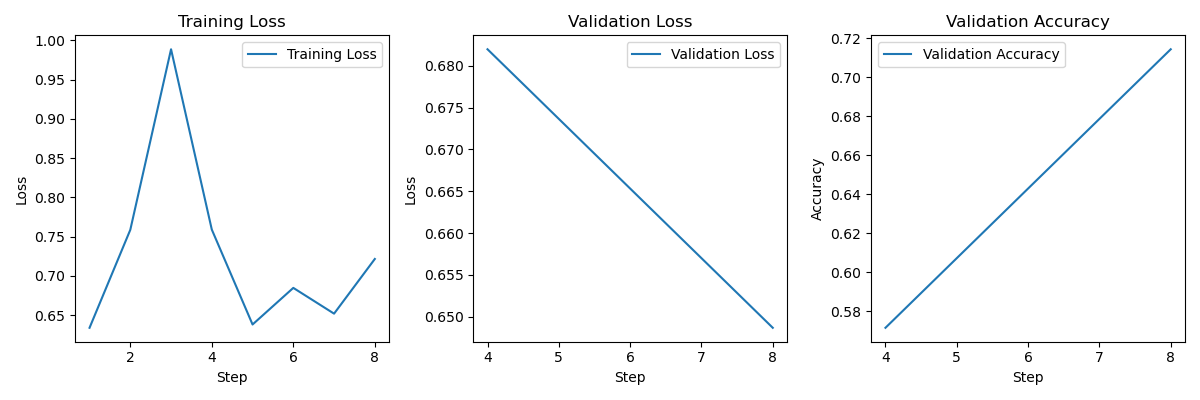


Fig. 2. Training curves showing training loss, validation loss, and validation accuracy over 10 epochs with incremental plotting for real-time monitoring.



Corresponding author: B. Bock; bmb646@nau.edu