Endophyte Systematic Review: Summary of Findings

Background:

- People often make the statement that all plants host endophytes
- This statement is either uncited or cites a source that does not find this
- Cited sources are often papers that find fungal endophytes in some plants

Question:

- Is it true that fungal endophytes are present in all plant species?
- Approach:
  - Download all abstracts from WoS, Scopus (and maybe also PubMed) about fungal endophytes in plants.
  - Train a model to sort between Relevant and Irrelevant abstracts
  - Train a model to sort between Presence (fungal endophytes found) and Absence (no fungal endophytes found)
  - Challenge: There are very few (<10) papers that do not find fungal endophytes in plants
  - Survey the plant lineages that do have evidence of fungal endophytes

1. Methods

Data Collection

- Literature search conducted (11/18/24)
- Search terms: ("fungal endophyte" OR "fungal endophytes" OR "endophytic fungi" OR "endophytic fungus") AND plant
- Databases: Web of Science and Scopus
- Will do another pull of the literature once this pipeline is approved by collaborators. Scopus is currently not allowing large downloads so also waiting on that. With the new search, I'm trying to capture historic names of endophytes and other terms for endophytes, as well as ways plants might be described in an abstract without saying plant. **Open to suggestions on improving this**

search

- Search terms: (
  "fungal endophyte" OR "fungal endophytes" OR "endophytic fungus" OR "endophytic fungi" OR "latent fungus" OR "latent fungi" OR "systemic fungus" OR "systemic fungi" OR "internal fungi" OR "resident fungi" OR "seed-borne fungi" OR "seed-transmitted fungi" OR "dark septate endophyte" OR "dark septate fungi" OR "DSE fungi"
)
  AND
  (
  plant* OR moss* OR bryophyte* OR liverwort* OR hornwort* OR fern* OR lycophyte* OR pteridophyte* OR tree* OR forest* OR shrub* OR grass* OR graminoid* OR herb* OR crop* OR seedling* OR sapling* OR seed* OR root* OR leaf* OR foliage OR shoot* OR stem* OR twig* OR rhizome* OR thallus OR frond* OR algae OR "green alga*" OR macroalga* OR
  "red alga*" OR "brown alga*" OR hydrophyte* OR kelp OR seaweed* OR seagrass* OR cyanobacteria OR cyanobiont* OR photobiont* OR lichen*
)
- Databases: Web of Science, Scopus, and PubMed
- Scripts: `Combo_abstracts.R` and `Combo_abstracts_pull2.R`

Machine Learning Pipeline

Detailed ML Workflow: Relevance and Presence/Absence Classification

Training Dataset Overview

Label	Count
Absence	117
Both	174
Other	73
Presence	312
Review	141

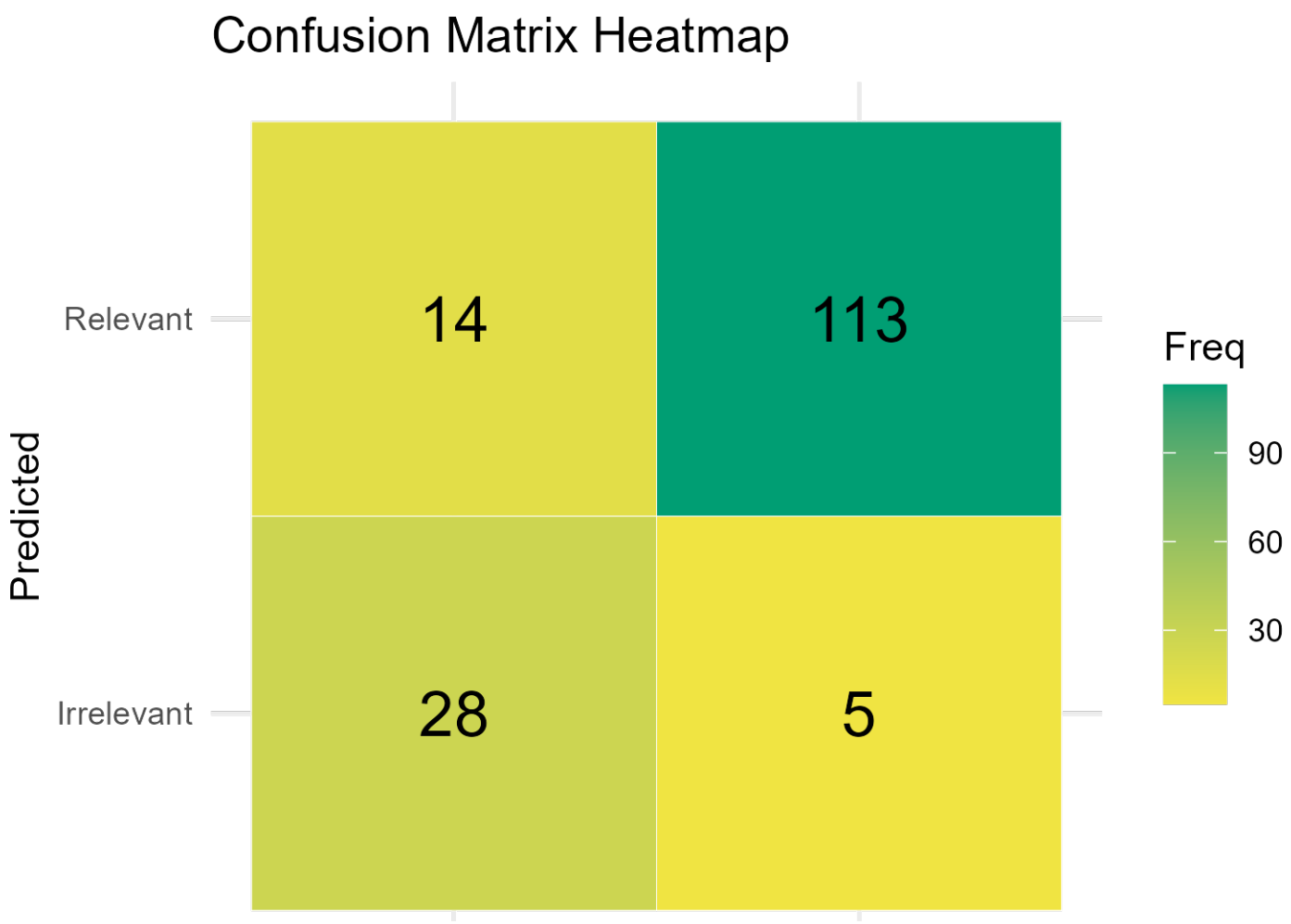
Training Dataset Details

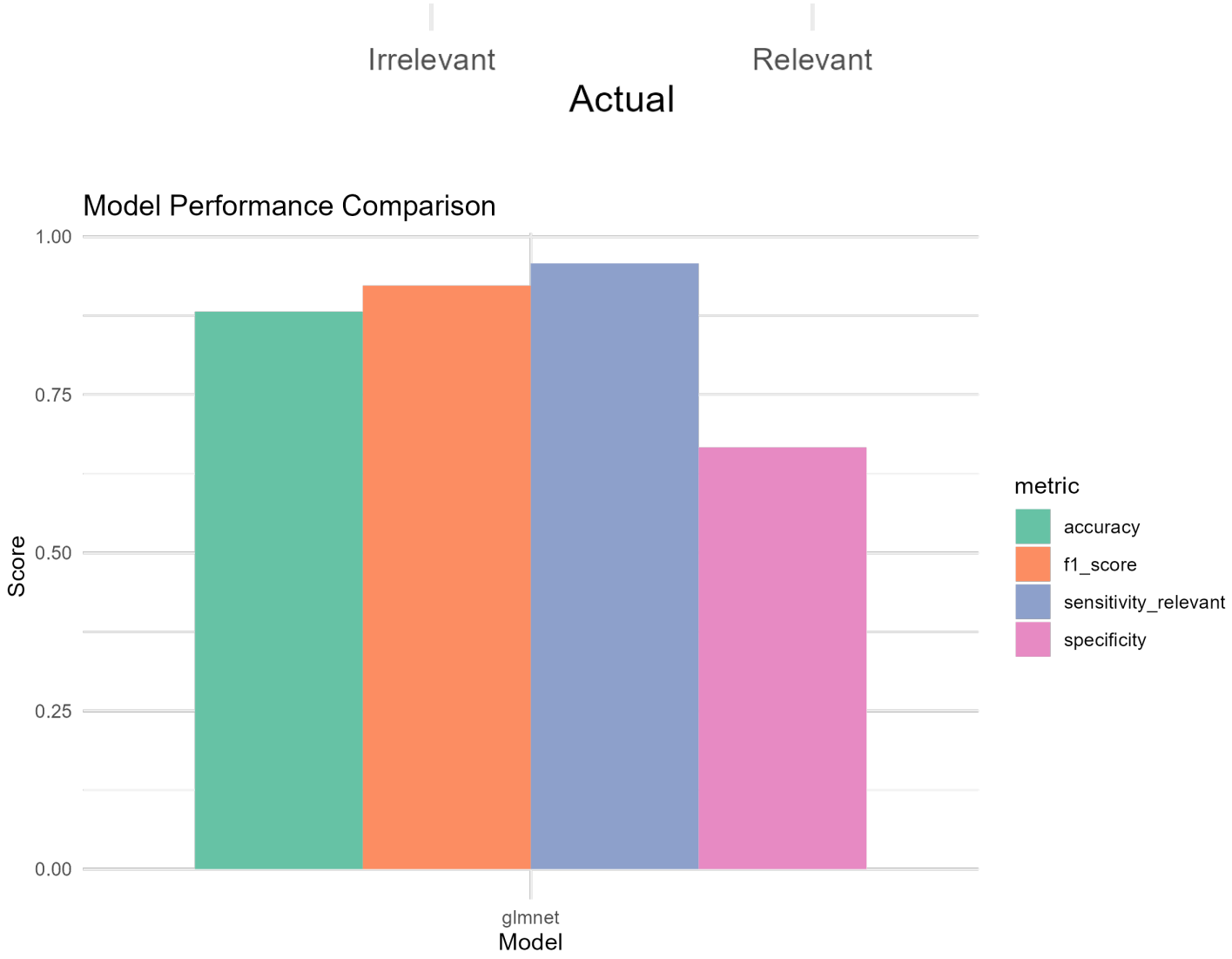
- I manually labeled all of those abstracts

- There are only a few actual Absence examples in the literature, so I fed those to ChatGPT and asked for fake Absence abstracts to increase my training dataset.

1. Relevance Classification

- Model: GLMNet (Elastic Net regularized logistic regression)
- Training: Distinguishes "Relevant" vs "Irrelevant" abstracts using labeled data.
- Relevant = Absence, Presence, Both (Presence)
- Irrelevant = Other, Review
- Feature Engineering: Abstracts are tokenized, stop words removed, and a document-term matrix (DTM) is constructed and harmonized to match training vocabulary.
- Prediction: Probabilities for "Relevant" and "Irrelevant" are predicted for all unlabeled abstracts.
- Thresholding: Multiple thresholds applied (Loose: 0.5, Medium: 0.6, Strict: 0.8). Abstracts labeled as "Relevant", "Irrelevant", or "Uncertain".
- Outputs:
  - `relevance_preds.csv` : All abstracts with predicted relevance probabilities and labels.
  - `irrelevant_uncertain_abstracts.csv` : Abstracts labeled as "Irrelevant" or "Uncertain" (for manual review).



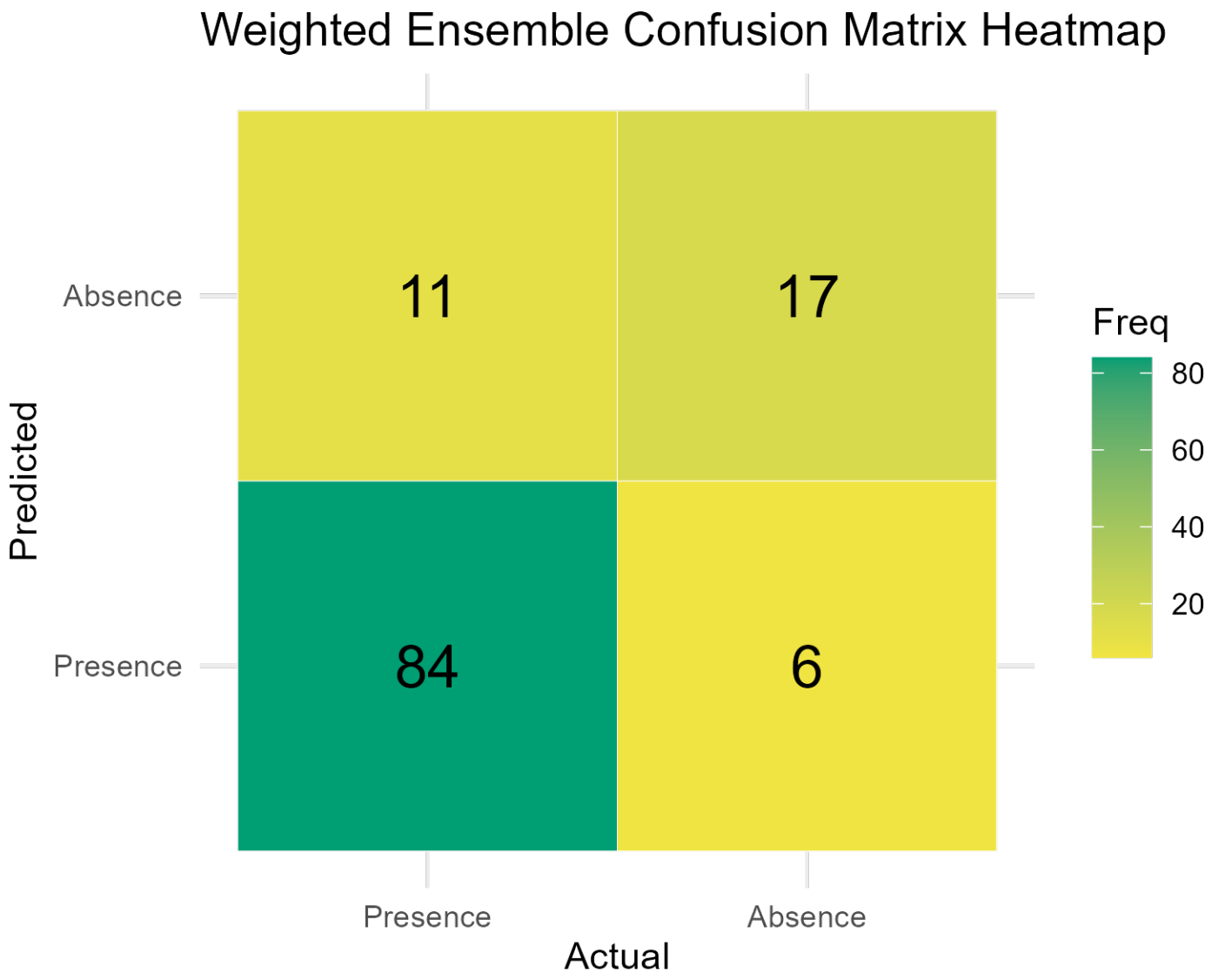


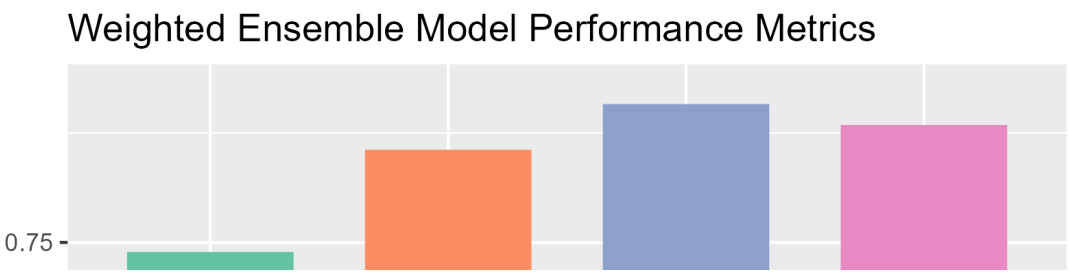
2. Presence/Absence Classification

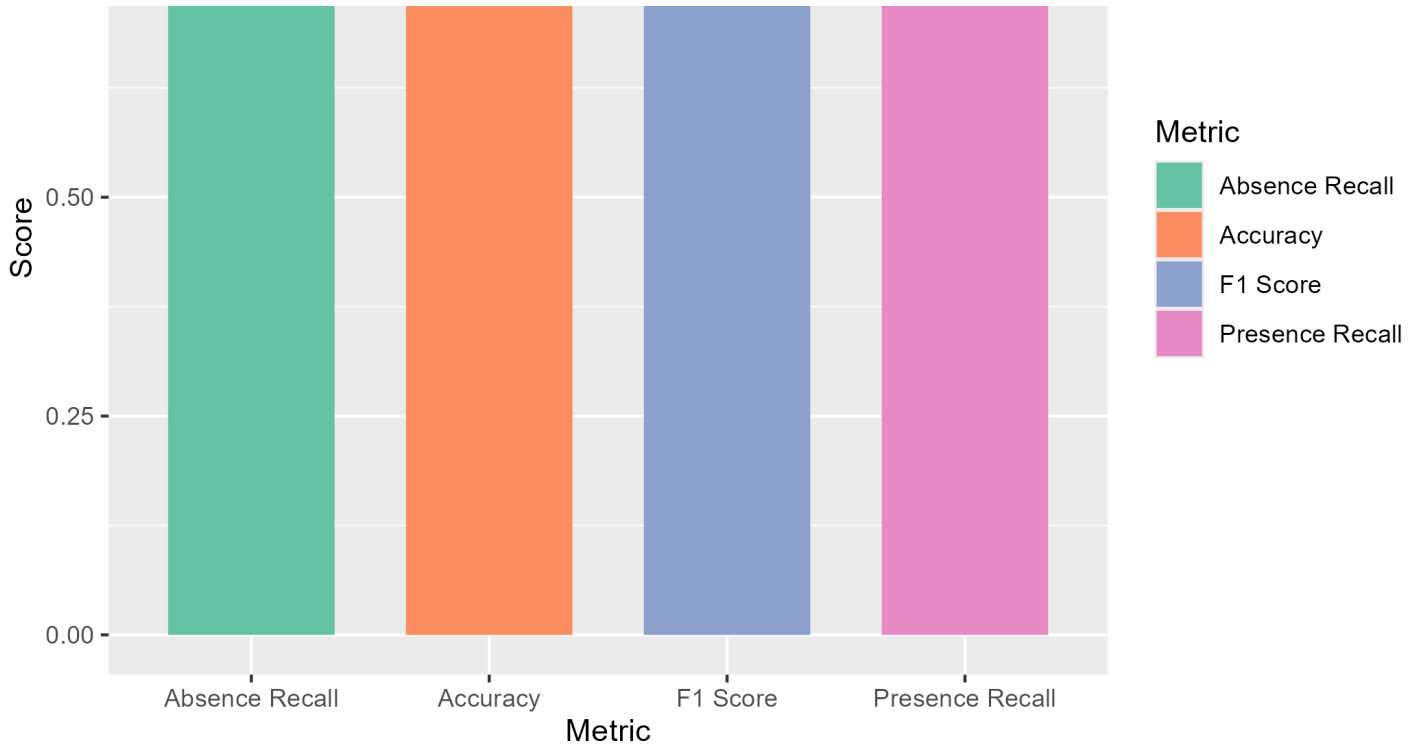
- Models: SVM (Linear) and GLMNet, with ensemble approaches.
- Training: Models trained to classify "Presence" vs "Absence" of fungal endophytes, using only abstracts labeled as "Relevant".
- Feature Engineering: DTM construction and harmonization as above.
- Prediction: Probabilities for "Presence" and "Absence" predicted for all relevant abstracts.
- Thresholding: Multiple thresholds applied (Loose: 0.5, Medium: 0.6, Strict: 0.8, Super Strict: 0.9). Abstracts labeled as "Presence", "Absence", or "Uncertain".
- Ensemble Strategies:
  - Weighted Ensemble: Combines SVM and GLMNet probabilities with configurable weights, prioritizing absence detection.
  - Threshold Optimization: Tests a range of thresholds (0.3–0.7) to maximize balanced recall.
  - Comparison: Performance of individual models and ensembles compared using accuracy, recall, specificity, and F1 score.

- Recommendations: Best approach selected based on research priorities (balanced, conservative, aggressive).

- Outputs:
  - `relevance_pa_preds_all_abstracts.csv` : Relevant abstracts with predicted presence/absence probabilities and labels.
  - Model files: `models/best_model_presence_glmnet_ensemble.rds` , `models/best_model_presence_svmLinear_ensemble.rds`
  - Ensemble prediction functions saved for future use.







3. Summary of Output Files

- Model RDS files: Saved in `models/` for reproducibility and future application.
- Prediction CSVs: Saved in the project root for downstream analysis and manual review.
- Figures: Confusion matrices, accuracy comparisons, and other evaluation plots saved as PNGs.

4. Prediction on Full Dataset

- After model training and evaluation, the best-performing relevance and presence/absence models were applied to the entire set of unlabeled abstracts.
- For relevance, all abstracts were scored and labeled as "Relevant", "Irrelevant", or "Uncertain" using the trained GLMNet model and multiple probability thresholds.
- Abstracts labeled as "Relevant" were then passed to the presence/absence models (SVM, GLMNet, and ensemble functions) to predict the likelihood of target taxa presence or absence, again using multiple thresholds.
- All predictions and probability scores were saved to CSV files for downstream analysis and manual review:
    - `relevance_preds.csv` : Relevance predictions for all abstracts
    - `irrelevant_uncertain_abstracts.csv` : Abstracts predicted as irrelevant or uncertain
    - `relevance_pa_preds_all_abstracts.csv` : Presence/absence predictions for all relevant abstracts
- This workflow enables systematic screening and prioritization of abstracts for further validation and extraction.

6. **Reproducibility:**
    - All random seeds set for reproducibility
    - Full code and session info saved for future reruns

Script: `ML_compare_models_subset.R` contains all steps above, including data loading, preprocessing, model training, evaluation, and export.

Extraction and Analysis

- Species and synonym extraction
- Geographic and temporal analysis
- Scripts: `extract_species_simple.R` , `visualize_taxa_results.R` , `temporal_trend_analysis.R` , `geographic_bias_analysis.R`

Extraction and Analysis Workflow

- Data sources include ML-classified abstracts (weighted ensemble predictions) and validated training data.
- Species detection is performed in batches using parallel processing and reference data from GBIF regarding all accepted species (`species.rds`), with results saved to `results/species_detection_weighted_ensemble.csv` .
- Plant parts are extracted using comprehensive keyword matching for structures and anatomical features, with both detected parts and counts recorded.
- Research methods (molecular, culture-based, microscopy) are identified using curated keyword lists, with each abstract annotated for method presence and summary.
- Geographic information is extracted using extensive keyword lists and regex for countries, continents, regions, and coordinates, with countries categorized as Global North/South.
- All extractions are merged with original metadata and predictions, and final results are saved to `results/comprehensive_extraction_results.csv` .
- Reporting includes extraction rates for species, methods, plant parts, and geography, breakdowns by prediction type and kingdom, and a comprehensive report (`results/comprehensive_extraction_report.txt`) with recommendations for manual review and study prioritization.
- Key outputs: `comprehensive_extraction_results.csv` (all extracted data), `species_detection_weighted_ensemble.csv` (species details), and `comprehensive_extraction_report.txt` (summary and recommendations).

Quality Control

- Manual validation sample (~200 abstracts)
- Script: `manual_validation_sample.R`

2. Main Results

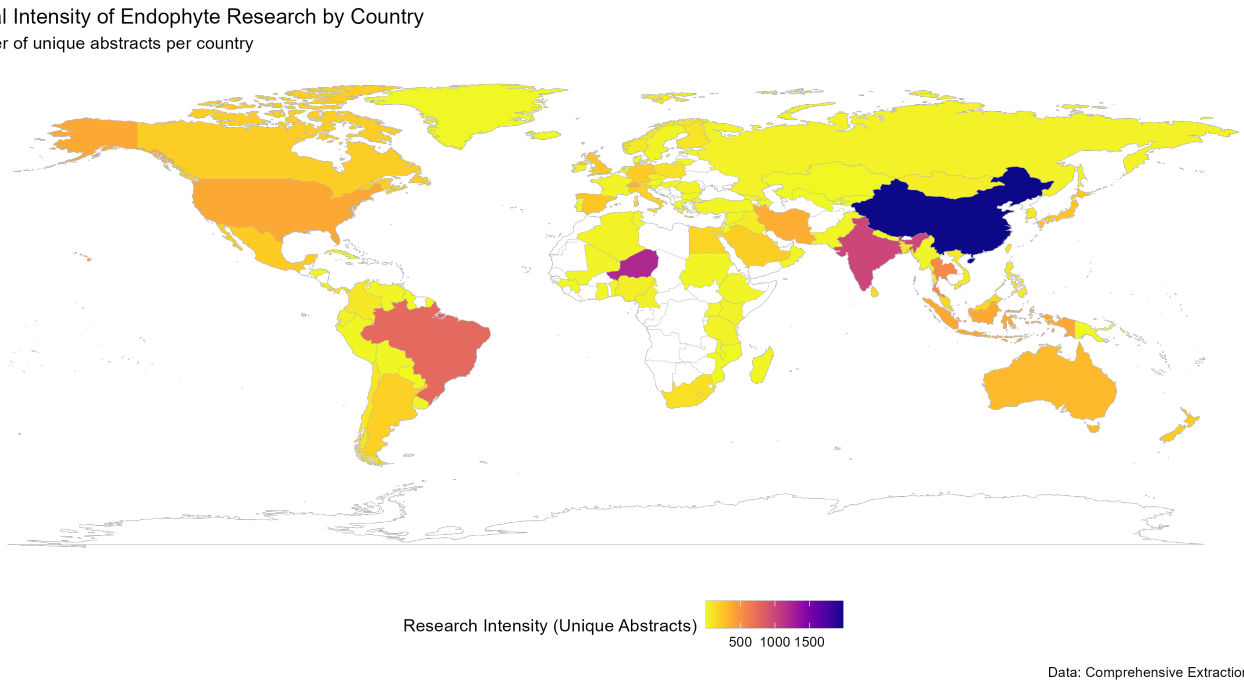
Absence of fungal endophytes?

- Big finding: No examples of plant species that have no evidence of fungal endophytes
- Massive differences in sampling intensity across plant taxa, plant parts, places in the world

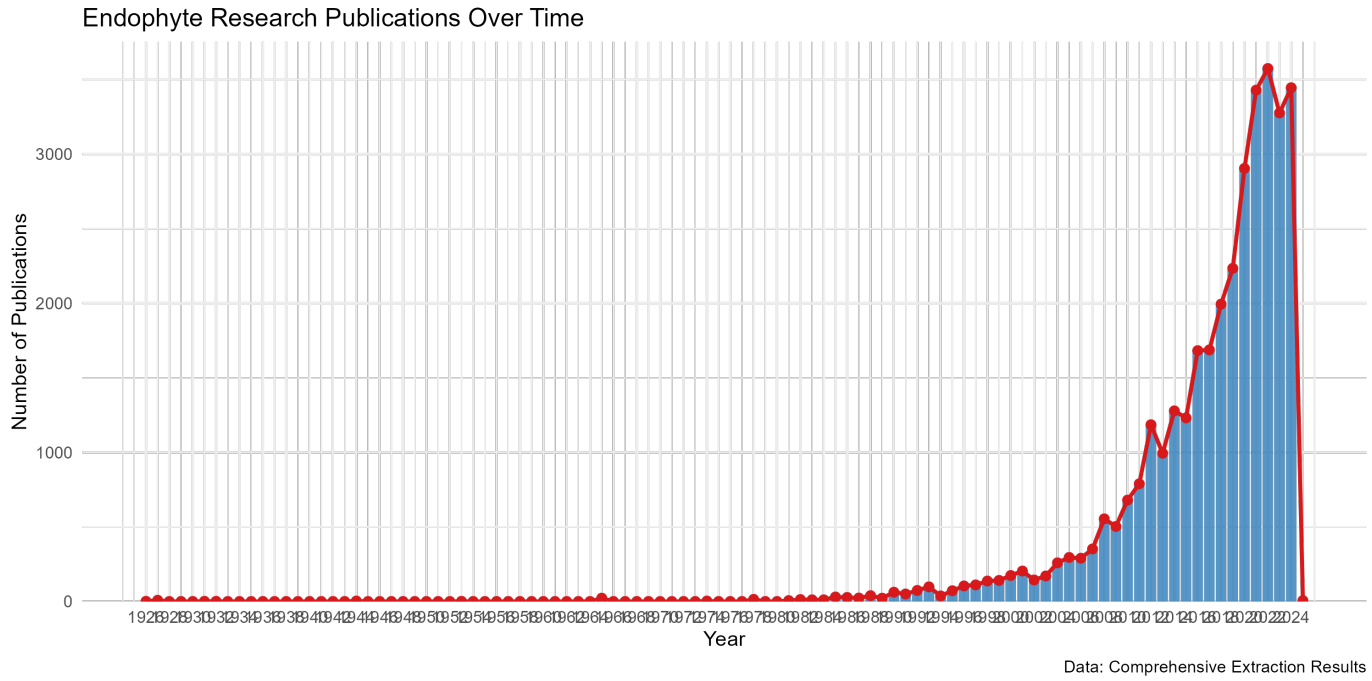
Literature Search and Screening

- Number of abstracts screened: 9778 (will increase with new pull)
- Number included (relevant, loosest labeling): 7851
- Number excluded (irrelevant: reviews, misc other): 1602

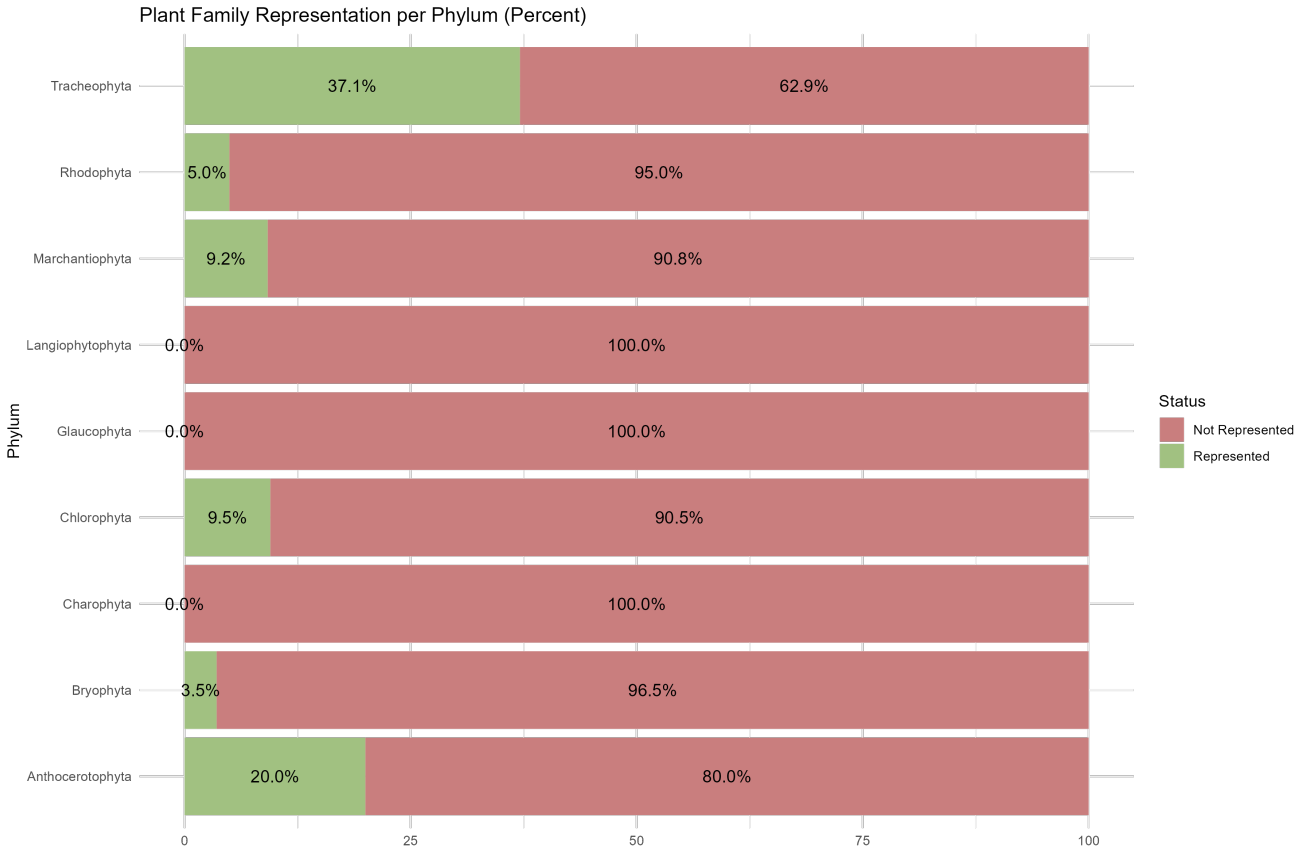
Geographic Distribution

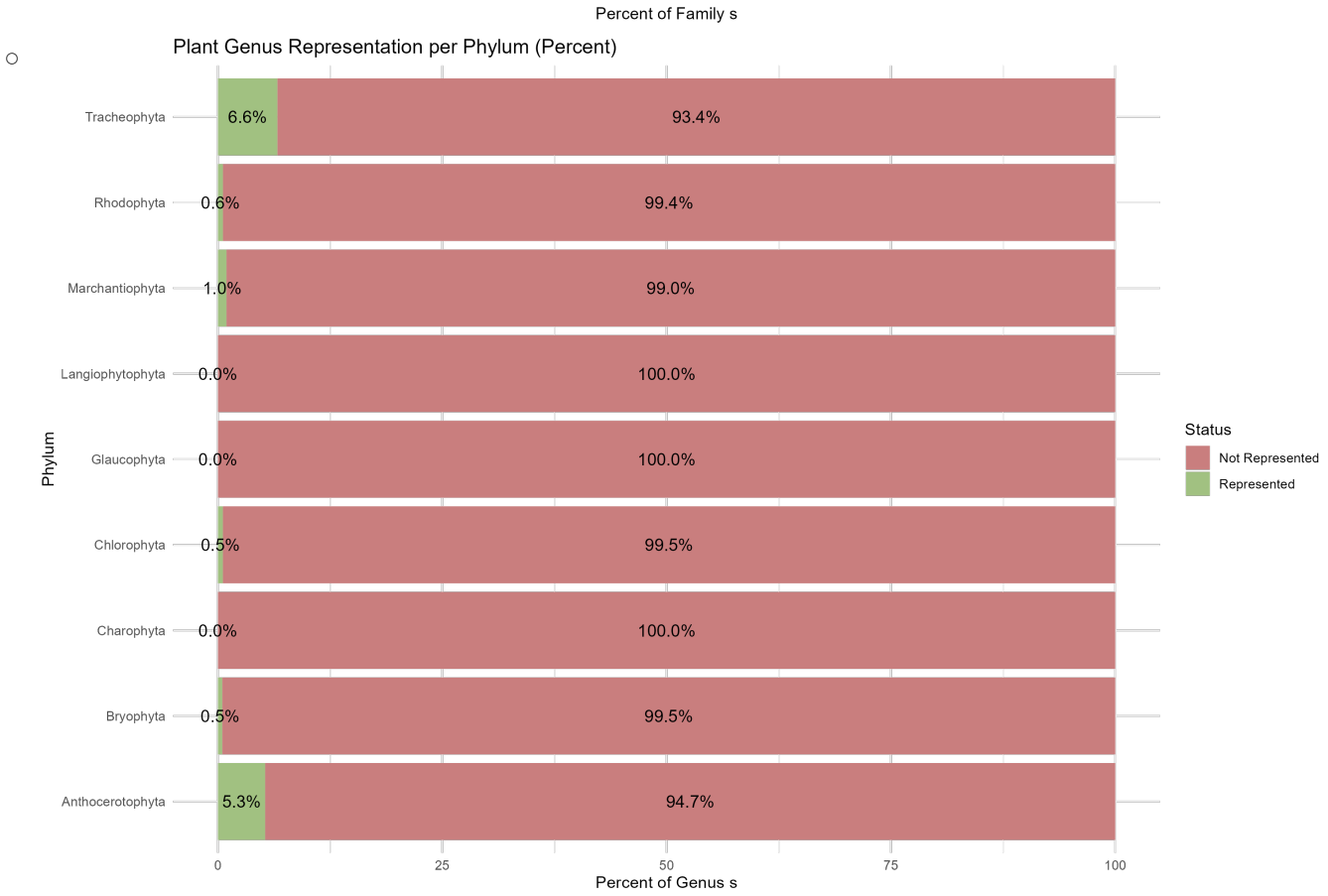
- Map of study locations and research concentration
- 

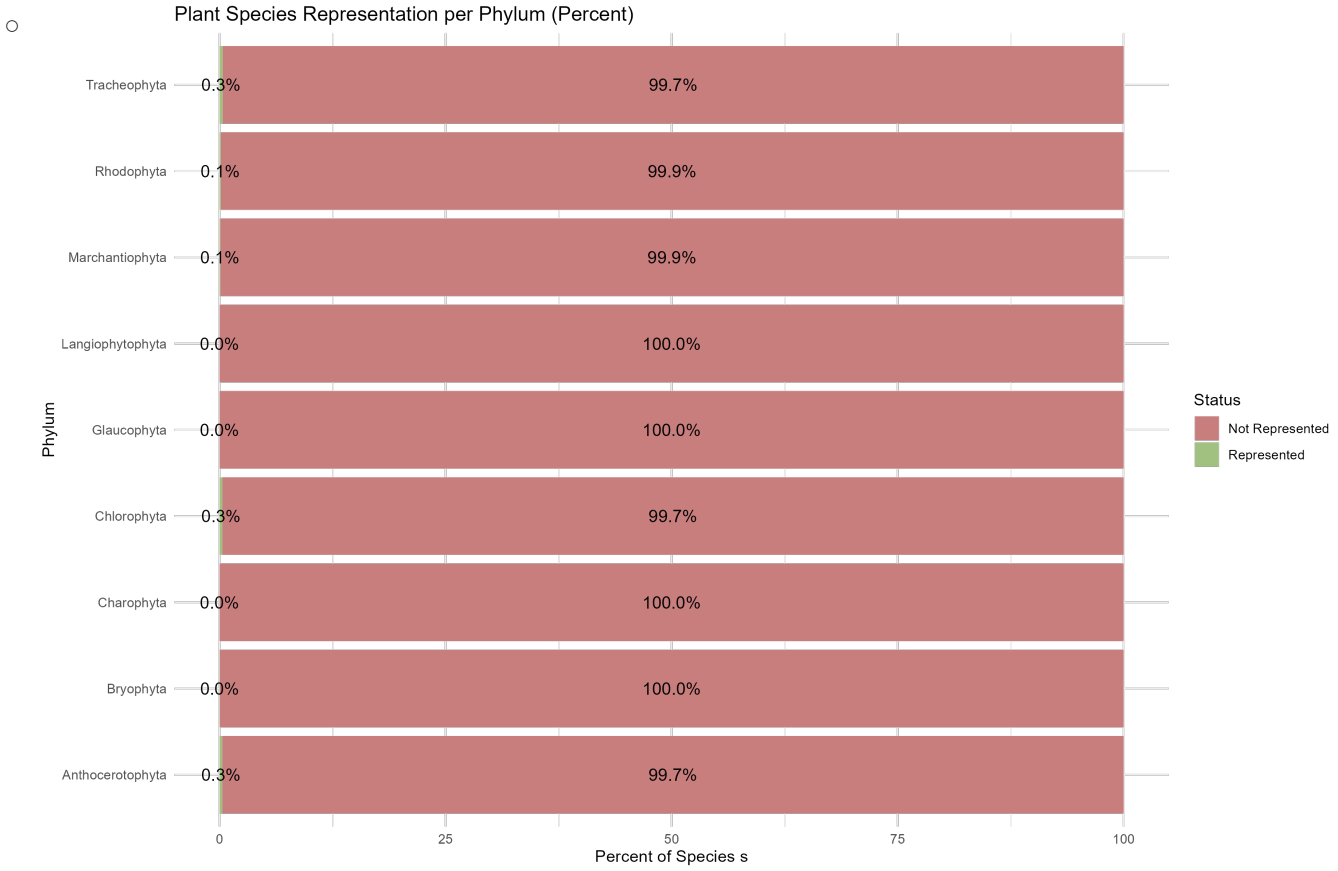
Temporal Trends

- Publication volume and method adoption over time
- 

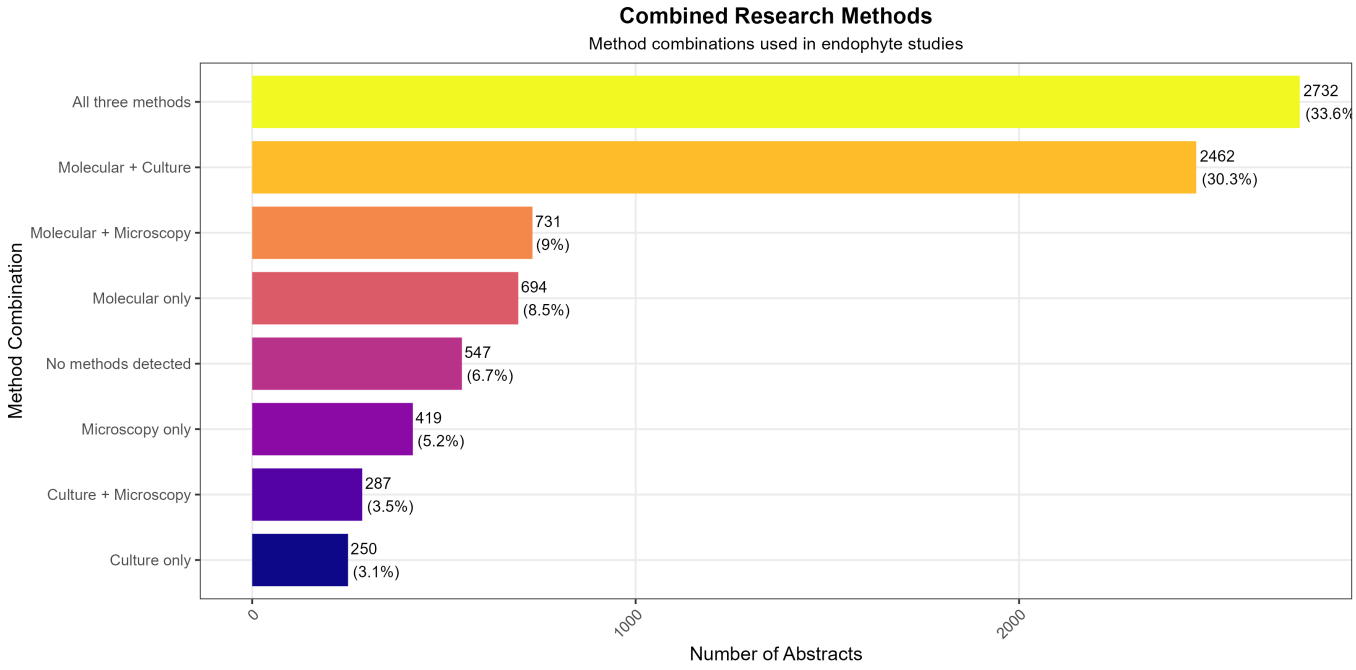
Species and Taxonomic Diversity

- Most studied plant families, genera, and species
- Representation by phylum:
  - 

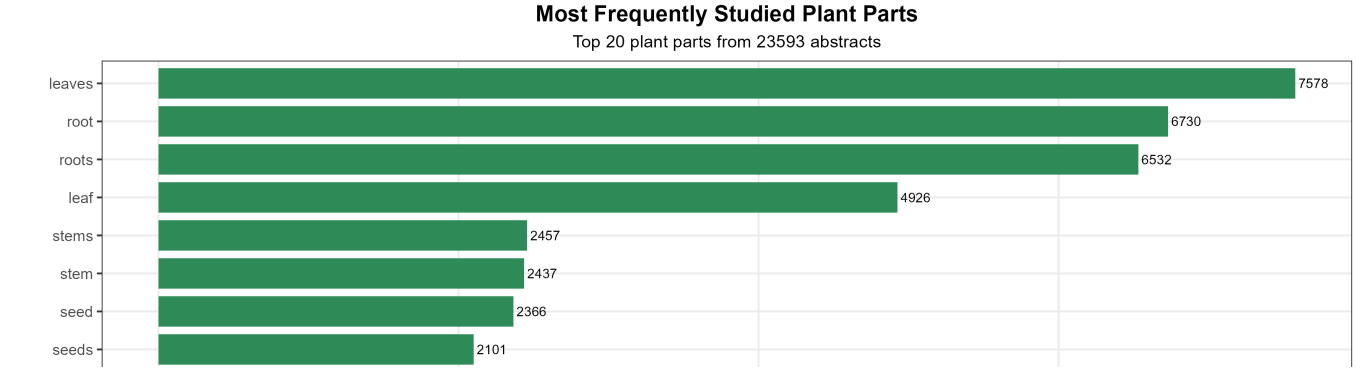


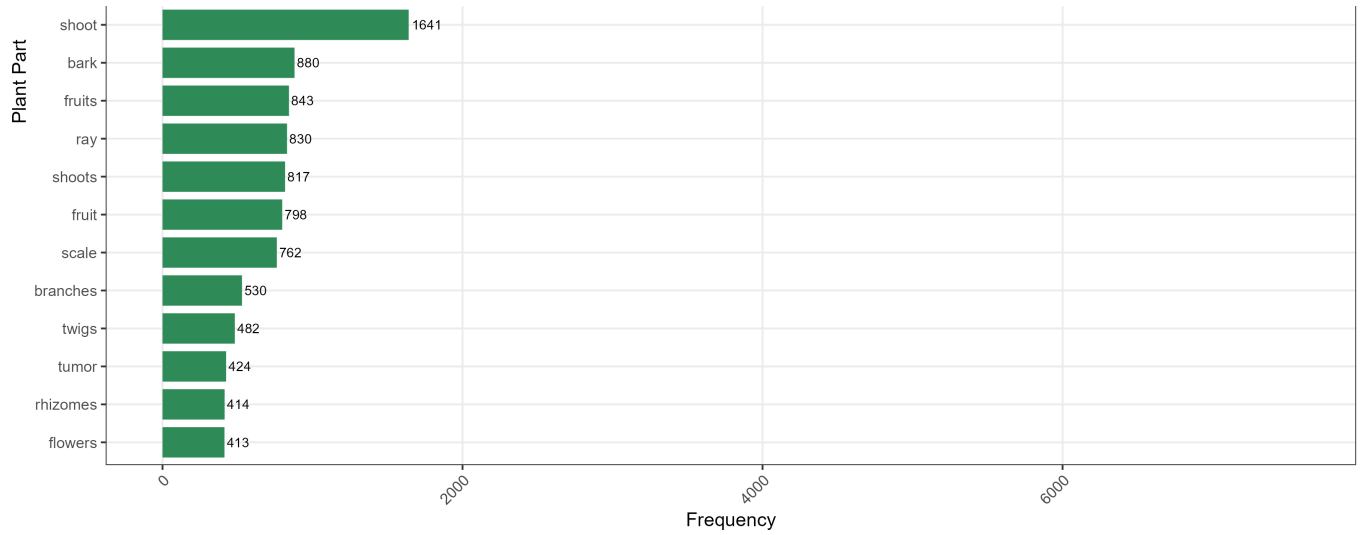


Research Methods

- Frequency of molecular, culture-based, and microscopy methods. **Could use input on making this better**.
- method_categories <- list(
  molecular = c("pcr", "dna", "rna", "sequenc", "primer", "amplif", "gene", "genom",
  "transcript", "clone", "phylogen", "molecular", "extraction", "isolat",
  "genetic", "marker", "polymorphism", "nucleotide", "hybridiz",
  "rrna", "18s", "28s", "rdna", "barcode", "phylogeny"),
  culture_based = c("culture*", "isolat", "plate", "medium", "agar", "petri", "colony",
  "incubat", "sterile", "aseptic", "axenic",
  "ferment", "broth", "in vitro", "cultivation"),
  microscopy = c("microscop", "stain", "section", "histolog", "morpholog", "ultrastructur",
  "sem", "tem", "scanning electron", "transmission electron", "light microscop",
  "confocal", "fluorescen", "magnification", "micrograph", "optical")
)
- 

Plant Parts

- 



Validation and Quality Assessment

- Model accuracy metrics
- Manual validation results
- Bias detection summary

3. Next Steps

- Get approval/input from co-authors on methods and final steps
- Do another pull of the literature using the new search string
- Finalize figures and tables for manuscript
- Complete manual validation and error analysis
- Draft methods and results sections
- Plan for supplementary materials (species lists, code, data dictionary)

Update this document as new results and figures are generated. All figures are embedded for direct review.