**Linking Literature and Trait Databases: A Pilot Workflow for Automated Classification of Fungal Trophic Modes**

**Beatrice M. Bock**  
Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011, USA  
Center for Adaptable Western Landscapes, Northern Arizona University, Flagstaff, AZ 86011, USA  
*Corresponding author:* bmb646@nau.edu

**Abstract**

Fungi exhibit diverse trophic strategies, ranging from obligate symbiosis to saprotrophy, with some taxa capable of occupying multiple ecological roles. Manually classifying trophic versatility from the literature is time-consuming and difficult to scale. Here, we present a pilot workflow to automate classification of fungal trophic modes using BioBERT, a transformer-based language model, applied to scientific abstracts. We curated 56 fungal research abstracts labeled as dual (occupying multiple trophic modes) or solo (restricted to a single mode) and fine-tuned BioBERT for binary classification. The model achieved 86% accuracy with balanced precision and recall. This workflow demonstrates that machine learning can reproduce literature-based trait assignments and lays the foundation for scaling to larger datasets and integration with trait databases such as FUNGuild. All code and the curated dataset are openly available to support reproducibility and reuse.

**Keywords:** fungal ecology, trophic modes, natural language processing, machine learning, trait databases, BioBERT

**Introduction**

Fungi play critical roles in ecosystems as decomposers, pathogens, and symbionts, and many taxa exhibit flexibility in trophic strategy (Berbee et al. 2017). Some species shift among modes depending on environmental or host context, while others remain restricted to a single ecological role (Martin and Tan 2025). Understanding and cataloging fungal trophic versatility is critical for biodiversity assessments, ecosystem modeling, and agricultural applications where the trophic mode a fungus occupies (e.g. a pathogen vs. a saprotroph) is directly relevant to plant functioning.

Traditional classification of fungal trophic modes relies on manual literature review (e.g., Brundrett and Tedersoo 2018), which is labor-intensive and subjective. Automated natural language processing (NLP) approaches offer a scalable alternative by extracting ecological information directly from research abstracts or full texts.

Transformer-based language models, such as BioBERT, pretrained on biomedical corpora, have demonstrated strong performance in text classification tasks (Lee et al. 2020). Their application to fungal ecology remains largely unexplored. This pilot study aims to demonstrate the feasibility of using BioBERT to classify fungal trophic modes from abstracts and to provide a workflow that can be extended to larger datasets and linked with fungal trait databases such as FUNGuild.

**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 restricted 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 Model 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) sets using stratified sampling.

We fine-tuned BioBERT v1.1 (monologg/biobert\_v1.1\_pubmed) for binary sequence classification using the Hugging Face Transformers library (version 4.56.1) in Python 3.9.23 on Windows 11. The model was trained using PyTorch 2.8.0 and evaluated with scikit-learn 1.6.1. Hyperparameters were set to a learning rate of 5 × 10⁻⁵, batch size of 8, and 10 epochs, with the AdamW optimizer.

**Evaluation**

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

**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 abstracts (10 dual, 12 solo).

**Model Performance**

BioBERT fine-tuning converged efficiently with no evidence of overfitting. Classification performance metrics on the test set are summarized in Table 1.

**Table 1. BioBERT classification performance on fungal trophic modes.**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Note** |
| Accuracy | 86.4% | Fraction of correctly predicted labels |
| Precision | 86.8% | Positive predictive value across classes |
| Recall | 85.8% | True positive rate across classes |
| F1-Score | 86.1% | Harmonic mean of precision and recall |

**Discussion**

This pilot study demonstrates that transformer-based NLP models can reproduce literature-based assignments of fungal trophic modes. Even with a small dataset, BioBERT achieved balanced accuracy and precision, suggesting it captures ecological nuance in text.

The workflow highlights the potential for scaling: larger corpora of abstracts could be classified automatically, accelerating the construction of trait datasets for fungi. Integration with databases such as FUNGuild would enable automated linkage between literature-derived trophic information and broader ecological metadata. This approach can facilitate meta-analyses, predictive modeling, and biodiversity monitoring while reducing manual curation effort.

Limitations include dataset size, binary simplification of complex trophic continua, and the subjectivity inherent in manual labeling. Future work will expand dataset coverage, incorporate multi-label classification for context-dependent taxa, and evaluate integration with other trait resources.

**Data Availability**

The curated abstract dataset, code, and trained models are openly available via Zenodo: <https://doi.org/10.5281/zenodo.17343492>

**Acknowledgements**

Thank you to Dr. Nancy Johnson and Dr. Kitty Gehring for guidance and support on this and other projects.

**Literature Cited**

* Berbee, M. L., James, T. Y., & 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., & 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., & Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.
* Lee, J., Yoon, W., Kim, S., et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36:1234–1240.
* Martin, F., & Tan, H. 2025. Saprotrophy-to-symbiosis continuum in fungi. *Current Biology* 35:R469–R475.
* Wolf, T., Debut, L., Sanh, V., et al. 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.

**Figures**

**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.