

**PREPRINT**

*Author-formatted, not peer-reviewed document posted on 03/11/2025*

DOI: <https://doi.org/10.3897/arphapreprints.e176591>

---

**Automated extraction of fungal trophic modes from literature using BioBERT: an open pilot workflow**

 **Beatrice Bock**

# Automated extraction of fungal trophic modes from literature using BioBERT: an open pilot workflow

Beatrice Margareta Bock ‡, §

‡ Department of Biological Sciences, Northern Arizona University, Flagstaff, United States of America

§ Center for Adaptable Western Landscapes, Northern Arizona University, Flagstaff, United States of America

Corresponding author: Beatrice Margareta Bock ([beabockm@gmail.com](mailto:beabockm@gmail.com))

Reviewed v 1

Academic editor: Editorial Secretary

## Abstract

Fungi exhibit diverse trophic strategies, ranging from obligate symbiosis to saprotrophy, with some taxa capable of occupying multiple ecological roles. Manually identifying trophic versatility from literature is time-consuming and difficult to scale. Here, we present a pilot workflow that automates the classification of fungal trophic modes using transformer-based language models. A curated dataset of 56 fungal ecology abstracts was manually labelled as dual (occupying multiple trophic modes) or solo (restricted to one mode) and used to fine-tune four models: BioBERT, BERT-base-cased, BERT-base-uncased and BiodivBERT. Stratified 5-fold cross-validation revealed that BioBERT and BERT-base-cased performed equally well (~ 89% accuracy, balanced precision and recall), highlighting the importance of case sensitivity in taxonomic text. BiodivBERT and uncased BERT models underperformed, indicating that domain adaptation alone is not sufficient. This pilot study emphasises reproducibility, transparency and open data integration, offering a generalisable proof-of-concept for linking literature-derived ecological information to existing fungal trait databases such as FUNGuild and FungalTraits. All code and data are openly available to support reuse and scaling to larger datasets.

## Keywords

fungal ecology, trophic modes, natural language processing, machine learning, trait databases, BioBERT, saprotrophy-symbiosis continuum

## Introduction

Fungi play essential roles in ecosystems as decomposers, pathogens and symbionts, and many taxa exhibit flexibility across these trophic modes (Berbee et al. 2017). Some species occupy multiple ecological strategies depending on host identity or environmental context, while others remain specialised to a single mode (Martin and Tan 2025). Understanding this trophic versatility is critical for biodiversity assessments, ecosystem modelling and agricultural applications where a fungus' ecological role affects plant performance.

Trait databases, such as FUNGuild (Nguyen et al. 2016) and FungalTraits (Pöhlme et al. 2021), have advanced fungal functional annotation at scale, but manual literature extraction remains a bottleneck. Human-curated classifications are time-intensive, subjective and limited by the accessibility of trait-relevant language in publications. Moreover, trait databases are valuable but can be limited in applicability, as annotations are often performed at the genus or family level despite often substantial interspecific variability (Violle et al. 2015).

Natural language processing (NLP) offers a scalable way to extract trait-relevant information directly from text. Transformer-based models, such as BioBERT, pretrained on large biomedical corpora, excel at contextual understanding and have achieved state-of-the-art results in various text classification tasks (Lee et al. 2019). However, their use in fungal ecology and trait data integration remains largely unexplored.

This pilot study tests the feasibility of fine-tuning BioBERT to classify fungal trophic modes from abstracts. The workflow is designed for transparency and future scaling, providing a reproducible pipeline that can complement or benchmark existing trait databases. By linking automated text classification with open trait resources, this work demonstrates a path towards more consistent, interoperable fungal functional data.

## Related Work

Recent advances in domain-specific NLP illustrate the potential for scaling and refining workflows like this one. BiodivBERT (Abdelmageed et al. 2023) represents the first pretrained language model tailored specifically for biodiversity research, achieving significant gains in named entity recognition and relation extraction. Its development from life-sciences corpora highlights how domain adaptation can substantially improve the precision and recall of ecological information retrieval. Similarly, Cornelius et al. (2025) demonstrated a machine-learning framework for extracting arthropod organismal traits, translating unstructured text into a machine-actionable database (ArTraDB). Their approach shows how targeted trait extraction can efficiently transform literature into structured ecological data.

Parallel work in plant functional ecology supports the broader utility of transformer-based extraction for trait data. Domazetoski et al. (2025) developed a natural language pipeline that automatically identifies both categorical and numerical plant traits from unstructured descriptions with high precision and recall. This scalability across morphological, life history and functional trait types underscores the potential to adapt such methods to fungal traits, where similar data gaps persist.

Beyond ecological trait extraction, innovations in model architecture and pretraining further extend applicability. BioT5 (Pei et al. 2023) introduced cross-modal pretraining to connect textual and molecular data using chemically informed representations, an approach that could eventually allow integration of genomic or metabolomic predictors into ecological models. ModernBERT (Warner et al. 2025) provides a computationally efficient encoder capable of handling long-context inputs, which is particularly relevant for large-scale biodiversity corpora. Together, these efforts suggest practical pathways to scale fungal trait text mining beyond small datasets, while maintaining interpretability and reproducibility. Foundational work by Gu et al. (2021) also reinforces the importance of domain-specific pretraining, demonstrating that models trained entirely within a target domain outperform general-domain models adapted later.

## Materials and Methods

### Dataset Curation

Fig. 1 shows the overall pipeline from data collection to evaluation. Fungal research abstracts were retrieved from the Web of Science Core Collection on 11 September 2025, using two Boolean search queries designed to capture fungi with distinct lifestyle classifications. For solo (single trophic mode) examples, we searched: '("obligate mycorrhizal" OR "strictly endophytic" OR "exclusive saprotroph") AND fungus' (119 results). For dual (multiple trophic mode) examples, we searched: '("dual lifestyle" OR "facultative lifestyle" OR "dual trophic mode" OR "lifestyle switching" OR "endophyte-saprotroph" OR "plant-associated saprotroph") AND fungi' (70 results).

From these 189 candidate articles, abstracts were manually reviewed by a single labeller (BMB) and 56 were selected, based on: (1) unambiguous description of trophic mode in the abstract text; (2) English language and (3) no duplicates between searches. Abstracts were labelled as:

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

Solo: taxa restricted to a single trophic mode (e.g. obligate symbionts or strict saprotrophs).

Ambiguous abstracts without explicit trophic mode statements were excluded. The final dataset contained 56 abstracts (28 dual, 28 solo). The dataset was balanced between

dual and solo classes, with abstract lengths ranging from 150–500 words (mean 360). A supplementary file listing the permissible trophic mode labels and their definitions is available in the repository (`datasets/trophic_mode_labels.md`).

## Preprocessing and Model Training

Abstracts were cleaned and tokenised using model-specific tokenisers (maximum sequence length: 512 tokens). Token length analysis revealed that three of 56 abstracts (5.4%) exceeded this limit and were truncated, though truncation occurred at the end of abstracts where contextual information for classification is typically less concentrated.

We compared four transformer-based language models to assess the impact of domain-specific pretraining and case sensitivity: (1) BERT-base-uncased ('google-bert/bert-base-uncased'; Devlin et al. (2018)), (2) BERT-base-cased ('google-bert/bert-base-cased'), (3) BioBERT v.1.1 ('monologg/biobert\_v.1.1\_pubmed', biomedical domain-adapted; Lee et al. (2019)) and (4) BiodivBERT ('NoYo25/BiodivBERT', biodiversity domain-adapted; Abdelmageed et al. (2023)). All models were fine-tuned for binary sequence classification using the Hugging Face Transformers library (v.4.40.0; Wolf (2020)) in Python 3.9 (Python Software Foundation 2020).

To maximise statistical robustness with the small dataset, we employed stratified 5-fold cross-validation on all 56 abstracts rather than a single train-test split. This approach ensures that each abstract is used once for validation while maintaining class balance across folds (seed = 42 for reproducibility). Models were trained using identical hyperparameters: learning rate =  $5 \times 10^{-5}$ , batch size = 8, maximum epochs = 20 with early stopping (patience = 3 epochs), dropout = 0.2, optimiser = AdamW with class-weighted loss to account for any fold-level imbalances (PyTorch v.2.0.1; Paszke et al. (2019)). Training was executed on NAU's Monsoon HPC cluster using Tesla K80 GPUs with CUDA 11.4.

## Evaluation

Performance was evaluated across all five folds for each model. For each fold, accuracy, precision, recall and F1-score were computed using scikit-learn v.1.4.2 (Pedregosa et al. 2011) and results were aggregated as mean  $\pm$  standard deviation. Precision, recall and F1-score are reported as macro averages (unweighted mean across both classes), which treats solo and dual classes equally and is appropriate for balanced binary classification.

Comparative visualisations include: (1) model performance bar charts with error bars representing cross-fold variation (Fig. 2); (2) aggregated confusion matrices for BioBERT showing cumulative predictions across all folds (Fig. 3) and (3) training time comparison across models (Fig. 4). Error analysis identified abstracts misclassified by multiple models, highlighting particularly ambiguous cases (`results/error_analysis.csv`).

## Code and Reproducibility

All code and data are openly available via: [Zenodo](#) and [GitHub](#) (Bock 2026).

The repositories include scripts for dataset curation, model fine-tuning and evaluation, along with example outputs and documentation to facilitate reuse or adaptation for other trait-related classification tasks.

## Results

### Model Performance

We compared four transformer-based language models using stratified 5-fold cross-validation on all 56 abstracts: (1) BERT-base-uncased; (2) BERT-base-cased; (3) BioBERT v.1.1 (biomedical domain-adapted) and (4) BiodivBERT (biodiversity domain-adapted). All models were trained with identical hyperparameters (20 epochs maximum with early stopping, learning rate  $5e-5$ , batch size 8, dropout 0.2).

BioBERT achieved the highest overall performance (F1 =  $0.892 \pm 0.120$ , Accuracy =  $0.894 \pm 0.116$ ), marginally outperforming BERT-base-cased (F1 =  $0.892 \pm 0.100$ , Accuracy =  $0.892 \pm 0.100$ ). Case sensitivity proved critical: cased models substantially outperformed their uncased counterparts (BERT-base-uncased: F1 =  $0.700 \pm 0.241$ , Accuracy =  $0.749 \pm 0.177$ ), likely because taxonomic nomenclature capitalisation provides important classification signals. Surprisingly, BiodivBERT underperformed (F1 =  $0.747 \pm 0.198$ , Accuracy =  $0.771 \pm 0.166$ ) despite its biodiversity-specific pre-training, suggesting that domain alignment alone does not guarantee superior performance on specialised classification tasks.

Classification metrics for BioBERT are summarised in Table 1. Results are reported as mean  $\pm$  standard deviation across five folds. Comparative model performance (Fig. 2) demonstrates that cased models substantially outperform uncased variants, while BioBERT and BERT-cased achieve statistically equivalent results. The aggregated confusion matrices (Fig. 3) shows balanced performance across both classes for BioBERT and BERT-cased. Training efficiency varied substantially (Fig. 4), with BioBERT and BERT-cased completing training in  $\sim 10$ -11 minutes while BiodivBERT and BERT-uncased required  $\sim 35$  minutes, likely due to differences in tokenisation efficiency and convergence patterns.

## Discussion

This pilot study demonstrates that transformer-based NLP models can extract ecological information embedded in scientific text. With a small but carefully curated dataset, BioBERT and BERT-base-cased achieved  $\sim 89\%$  accuracy in classifying fungal trophic

modes, indicating that pretrained language models, whether biomedical or general-purpose, can generalise to ecological contexts with minimal fine-tuning.

The comparative model analysis revealed three important insights. First, BioBERT's marginal advantage over BERT-base-based suggests that biomedical domain adaptation provides limited benefit for this specific task, possibly because fungal ecology vocabulary differs substantially from clinical biomedical text. Second, case sensitivity proved critical: models trained on cased text outperformed uncased variants by ~ 15 percentage points, likely because taxonomic nomenclature capitalisation (e.g. *\*Fusarium\** vs. *\*fusarium\**) carries important classification signals. Third, BiodivBERT's underperformance, despite biodiversity-specific pre-training, indicates that domain alignment alone does not guarantee superior performance; the pre-training corpus must closely match the downstream task domain.

The workflow presented here provides a proof-of-concept for trait-orientated text mining in fungi. Discrepancies between automated and curated classifications can highlight ambiguous or conflicting entries, which is particularly important given the known limitations in trait data provenance (Violle et al. 2015). When scaled, this approach may support more dynamic updating and cross-validation of trait resources such as FUNGuild and FungalTraits.

Error analysis across all models (figures/error\_analysis.csv) reveals that 39 total misclassifications occurred across 280 predictions (4 models × 5 folds × 56/5 samples per fold). Seven abstracts were misclassified by at least two models, suggesting inherent ambiguity in how trophic modes are described in these cases. For example, abstracts describing endophytes with both plant-beneficial and saprotrophic capabilities proved challenging for all models, likely because the text emphasises ecological context over explicit trophic mode labels. These consistently problematic cases highlight where human curation remains essential and where future annotation guidelines could improve clarity. Improving model learning, such as by increasing the dataset size, refining annotation or using more advanced models, may also help resolve these ambiguous cases in future work.

## Limitations

This pilot study has several deliberate constraints. The small sample size (56 abstracts) limits statistical power and generalisability, though stratified cross-validation helps mitigate overfitting concerns. The binary simplification (solo vs. dual) reduces the rich spectrum of fungal trophic strategies to a coarse categorisation; real ecological roles often exist along gradients rather than discrete classes. Manual label subjectivity by a single labeller may introduce bias, though this was mitigated by conservative inclusion criteria requiring explicit trophic mode statements.

## Future Work

Several directions could extend this pilot into a more comprehensive tool:

1. Expanding dataset scope: Increasing the corpus to hundreds or thousands of abstracts would improve model robustness and allow detection of rarer trophic patterns. This could include full-text articles rather than abstracts alone and taxa beyond fungi to establish cross-kingdom applicability;

2. Multi-label classification for ecological gradients: Rather than binary solo/dual classification, future models could predict specific trophic modes (e.g. saprotroph, symbiont, pathogen, endophyte) as non-exclusive labels. This would capture the continuous nature of ecological roles and better reflect biological reality where taxa may simultaneously occupy multiple niches;

3. Integration with environmental metadata: Linking text-derived traits with geographic, climatic or substrate data could enable context-aware predictions; for example, predicting how a taxon's trophic mode might shift under different environmental conditions;

4. Genomic and metabolomic predictors: Combining textual information with molecular data (e.g. gene content, secondary metabolite profiles) could improve prediction accuracy and provide mechanistic insights into trophic flexibility;

5. Domain-specific pretraining: Training a BERT-style model from scratch on ecological and mycological corpora (following the BiodivBERT approach) could yield better performance than adapting biomedical models.

## Conclusions

This pilot study demonstrates that transformer-based NLP can successfully extract fungal trophic mode information from scientific abstracts. BioBERT and BERT-cased achieved ~89% accuracy in classifying abstracts as describing single or multiple trophic modes, validating the feasibility of automated trait extraction for fungal ecology. The key contributions of this work are:

1. Proof-of-concept: Pretrained biomedical language models can generalise to ecological classification tasks with minimal fine-tuning;

2. Reproducible workflow: All code and data are openly available, enabling replication and extension by other researchers;

3. Trait database integration: The approach complements existing resources like FUNGuild and FungalTraits by providing a scalable method to flag taxa with lifestyle plasticity. As fungal trait databases continue to grow in importance for biodiversity assessments and ecosystem modelling, automated text mining offers a path towards more efficient, consistent and comprehensive trait annotation. This workflow provides a foundation for scaling to larger datasets and more nuanced ecological classifications.

## Acknowledgements

Thank you to Dr. Nancy Johnson and Dr. Kitty Gehring for guidance and support on this and other projects. Special thanks to Anne-Marie Cooper, as well.

## Conflicts of interest

The authors have declared that no competing interests exist.

## References

- Abdelmageed N, Löffler F, König-Ries B (2023) BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain. In: Yamaguchi A, et al. (Ed.) SWAT4HCLS. 62-71 pp. URL: <https://ceur-ws.org/Vol-3415/paper-7.pdf>
- Berbee M, James T, Strullu-Derrien C (2017) Early Diverging Fungi: Diversity and Impact at the Dawn of Terrestrial Life. *Annual Review of Microbiology* 71 (1): 41-60. <https://doi.org/10.1146/annurev-micro-030117-020324>
- Bock B (2026) beabock/biobert\_dualsolo: Reproducible BioBERT & BERT model comparison (4 models, 5-fold CV). Zenodo <https://doi.org/10.5281/zenodo.17343492>
- Cornelius J, Detering H, Lithgow-Serrano O, Agosti D, Rinaldi F, Waterhouse R (2025) From literature to biodiversity data: mining arthropod organismal traits with machine learning. *Biodiversity Data Journal* 13 <https://doi.org/10.3897/bdj.13.e153070>
- Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR URL: <http://arxiv.org/abs/1810.04805>
- Domazetoski V, Kreft H, Bestova H, Wieder P, Koynov R, Zarei A, Weigelt P (2025) Using large language models to extract plant functional traits from unstructured text. *Applications in Plant Sciences* 13 (3). <https://doi.org/10.1002/aps3.70011>
- Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H (2021) Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare* 3 (1): 1-23. <https://doi.org/10.1145/3458754>
- Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36 (4): 1234-1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Martin F, Tan H (2025) Saprotrophy-to-symbiosis continuum in fungi. *Current Biology* 35 (11). <https://doi.org/10.1016/j.cub.2025.01.032>
- Nguyen N, Song Z, Bates S, Branco S, Tedersoo L, Menke J, Schilling J, Kennedy P (2016) FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology* 20: 241-248. <https://doi.org/10.1016/j.funeco.2015.06.006>
- Paszke A, Gross S, Massa F, Lerer G, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Chintala S, et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32: 8024-8035.

- 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, Duchesnay E (2011) Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research* 12: 2825-2830.
- Pei Q, Zhang W, Zhu J, Wu K, Gao K, Wu L, Xia Y, Yan R (2023) BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* <https://doi.org/10.18653/v1/2023.emnlp-main.70>
- Pölme S, Abarenkov K, Henrik Nilsson R, Lindahl B, Clemmensen KE, Kausrud H, Nguyen N, Kjölller R, Bates S, Baldrian P, Frøslev TG, Adojaan K, Vizzini A, Suija A, Pfister D, Baral H, Järv H, Madrid H, Nordén J, Liu J, Pawlowska J, Pöldmaa K, Pärtel K, Runnel K, Hansen K, Larsson K, Hyde KD, Sandoval-Denis M, Smith M, Toome-Heller M, Wijayawardene N, Menolli N, Reynolds N, Drenkhan R, Maharachchikumbura SN, Gibertoni T, Læssøe T, Davis W, Tokarev Y, Corrales A, Soares AM, Agan A, Machado AR, Argüelles-Moyao A, Detheridge A, de Meiras-Ottoni A, Verbeken A, Dutta AK, Cui B, Pradeep CK, Marín C, Stanton D, Gohar D, Wanasinghe D, Otsing E, Aslani F, Griffith G, Lumsch T, Grossart H, Masigol H, Timling I, Hiiesalu I, Oja J, Kupagme J, Geml J, Alvarez-Manjarrez J, Ilves K, Loit K, Adamson K, Nara K, Küngas K, Rojas-Jimenez K, Bitenieks K, Irinyi L, Nagy L, Soonvald L, Zhou L, Wagner L, Aime MC, Öpik M, Mujica MI, Metsoja M, Ryberg M, Vasar M, Murata M, Nelsen M, Cleary M, Samarakoon M, Doilom M, Bahram M, Hagh-Doust N, Dulya O, Johnston P, Kohout P, Chen Q, Tian Q, Nandi R, Amiri R, Perera RH, dos Santos Chikowski R, Mendes-Alvarenga R, Garibay-Orijel R, Gielen R, Phookamsak R, Jayawardena R, Rahimlou S, Karunarathna S, Tibpromma S, Brown S, Sepp S, Mundra S, Luo Z, Bose T, Vahter T, Netherway T, Yang T, May T, Varga T, Li W, Coimbra VRM, de Oliveira VRT, de Lima VX, Mikryukov V, Lu Y, Matsuda Y, Miyamoto Y, Kõljalg U, Tedersoo L (2021) FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. *Fungal Diversity* 105 (1): 1-16. <https://doi.org/10.1007/s13225-020-00466-2>
- Python Software Foundation (2020) Python 3.9.0. URL: <https://docs.python.org/3.9>
- Violle C, Borgy B, Choler P (2015) Trait databases: misuses and precautions. *Journal of Vegetation Science* 26 (5): 826-827. <https://doi.org/10.1111/jvs.12325>
- Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, Gallagher A, Biswas R, Ladhak F, Aarsen T, Adams GT, Howard J, Poli I (2025) Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2526-2547. <https://doi.org/10.18653/v1/2025.acl-long.127>
- Wolf T, 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*. Online. Association for Computational Linguistics URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>

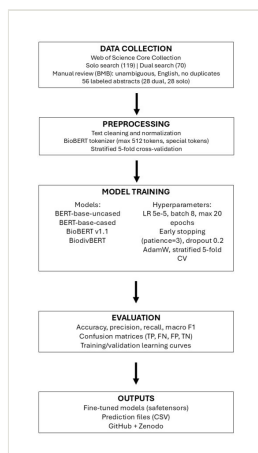


Figure 1.

Workflow diagram of the classification pipeline. The diagram summarises the full end-to-end process: literature search (Web of Science queries), manual curation (56 labelled abstracts), preprocessing (text cleaning, tokenisation with truncation at 512 tokens and token-length QC), stratified 5-fold cross-validation, model fine-tuning across four models (BERT-base-uncased, BERT-base-cased, BioBERT v.1.1, BiodivBERT) with standardised hyperparameters, evaluation (metrics, confusion matrices, learning curves) and outputs (fine-tuned models, predictions).

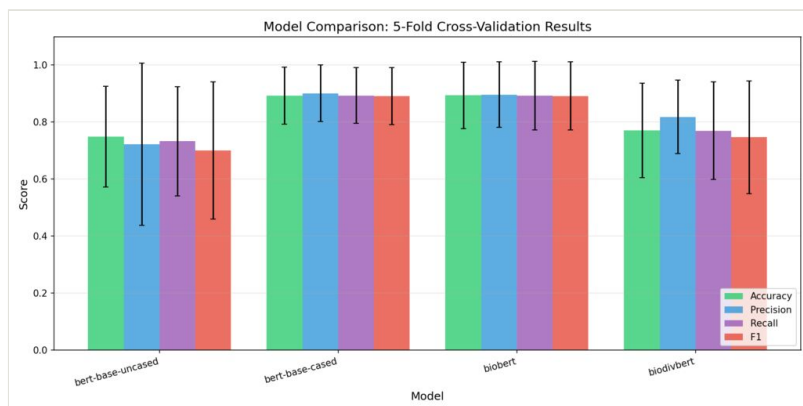


Figure 2.

Comparative model performance across four transformer-based architectures. Bar charts show mean classification metrics (accuracy, precision, recall, F1-score)  $\pm$  standard deviation from stratified 5-fold cross-validation ( $n = 56$  abstracts). BioBERT (biomedical domain-adapted) and BERT-base-cased achieved statistically equivalent performance ( $\sim 89\%$  accuracy), substantially outperforming BERT-base-uncased ( $\sim 75\%$ ) and BiodivBERT ( $\sim 77\%$ ). Case sensitivity proved critical, with cased models outperforming uncased by  $\sim 15$  percentage points. Metrics are calculated as macro averages (unweighted mean across dual and solo classes).

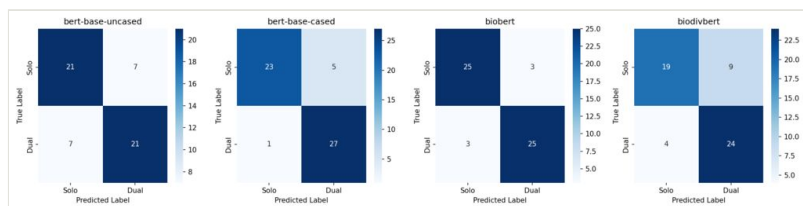


Figure 3.

Aggregated confusion matrices for all four models (BioBERT, BERT-base-cased, BERT-base-uncased, BiodivBERT) across all five folds (total 56 predictions per model). Each matrix shows true labels (Solo = single trophic mode, Dual = multiple trophic modes) versus predicted labels, allowing direct comparison of error patterns and class balance for each model. BioBERT and BERT-base-cased show balanced performance, while uncased and BiodivBERT models display more misclassifications. Colour intensity indicates prediction frequency; diagonal cells represent correct predictions.

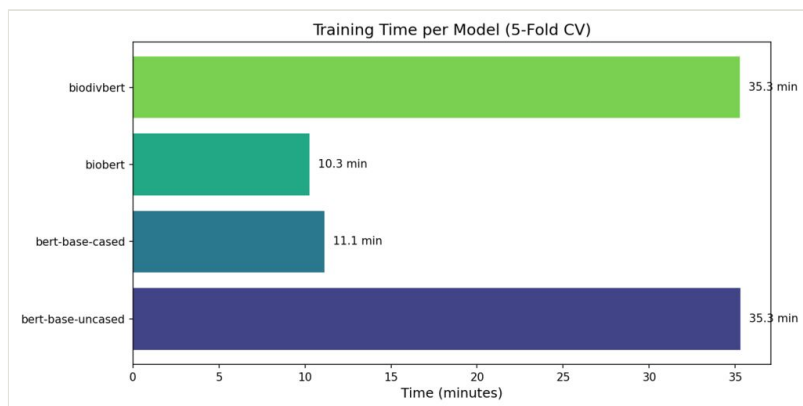


Figure 4.

Training time comparison across models. BioBERT and BERT-base-cased completed 5-fold cross-validation training in ~ 10-11 minutes, while BiodivBERT and BERT-base-uncased required ~ 35 minutes. Differences likely reflect tokenisation efficiency and convergence patterns rather than model size (all models have ~ 110M parameters). Faster convergence in cased models correlates with higher classification accuracy, suggesting that case-preserving tokenisation provides stronger learning signals for this taxonomic text classification task. Training performed on NAU Monsoon HPC cluster (Tesla K80 GPU, CUDA 11.4).

Table 1.

BioBERT classification performance on fungal trophic modes (5-fold CV). Precision, recall and F1-score are reported as macro averages (unweighted mean across both classes), which is appropriate for balanced binary classification and treats both classes equally regardless of support.

Metric	Value	Note
Accuracy	89.4% ± 11.6%	Fraction of correctly predicted labels
Precision	89.9% ± 11.5%	Positive predictive value (macro average)
Recall	88.8% ± 12.4%	True positive rate (macro average)
F1-Score	89.2% ± 12.0%	Harmonic mean of precision and recall (macro average)