Shipley Vetting Checklist

# Stage 1–3 Vetting Checklist (Prof. Bill Shipley)

Date: 2025-10-30  
Scope: Stage 1 (data), Stage 2 (models), Stage 3 (indices)  
Canonical context: results/summaries/phylotraits/

## How to Use

* This checklist mirrors the Stage 1–3 documentation structure so you can review each sub-stage in place.
* For each item, tick boxes, add notes, and link evidence (file paths, plots, small calculations).
* Severity tags: **CRITICAL** (blocker), **IMPORTANT** (investigate).
* Prefer file-backed evidence: CSV snapshots, small DuckDB queries, or saved plots in results/verification/.

Legend for checkboxes: [ ] pending, [x] done, [!] flagged

## Priority Scripts to Review (High Impact)

* Stage 1 — mixgb (High Priority):
  + src/Stage\_1/mixgb/run\_mixgb.R — Production driver (imputation m1–m10, tuned params, GPU); verify hyperparams and fold logic.
  + src/Stage\_1/build\_final\_imputed\_dataset.py — Materializes canonical imputed outputs from mixgb runs; verify column mapping and IDs.
  + src/Stage\_1/verify\_xgboost\_production.py — End-to-end QA for imputed datasets (structure, coverage, drift).
  + src/Stage\_1/mixgb/mixgb\_cv\_eval\_parameterized.R — CV evaluation wrapper; confirm no leakage in CV setup and metrics.
  + src/Stage\_1/verification/verify\_xgboost\_imputation.py — Post-imputation integrity check (completeness, ranges, relationships).
* src/Stage\_2/build\_tier2\_features.py — Assemble per-axis feature tables from Stage 1.10 master, excluding the target-axis EIVE.
* src/Stage\_2/calculate\_tier2\_cv\_phylo.R — Compute context-matched phylogenetic predictors on pruned trees (per axis).
* src/Stage\_2/update\_tier2\_features\_with\_cv\_phylo.py — Merge context-matched phylo into per-axis feature tables.
* src/Stage\_2/build\_tier2\_no\_eive\_features.py — Create NO-EIVE feature tables for imputation (removes cross-axis EIVE and p\_phylo\_\*).
* src/Stage\_2/run\_tier2\_production\_corrected\_all\_axes.sh — Train Tier 2 full models (WITH cross-axis EIVE).
* src/Stage\_2/run\_tier2\_no\_eive\_all\_axes.sh — Train Tier 2 NO-EIVE models (WITHOUT any EIVE-related predictors).
* src/Stage\_2/impute\_eive\_no\_eive\_only.py — Batch imputation for all species needing EIVE (uses NO-EIVE models).

Key SQL joins to review (DuckDB, Stage 1.3):

* tmp\_duke\_eive\_mabberly\_union — Canonical union across Duke/EIVE/Mabberly with WFO keys; check deduplication and key uniqueness.
* tmp\_master\_taxa — Master taxon union over Duke/EIVE/Mabberly/TRY Enhanced/AusTraits; verify source flags and counts.
* stage1\_shortlist\_candidates — Trait-rich shortlist criteria; verify numeric-count logic and joins to AusTraits taxa.
* stage1\_modelling\_shortlist — Full-EIVE + ≥8 numeric TRY traits; verify de-duplication and inclusion thresholds.

## Stage 1 — Data, Normalisation, Imputation Inputs

Reference index: results/summaries/phylotraits/Stage\_1/1.0\_Data\_Pipeline\_Index.md

### 1.0 Data Pipeline Index

* CRITICAL — Verify that every linked Stage 1 summary exists and open without errors.  
  Evidence: Open each path listed under “Stage Catalogue (non-legacy)”.
* IMPORTANT — Cross-check “Verified Artefact Inventory” paths and sizes against filesystem.  
  Evidence: ls -lh of each artefact; note mismatches.
* IMPORTANT — Command references reproduce successfully (dry run or on small subset).  
  Evidence: Rerun a representative command; capture log excerpt.

### 1.1 Raw Data Preparation & WorldFlora Matching

File: results/summaries/phylotraits/Stage\_1/1.1\_Raw\_Data\_Preparation.md

* CRITICAL — WorldFlora match rates are plausible for each source (Duke, EIVE, Mabberly, AusTraits, TRY, GBIF, GloBI).  
  Check: unmatched proportion reasonable; synonyms resolved; no explosion of duplicates after merges.
* CRITICAL — Stable, deterministic merges: no cartesian joins; all joins on intended keys (WFO IDs).  
  Check: unique key constraints hold before/after merge.
* IMPORTANT — Name-processing scripts preserve original names alongside WFO-normalised fields.  
  Check: columns for original and matched names retained.
* IMPORTANT — File counts and row counts match expectations in each sub-workflow (TRY Enhanced, TRY subset, GBIF, GloBI).  
  Evidence: stated vs observed counts; note deviations.

### 1.2 WFO Normalisation & Dataset Merges — Verification

File: results/summaries/phylotraits/Stage\_1/1.2\_WFO\_Normalisation\_Verification.md

* CRITICAL — Parquet conversion checks pass (schema stability; row counts) across all sources.  
  Evidence: section “1. Parquet Conversion Checks”.
* CRITICAL — WorldFlora name-table checks pass (match categories, synonym handling, unique mapping).  
  Evidence: section “2. WorldFlora Name Table Checks”.
* CRITICAL — Enriched merge checks pass (no row loss, expected new columns present, key uniqueness).  
  Evidence: section “3. Enriched Merge Checks”.
* IMPORTANT — GloBI subset and full datasets show consistent enrichment (Plants-only vs full).  
  Evidence: sections “GloBI Interactions — Plants Subset” and “— Full Dataset”.
* IMPORTANT — TRY Enhanced/Selected-traits and Mabberly checks pass with stable taxon coverage.  
  Evidence: corresponding sections.

### 1.3 Dataset Construction

File: results/summaries/phylotraits/Stage\_1/1.3\_Dataset\_Construction.md

* CRITICAL — Master union build scripts produce the documented row/column counts.  
  Evidence: compare to “Coverage Summary”.
* CRITICAL — Shortlist construction (Trait-Rich; Modelling) meets inclusion rules; no leakage from targets into predictors.  
  Evidence: inspect feature lists; confirm exclusion of target EIVE from predictors where required.
* IMPORTANT — GBIF coverage figures (occurrence coverage subsections) match outputs.  
  Evidence: quick DuckDB queries on occurrence tables.

### 1.4 Shortlisting Verification

File: results/summaries/phylotraits/Stage\_1/1.4\_Shortlisting\_Verification.md

* CRITICAL — Verification steps for union, master union, and modelling shortlist match “Expected Results”.  
  Evidence: reproduce counts and key constraints.
* IMPORTANT — GBIF occurrence coverage checks for shortlisted sets match expectations.  
  Evidence: section “GBIF Occurrence Coverage Checks”.

### 1.5 Environmental Sampling Workflows

File: results/summaries/phylotraits/Stage\_1/1.5\_Environmental\_Sampling\_Workflows.md

* CRITICAL — Aggregations (mean, sd, min, max) and quantiles (q05/q50/q95/IQR) are internally consistent.  
  Check: quantile ordering; IQR ≥ 0; min ≤ q05 ≤ q50 ≤ q95 ≤ max.
* IMPORTANT — Sampling logs exist; no tile/cell read errors; retries handled.  
  Evidence: run logs referenced under “Run Log & Archival”.

### 1.6 Environmental Verification

File: results/summaries/phylotraits/Stage\_1/1.6\_Environmental\_Verification.md

* CRITICAL — Baseline file checks pass (schema, row counts, presence of species keys).  
  Evidence: “1.1 Baseline File Checks”.
* CRITICAL — Null fraction guardrails respected across datasets.  
  Evidence: “Null Fraction Thresholds” vs observed nulls.
* CRITICAL — WorldClim, SoilGrids, Agroclim value ranges within documented intervals.  
  Evidence: “WorldClim/SoilGrids/Agroclim Variable Ranges”.
* IMPORTANT — Quantile ordering constraint holds globally; aggregation tolerance within bounds.  
  Evidence: “Quantile Ordering Constraint”, “Aggregation Tolerance”.

### 1.7a Imputation Dataset Preparation

File: results/summaries/phylotraits/Stage\_1/1.7a\_Imputation\_Dataset\_Preparation.md

* CRITICAL — Column definitions per permutation (Perm1/2/3) match documented specs; no target leakage.  
  Evidence: “3.1–3.3” definitions; anti-leakage rules.
* Evidence: section “2. Canonical Sources”.

### 1.7b XGBoost Anti-Leakage Experiments

Removed from vetting: preparatory experiments not part of the production pipeline.

### 1.7c BHPMF Gap-Filling and Imputation

Removed from vetting: alternative method not used in production.

### 1.7d XGBoost Production Imputation

File: results/summaries/phylotraits/Stage\_1/1.7d\_XGBoost\_Production\_Imputation.md

* CRITICAL — 10-fold CV results match documented metrics; sanity checks vs earlier datasets pass.  
  Evidence: “2. Cross-Validation Results”.
* CRITICAL — Production imputation outputs are complete (no missing values), dimensions exact.  
  Evidence: “3.2 Data Completeness Verification”.
* IMPORTANT — Coverage summaries are sensible across taxa; outliers documented.  
  Evidence: “3.3 Imputation Coverage Summary”.

### 1.7e XGBoost Verification Pipeline

File: results/summaries/phylotraits/Stage\_1/1.7e\_XGBoost\_Verification\_Pipeline.md

* CRITICAL — Pre-imputation dataset checks pass (column structure, feature composition, EIVE exclusion, log transforms, types).  
  Evidence: “1. Dataset Construction Validation”.
* CRITICAL — Post-imputation biological plausibility checks pass. Evidence: “2.2 Trait Relationship Validation”.
* CRITICAL — Completeness: ZERO missing in imputed datasets; log consistency holds.  
  Evidence: “2.3–2.4”.
* CRITICAL — Data leakage audit clears (CV standardisation uses training-only stats).  
  Evidence: “2. Data Leakage Prevention Audit”.

### 1.8 iNaturalist Media QA

Removed from vetting: non-essential for production EIVE/CSR outputs.

### 1.9 Phylogenetic Predictors

File: results/summaries/phylotraits/Stage\_1/1.9\_Phylogenetic\_Predictors.md

* CRITICAL — Formula and parameters match documented  
  Evidence: “Formula (Shipley)” and “Parameters”.
* CRITICAL — Coverage meets stated proportions (Tier 1: ~99.2%).  
  Evidence: “1.3 Output — Coverage”.
* CRITICAL — Tier 2 context-matched predictors calculated per axis over pruned trees (no context leakage).  
  Evidence: “2.1 Why Context-Matched?”
* IMPORTANT — Merges into feature tables verified; no row loss.

### 1.10 Modelling Master Tables

File: results/summaries/phylotraits/Stage\_1/1.10\_Modelling\_Master\_Tables.md

* CRITICAL — Feature inventory count (741) exact for Tier 1 and Tier 2.  
  Evidence: “1. Feature Inventory”.
* CRITICAL — Tier 1 (1,084) and Tier 2 (11,680) dimensions exact; species coverage matches.  
  Evidence: “2.2 Output”, “3.2 Output”.
* IMPORTANT — Inclusion of full environmental quantiles is justified and consistent with Stage 2 usage.  
  Evidence: “Environmental Quantiles Rationale”.
* IMPORTANT — No-EIVE variants documented for Stage 2 imputation workflow.  
  Evidence: “3.3 Usage in Stage 2”.

### Annex A — TRY ↔ AusTraits Mapping

Removed from vetting: optional mapping review.

## Stage 2 — Modelling, Validation, SHAP, Imputation

Reference overview: results/summaries/phylotraits/Stage\_2/2.0\_Modelling\_Overview.md

### Priority Focus — Ecological Plausibility (SHAP + Domain)

* CRITICAL — Per-axis top SHAP predictors are ecologically consistent with the target axis (e.g., climate drivers for T, soil/pH proxies for R, moisture signals for M) and known trait–environment relationships.  
  Evidence: Axis reports’ “Top SHAP Predictors” and model directories’ xgb\_{axis}\_shap\_importance.csv.
* CRITICAL — Category-level SHAP contributions (EIVE/env/phylo/traits/categorical) are proportionate and interpretable; no implausible dominance by a single proxy category.  
  Evidence: results/summaries/phylotraits/Stage\_2/2.8\_SHAP\_Category\_Analysis.md.
* IMPORTANT — Flag any ecological inconsistencies (e.g., inverse signs against theory, habitat-incongruent features) with [!] and suggest constraints or feature reviews.

### 2.0 Modelling Overview

* CRITICAL — Context-matched phylogenetic predictors confirmed and used for Tier 2 CV.  
  Evidence: section “CRITICAL: Context-Matched Phylo Predictors Required”.
* CRITICAL — Inputs from Stage 1.10 (Tier 1 and Tier 2) match exact feature inventory.  
  Evidence: “2. Input Datasets”.
* IMPORTANT — Two-tier workflow and two-model hybrid (Full vs No-EIVE) are consistently applied across axes.  
  Evidence: “3.2–3.4”.

### 2.1–2.5 Axis Reports (L, T, M, N, R)

Files:

* results/summaries/phylotraits/Stage\_2/2.1\_L\_Axis\_XGBoost.md
* results/summaries/phylotraits/Stage\_2/2.2\_T\_Axis\_XGBoost.md
* results/summaries/phylotraits/Stage\_2/2.3\_M\_Axis\_XGBoost.md
* results/summaries/phylotraits/Stage\_2/2.4\_N\_Axis\_XGBoost.md
* results/summaries/phylotraits/Stage\_2/2.5\_R\_Axis\_XGBoost.md
* IMPORTANT — Tier 2 Full vs No-EIVE performance consistent with overview; differences ecologically interpretable.  
  Evidence: “Tier 2 Production CV”.
* IMPORTANT — Top SHAP predictors per axis biologically plausible Evidence: “Top SHAP Predictors”.

### 2.6 CLM vs XGBoost Comparison

File: results/summaries/phylotraits/Stage\_2/2.6\_CLM\_vs\_XGBoost\_Production.md

* IMPORTANT — Fidelity of replication based on (Shipley, 2017) Evidence: “Performance Comparison”.

### 2.7 Production and Imputation

File: results/summaries/phylotraits/Stage\_2/2.7\_Production\_and\_Imputation.md

* CRITICAL — Feature tables built correctly; context-matched phylo merged; No-EIVE tables strip cross-axis EIVE.  
  Evidence: “3. Feature Table Build”.
* CRITICAL — Trained model artefacts present; CV metrics recorded and plausible.  
  Evidence: “4. Model Training”.
* CRITICAL — Imputation outputs complete; no missing values; metadata and source breakdown consistent.  
  Evidence: “5.6 Output Files and Verification”.

### 2.8 SHAP Category Analysis

File: results/summaries/phylotraits/Stage\_2/2.8\_SHAP\_Category\_Analysis.md

* IMPORTANT — Category-level SHAP distributions match ecological intuition (e.g., EIVE vs env vs phylo vs trait contributions).  
  Evidence: per-axis category importance sections.

### Annex A — Experimental Joint Imputation (mixgb)

Removed from vetting: experimental branch not used in production. However, this is of interest as an alternative ONE stage method to impute EIVE and traits simultaneously. Metrics lag that of canonical two-stage run.

## Stage 3 — CSR Indices and Ecosystem Services

Index and runbook: results/summaries/phylotraits/Stage\_3/3.0\_Stage3\_Overview.md

### 3.0 Overview & Runbook

* CRITICAL — Single-command execution reproduces final outputs; file manifest matches locations.  
  Evidence: “Single Command Execution”, “Data Outputs”.
* IMPORTANT — Key validation results (NPP stratification, N-fixation, CSR patterns, data quality) reproduced or spot-checked.  
  Evidence: “Key Validation Results”.

### 3.1 Nitrogen Fixation Methodology

File: results/summaries/phylotraits/Stage\_3/3.1\_Nitrogen\_Fixation\_Methodology.md

* CRITICAL — Extraction from TRY TraitID 8 yields stated coverage and rating distributions.  
  Evidence: “Results” and “Verification” sections.
* IMPORTANT — Fabaceae comparison and actinorhizal detections match expectations; false positives/negatives noted.  
  Evidence: “4.1–4.2”.

### 3.2 CSR Pipeline & Verification

File: results/summaries/phylotraits/Stage\_3/3.2\_CSR\_and\_Verification.md

* CRITICAL — Formula correctness verified (Pierce et al. 2016 StrateFy).  
  Evidence: “✓ Formula Correctness”.
* CRITICAL — Life-form stratification applied per guidance; NaN handling for edge cases documented.  
  Evidence: “✓ Life Form Stratification”, “Edge Case Analysis”.
* IMPORTANT — Patterns match Shipley Part I; Part II enhancements implemented.  
  Evidence: “✓ CSR Patterns Match…”, “✓ Shipley Part II Enhancements”.

## Cross-Cutting Biological/Statistical Red Flags (Apply Anywhere)

* CRITICAL — Any evidence of target leakage between predictors and targets (any stage).  
  Action: isolate and document the channel; propose fix.
* CRITICAL — Unit inconsistencies (e.g., LMA/SLA transforms) or log/linear mismatches.  
  Action: confirm transforms; re-derive a few examples.
* IMPORTANT — Over-dominance of single feature category in SHAP or GAIN (in imputation) without biological basis.  
  Action: check for duplicated proxies; collinearity artefacts.
* IMPORTANT — Phylogenetic predictors computed on wrong context (full vs pruned tree) for CV training.  
  Action: verify per-axis pruning.

## Reviewer Notes and Additional Checks

Use these placeholders to add expert-driven checks beyond automation.

### New Check 1

* Description:  
  Evidence:  
  Notes:

### New Check 2

* Description:  
  Evidence:  
  Notes:

### New Check 3

* Description:  
  Evidence:  
  Notes: