LifeDive: Cross-kingdom phenotype estimates for 35,856 species

Tyler M. Moore

Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania; Lifespan Brain Institute, Penn Medicine and Children's Hospital of Philadelphia

Permanent preprint · 10 November 2026
tymoore@pennmedicine.upenn.edu ORCID 0000-0002-1384-0151
DOI 10.17605/OSF.IO/AQKHN

Recommended citation: Moore, T. M. (2026). *LifeDive: Cross-kingdom phenotype estimates for 35,856 species.* Permanent preprint. https://doi.org/10.17605/OSF.IO/AQKHN

Abstract

Dense, standardized species-by-trait matrices support comparative biology and ecology, but their coverage remains uneven. LifeDive uses a general-purpose, web-search-enabled language model within a scripted annotation pipeline to estimate traits across animals, plants, and fungi. The workflow assembled 35,856 species by 196 phenotype questions, yielding 7,027,776 entries in a completed matrix; 16.6% of entries were imputed from other machine-derived annotations. Four ordered substantive responses express model judgments about trait statements, with a separate code for unknown or inapplicable cases. An exploratory 14-factor representation summarized broad biological contrasts, although statistical factor-retention criteria suggested substantially more dimensions. A benchmark screened 168 external criteria, with selected results reported for 44; many biologically direct associations were strong, particularly for taxonomy and selected ecological traits. A separate model-assisted audit reported 96.5% directional agreement in a selected taxonomy-rule subset and approximately 237 agreements among 240 selected species-specific checks. These audit subsets do not estimate accuracy across the whole matrix, and the AmphiBIO-derived benchmark results require qualification because unrecorded categories were treated as negative indicators. Trait profiles also produced recognizable taxonomic and functional groupings. LifeDive offers a broad computational phenotype resource for exploration, hypothesis generation, and prioritizing expert review. Individual values remain estimates, and downstream inference requires attention to imputation, sampling, source dependence, and trait-specific verification. The reported application programming interface and computing expenditure was approximately US$10,000.

Keywords: comparative biology; computational phenotype; cross-kingdom traits; large language models; trait imputation; validation

Contents

Main text and supplementary information are integrated in this single document. Contents entries and PDF bookmarks provide direct navigation.

1 Introduction . . . 3

2 Methods . . . 4

3 Results . . . 8

4 Discussion . . . 16

5 Conclusion . . . 18

Transparency and availability . . . 18

References . . . 19

S1 Supplementary methods . . . 22

S2 Supplementary results . . . 26

S3 Model-assisted audit: methods and reported results . . . 32

Table S1. Species in the 200-species subset . . . 35

Table S2. Retained questions and displayed pattern coefficients . . . 41

Table S3. Selected external-criterion associations . . . 51

Reading the evidence. The main text distinguishes the completed data product from original annotations. External correlations are selected associations; the model-assisted audit is a qualified exploratory assessment. The supplementary tables preserve reported outputs and identify unavailable or coding-sensitive entries.

1 Introduction

Phenotypes connect evolutionary history to ecological function. Morphology, physiology, behavior, life history, habitat use, trophic strategy, stress tolerance and reproductive mode determine how organisms interact with one another and with their environments, and they supply the raw material for comparative biology, macroecology, conservation prioritization and functional biodiversity science. These fields increasingly depend on large, standardized matrices of organismal characteristics: whether a species is woody, venomous, photosynthetic, aquatic, parasitic, drought-tolerant, long-lived, edible, human-associated or sensitive to pollution. Yet the underlying data remain uneven. Landmark trait resources have transformed plant ecology, ornithology and mammalian macroecology, but dense species-by-trait matrices are hard to assemble: expert curation is slow, taxonomic coverage is biased, and missingness is structured by clade, geography, body size, charisma and research effort (Domazetoski et al., 2025; Kattge et al., 2020; Marcos et al., 2025).

Existing databases show both the value and the constraint. PanTHERIA, EltonTraits, COMBINE and AVONET made broad comparative analyses possible by harmonizing life-history, ecological, morphological and geographic attributes across mammals or birds while carefully documenting taxonomy, synonymy, missingness and provenance (Jones et al., 2009; Soria et al., 2021; Tobias et al., 2022; Wilman et al., 2014). Their success also exposes a bottleneck: each new trait domain or clade typically requires renewed expert reading, taxonomic reconciliation and quality control. Cross-kingdom work is especially demanding, because a useful matrix must place traits as different as cell walls, venom, gills, lignin degradation, seed dispersal and endosymbiosis onto a single response framework. In such settings, imputation and uncertainty are not peripheral technicalities; they become part of the resource itself.

This constraint binds most tightly at the scales at which macroecology and biogeography operate. Global analyses of functional diversity, trait-environment relationships, community assembly and extinction risk require trait values for every species in an assemblage, yet the assemblages that matter are rarely confined to one well-curated clade. A forest plot contains trees, fungi, arthropods and vertebrates; a lake contains macrophytes, invertebrates and fishes. Analysts therefore either restrict their questions to the clades for which matrices exist, or impute across gaps whose structure is itself biased by the same taxonomic and geographic factors that produced the gaps. A resource that applies one response framework across kingdoms addresses a different problem from the one solved by any single expert database: it trades trait-specific precision for comparability, so that the same question can be put to a tree, a bird, a fungus, a mollusc and an insect.

Automated text mining can accelerate trait mobilization by converting floras, monographs, web pages, and articles into structured records (Endara et al., 2018; Farrell et al., 2024; Folk et al., 2024). Recent large language model (LLM) systems extract ecological records, plant traits, and species interactions from unstructured text (Castro et al., 2024; Gougherty & Clipp, 2024; Keck et al., 2026). Extraction remains constrained by source access, wording, and controlled vocabularies. A complementary approach is to ask a model to estimate a trait even when a particular source does not state it explicitly; doing so broadens coverage but changes the evidentiary status of the resulting value from an extracted observation to an inference.

Pretrained language models encode factual and relational information, but they can also produce fluent errors and are sensitive to prompt wording, training distribution, and retrieval context

(Petroni et al., 2019; Wysocka et al., 2024). Work on model-generated quantitative data and its possible uses motivates examination of both individual values and multivariate structure (Moore, 2025; Varnum et al., 2024). Studies of scientific extraction emphasize explicit schemas, constrained outputs, parsers, missing-data rules, and comparison with trusted sources (Dagdelen et al., 2024; Lai et al., 2025; Mammides et al., 2026; Polak & Morgan, 2024). A computational trait resource therefore needs separate assessments of recognizable structure, external agreement, and the reliability of individual entries.

LifeDive applies one question set across kingdoms and uses a scripted batch workflow to make large-scale annotation feasible for an individual investigator. The reported API and computing expenditure was approximately US$10,000. This expenditure describes the present workflow; no matched comparison of labor, cost, or accuracy with expert curation was conducted.

This paper has two aims: to document the annotation and processing workflow, and to assess whether its output preserves recognizable biological structure. The evidence consists of factor analysis, selected external benchmark associations, a model-assisted audit of subsets of cells from 200 randomly selected species, and a phenotype-similarity network. These complementary analyses assess internal structure, external agreement, and agreement for selected cells. Detailed methods, additional results, the audit prompt and an edited account of its findings, and supplementary tables are included in this PDF.

2 Methods

2.1 Study design and species selection

Species candidates came from the Catalogue of Life checklist, version 2026-04-18 XR (Catalogue of Life, 2026). The compilation retained accepted or provisionally accepted records at species rank, with available higher taxonomy and English vernacular names. Sampling focused on Animalia, Plantae, and Fungi and included approximately 5% extinct species. A family-first algorithm aimed to represent every eligible family, then filled the remaining quota from species not already chosen. Priorities favored fully accepted names, English common names, and Animalia and Plantae over Fungi; a small random component broke ties under a fixed seed. This procedure returned 35,000 species. GPT 5.5 Pro was then asked to identify familiar organisms missing from the list, including examples such as the barn owl (*Tyto alba*) and Brazilian wandering spider (*Phoneutria nigriventer*), producing the final sample of 35,856. The sample is designed for breadth and familiarity, rather than as a probability sample of global biodiversity.

2.2 Phenotype questions and response coding

The final analysis retained 196 questions spanning cellular biology, anatomy and structure, physiology, behavior, reproduction, life history, ecology, environmental tolerance, and human salience. Table 1 gives examples; Table S2 preserves the complete retained question wording. Responses used four ordered substantive categories: 0, Definitely No; 1, Probably No; 3, Probably Yes; and 4, Definitely Yes. Code 2 denoted Unknown or not applicable and was not treated as an ordinal midpoint in analysis. This response format expresses the model's endorsement of a proposition, rather than a common physical quantity or a measured amount of a trait. The same questions can be applied across kingdoms, but some are ambiguous or biologically inapplicable for particular taxa.

Table 1. Representative phenotype questions. Eleven of the 196 retained questions. Original question wording is preserved.

Trait domain	Example phenotype question
Cellular biology	Does this organism produce cellulose?
Cellular biology	Can this organism photosynthesize?
Anatomy and structure	Does this organism use chitin as a structural material?
Anatomy and structure	Does this organism have tentacles or tentacle-like appendages?
Physiology	Can this organism transmit electrical signals between its cells?
Behavior	Can this organism produce complex vocalizations or sounds?
Reproduction	Can this organism change its sex or mating type during its lifetime?
Life history	Can this organism live longer than 100 years?
Ecology	Does this organism depend on a close symbiotic relationship with another species to survive?
Environmental tolerance	Can this organism survive at 50 °C for 100 hours or more?
Human salience	Can this organism pose a serious threat to human life?

Substantive responses: 0 = Definitely No; 1 = Probably No; 3 = Probably Yes; 4 = Definitely Yes. Code 2 = Unknown or not applicable. The complete retained questions appear in Table S2.

2.3 Language-model annotation and quality control

Phenotype annotations were collected with the xAI Grok Batch application programming interface (xAI, 2026), using the scripted model grok-4-1-fast-reasoning with web search enabled so that the model could retrieve information on obscure taxa. Batch processing was chosen because the task consisted of many independent species-by-question inference requests. Species were processed in groups of three; each prompt listed the species and questions, defined the response scale and required a complete comma-separated row for every species, with every cell populated and the Unknown code used where an answer could not be determined.

The scripted workflow validated requests before upload, saved and polled batch identifiers, retrieved completed results with pagination, and checkpointed parsed data. The parser removed code fences, aligned question columns by identifier, padded or trimmed responses with too few or too many species rows, and converted nonnumeric, missing, or unreliably aligned cells to the Unknown code. Requests with extraction errors, missing response text, or unreliable column alignment were logged for retry. These checks address format and alignment; they do not establish biological correctness. Row padding or trimming also requires reliable identity matching. The results are computational phenotype estimates, not direct measurements or expert-curated observations.

2.4 Data assembly and imputation

Annotation was conducted in 11 question subsets and merged by species. Unknown/not applicable and out-of-range values were treated as missing. Columns were retained when they had fewer than 25,000 missing values, leaving 196 analyzed questions. This permissive item-level threshold is distinct from the overall missing-cell rate of 16.6%. Missing responses were imputed item by item with multinomial random forests (Breiman, 2001) in H2O version 3.46.0.9 (H2O.ai, 2026). The model for each item used its observed values as training outcomes and the other retained items as predictors, with a fixed seed and randomized item order. Predicted classes were inserted only into missing target cells.

The completed matrix was used for the latent-structure analyses. Original unimputed responses remain the appropriate starting point for interpreting individual annotations. Imputation predicts a missing response category from other machine-generated categories; it does not create a new biological observation or distinguish an unknown trait from an inapplicable question. The reported analyses do not separately establish imputation accuracy or propagate imputation uncertainty.

2.5 Factor analysis and score construction

After imputation, the four substantive categories were recoded as 0, 1, 2, and 3, corresponding to original codes 0, 1, 3, and 4. Exploratory factor analysis of the polychoric correlation matrix was performed with *psych::irt.fa*, which also expresses resulting parameters in item-response-theory terms (Revelle, 2026). The reported environment used R version 4.6.0 and psych version 2.6.5. Promax rotation allowed correlated dimensions (Fabrigar et al., 1999; Hendrickson & White, 1964). Fourteen factors were chosen as an interpretable descriptive reduction. Parallel analysis suggested 58 factors and the minimum-average-partial criterion suggested 41 (Horn, 1965; Velicer, 1976); the 14-factor solution is therefore not presented as a uniquely supported estimate of dimensionality.

Unit-weighted sum scores based on the loading patterns were used instead of theta scores because a small number of highly discriminating items otherwise dominated score levels. The Parasitic and Cold-tolerant score directions were oriented so that higher scores indicated more of the named profile. Figure 1 uses scores winsorized with a trim of 0.0005 and standardized, with small uniform random jitter for plotting only. The plotting workflow used ggplot2 version 4.0.3, ggrepel version 0.9.8, and patchwork version 1.3.2 (Pedersen, 2025; Slowikowski, 2026; Wickham, 2016). The rounded, selectively displayed pattern coefficients in Table S2 do not alone specify the complete score-construction algorithm; the analysis code is needed for exact reproduction.

2.6 External benchmark comparisons

The benchmark contained 168 numeric criteria from public trait resources and a standardized taxonomic backbone. Sources included bird and mammal traits from EltonTraits 1.0 (Wilman et al., 2014), mammal traits from PanTHERIA (Jones et al., 2009) and MammalDIET (Kissling et al., 2014), amphibian traits from AmphiBIO (Oliveira et al., 2017), marine-invertebrate traits from SeaLifeBase (Palomares & Pauly, 2024), fish traits from FishBase (Froese & Pauly, 2024), and taxonomy from the traitdata package and GBIF Backbone Taxonomy (GBIF Secretariat, 2023; RS-eco, 2023). The supplementary workflow also identifies the Amniote Life-History Database and AnAge among overlapping trait sources (Myhrvold et al., 2015; Tacutu et al., 2018).

Species were matched by genus-species binomial after normalizing whitespace and case. Separate source columns were retained when sources covered the same trait. Categorical variables were expanded into indicators, preserving missingness for species absent from a source. However, the original AmphiBIO processing assigned 0 to unflagged categories among matched species. AmphiBIO defines these entries as unrecorded, rather than confirmed absence (Oliveira et al., 2017). Such recoding yields documented-versus-undocumented indicators. The AmphiBIO-derived results are retained in Table S3 with explicit flags, but do not support claims of verified presence/absence accuracy; source coding and missingness must be resolved before using them for that purpose.

The correlation workflow selected Pearson, biserial, polyserial, tetrachoric, or polychoric estimators according to automatically classified variable types, using pairwise complete observations and a minimum overlap of 30 species (Olsson, 1979; Revelle, 2026). Constant variables and numerical failures were assigned missing correlations. For each criterion, the five largest absolute item correlations were retained for inspection. Table S3 reports 44 selected criteria, not the complete 168-criterion screen. Table 3 highlights biologically interpretable examples from those results. Selecting large correlations is an exploratory procedure; the displayed values are neither average predictive performance nor cell-level agreement percentages. Latent correlation estimators require distributional assumptions beyond the fact that responses are ordinal.

2.7 Model-assisted audit

A random sample of 200 species (Table S1) was examined with Claude Opus 4.7 extended, which had not generated the LifeDive annotations. The model assigned broad taxonomic groups and constructed expected answers for approximately 50 questions it judged largely predictable from those groups. It also assessed 240 selected species-specific cells, spanning 20 species and 12 questions, using its knowledge and targeted web checks. No independent human expert adjudication was reported. The two procedures used the same auditing model, and neither the rule-covered questions nor the smaller spot-check sample constituted a uniform random sample of all cells. Section S3 gives the original prompt, an edited account of the procedures, and the limitations of the reported counts.

2.8 Phenotype-similarity network

For the same 200 species, pairwise phenotype-profile similarity was summarized using polychoric correlations across the 196 questions (Olsson, 1979; Revelle, 2026). In this calculation the paired observations are different questions, not independent species observations. The resulting coefficients are descriptive profile similarities whose interpretation depends on question content, redundancy, shared absences, and the latent-distribution assumptions. Negative correlations were retained in numerical summaries and set to zero for plotting. The nonnegative weighted matrix was drawn with qgraph version 1.10.1 using a force-directed spring layout (Epskamp et al., 2012; Fruchterman & Reingold, 1991), with nodes grouped as Arthropoda, Ascomycota, Chordata, Cnidaria, Mollusca, Tracheophyta, or Other. The layout is not a phylogenetic tree or a calibrated biological distance.

3 Results

3.1 A cross-kingdom phenotype matrix

The workflow produced a completed matrix of 35,856 species by 196 retained questions, comprising 7,027,776 entries. Of these entries, 16.6% were filled by statistical imputation after Unknown/not applicable and out-of-range responses were treated as missing. The retained questions include cellular and structural traits, anatomy, ecological associations, life history, environmental tolerances, and human-associated descriptions. Human-salience variables, including perceived beauty and nuisance status, are part of the resource but should not be confused with direct organismal measurements.

3.2 Fourteen interpretable phenotype dimensions

The 14-factor descriptive solution was interpretable in terms of broad biological profiles (Table 2). The reported primary-loading counts were 82 Plant-like, 12 Fungus-like, 19 Aquatic, 15 Insect-like, 5 Simple, 10 Robust, 12 Durable, 10 Bird-like, 5 Dangerous, 5 Shelled, 5 Drought-tolerant, 7 Parasitic, 6 Cold-tolerant, and 3 Environmentally-sensitive items, summing to 196. The pattern table is selectively displayed and rounded; Table S2 marks three unavailable Bird-like coefficients as NR. Counts are preserved from the analysis summary rather than reconstructed from the incomplete rounded table.

Fungus-like captured a distinct nonanimal syndrome, comprising lignin-degrading enzymes, siderophores, multinucleate cells, decomposer nutrition, predominantly haploid life stages, filamentous growth and extracellular digestion. Aquatic loaded on underwater survival, aquatic habitat, gill-like gas exchange, salinity tolerance and indeterminate growth, and negatively on dead-mass shedding, post-fire recovery and thermogenesis, separating aquatic profiles from terrestrial plant regrowth and endothermic animals. Insect-like loaded on chitin, chitinase, very small body mass, short lifespan and surface gas-exchange openings, and negatively on closed vascular systems, rigid internal skeletons and hemoglobin-like physiology. These dimensions summarize recognizable cross-kingdom contrasts in the selected question set.

Plant-like was the broadest dimension, contrasting cellulose production, photosynthesis, large vacuoles, and plant-associated chemistry with animal-associated physiology. The narrower profiles included nuisance and invasiveness under Robust; body size, longevity, and durable structures under Durable; appearance, aerial, and reproductive features under Bird-like; and venom, toxicity, and perceived hazard under Dangerous. Simple summarized budding, gelatinous tissues, and colonial organization. These labels describe covariance among the particular questions; they do not measure evolutionary advancement or define mutually exclusive kinds of organisms. A species can score highly on multiple profiles.

Table 2. Fourteen descriptive phenotype dimensions. Primary-loading counts and selected pattern coefficients. Factor labels summarize the included questions and do not define exclusive taxonomic groups.

Dimension	n	Defining LifeDive items (examples; loading)	Interpretation
Plant-like	82	+ cellulose (.97); + low-light response (.96); + large vacuoles (.94); + mycorrhizal symbiosis (.92); + photosynthesis / cell walls (.91); - centrioles (-.94)	Land-plant cellular, structural and ecological syndrome, contrasting with animal-like physiology.
Fungus-like	12	+ lignin-degrading enzymes (.93); + siderophores (.84); + multinucleate cells (.80); + decomposer nutrition (.69); + filamentous growth (.58)	Fungal and decomposer morphology, nutrient acquisition and physiology.
Aquatic	19	+ underwater survival (.86); + aquatic habitat (.82); + gill-like gas exchange (.82); + salinity tolerance (.65); + indeterminate growth (.63)	Aquatic habitat, gas exchange and environmental tolerance.
Insect-like	15	+ chitin (.92); + chitinase (.69); + very small body mass (.65); + short lifespan (.64); + surface gas-exchange openings (.62); - closed vascular system / internal skeleton / hemoglobin	Small arthropod-like or invertebrate body-plan syndrome.
Simple	5	+ budding (.85); + gelatinous tissue (.67); + colony or network resource sharing (.46); - intercellular cytoplasmic connections (-.55)	Simple, colonial, gelatinous or asexually reproducing forms.
Robust	10	+ weed or crop-pest status (.89); + invasive status (.88); + annoying to humans (.67); + urban survival (.61); + rapid population growth (.54)	Opportunistic, disturbance-tolerant and human-associated taxa.
Durable	12	+ hard or durable tissues sought by humans (.75); + very large mass (.63); + protective covering (.61); + lifespan over 100 years (.61); + fossil-preserving structures (.57)	Large, long-lived, structurally robust or durable organisms.
Bird-like	10	+ structural coloration (.70); + perceived beauty (.68); + external eggs, seeds or spores (.59); + blue coloring (.56); + gliding (.49)	Visually conspicuous, aerial or oviparous, bird-associated traits.
Dangerous	5	+ serious threat to human life (.73); + venom (.65); + defensive toxins (.59); + poisonous to mammals (.57); + ferocity (.46)	Human danger, venom, poison and defensive or predatory hazard.
Shelled	5	+ calcium carbonate structure (.78); + tentacles (.71); + spiral or helical structures (.45); - elongated body form (-.38); - scales (-.34)	Shelled or calcified and tentaculate invertebrate-like structures.
Drought-tolerant	5	+ long periods without water (.75); + arid or desert thriving (.72); + survival at 50 °C (.64); + desiccation and rehydration (.45); + water storage (.40)	Tolerance of aridity, heat and desiccation.
Parasitic	7	+ parasitism (.76); + close symbiosis (.59); + human-body colonization (.41); + living on host surface (.35); - bioluminescence (-.38)	Parasitic, host-associated or symbiotically dependent life histories.
Cold-tolerant	6	+ polar occurrence (.56); + freezing tolerance (.47); + seasonal long-distance movement (.45); + temperature-triggered developmental transition (.31); - tropical occurrence (-.59)	Cold-climate occurrence and cold or seasonal physiological responses.
Environmentally-sensitive	3	+ pollution or environmental sensitivity (.76); + endangered status (.64); + freshwater habitat (.62)	Environmental sensitivity, conservation concern and freshwater association.

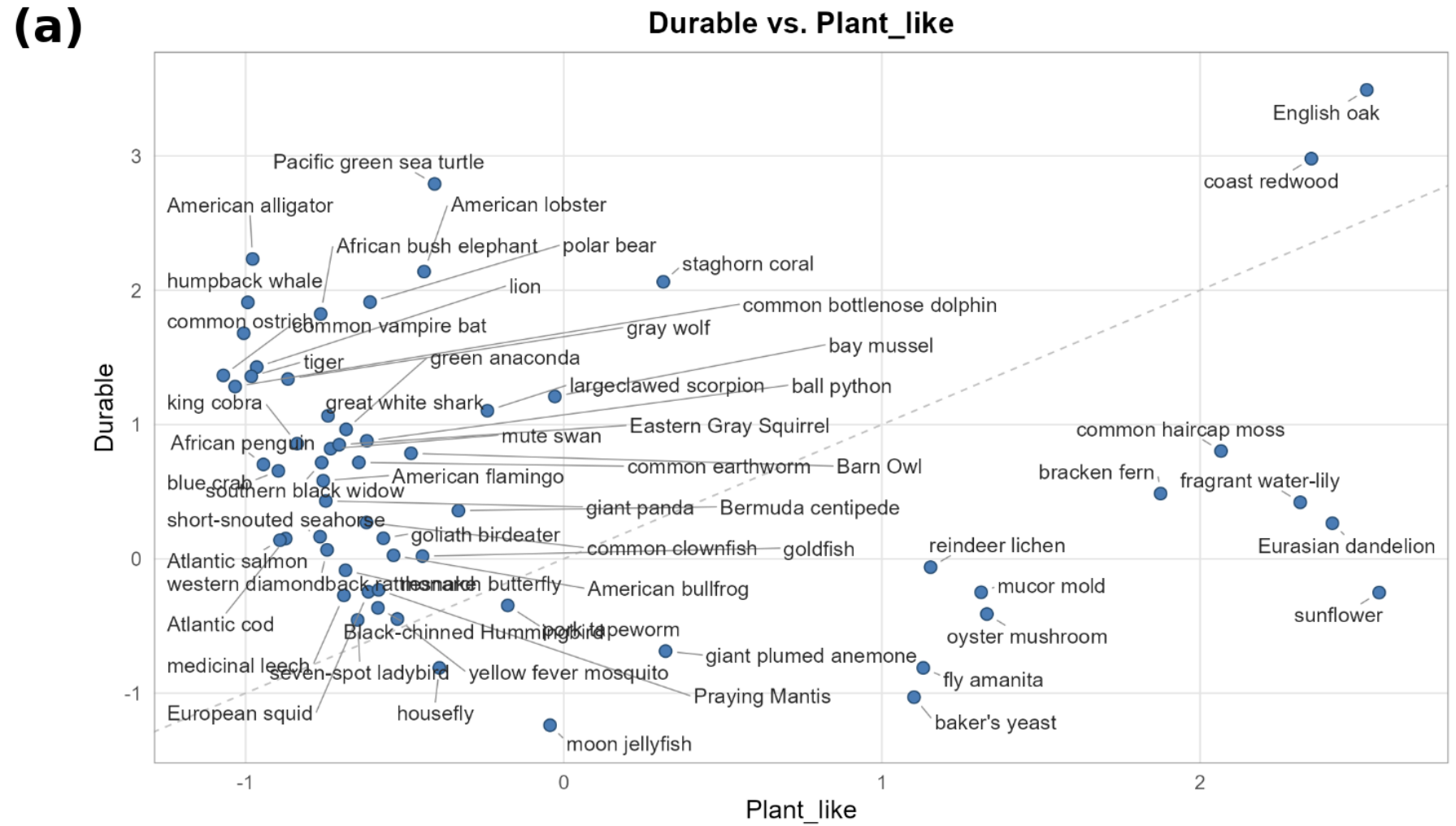
Counts sum to 196 and are preserved from the analysis summary. Coefficients are rounded to two decimals. Parasitic and Cold-tolerant directions are shown in the named orientation. The complete retained question set and displayed pattern coefficients are in Table S2; blank cells there are not zeros, and three unavailable coefficients are marked NR.

3.3 Familiar species in phenotype space

Figure 1 illustrates the scores for selected familiar species. English oak and coast redwood occupy the high-Plant-like, high-Durable region, whereas sunflower, dandelion, water-lily, and bracken fern remain high on Plant-like but lower on Durable. Many animals have negative Plant-like scores, while large or structurally robust animals tend to have higher Durable scores than smaller or delicate animals. Fungal examples occupy an intermediate Plant-like region and generally have low Durable scores.

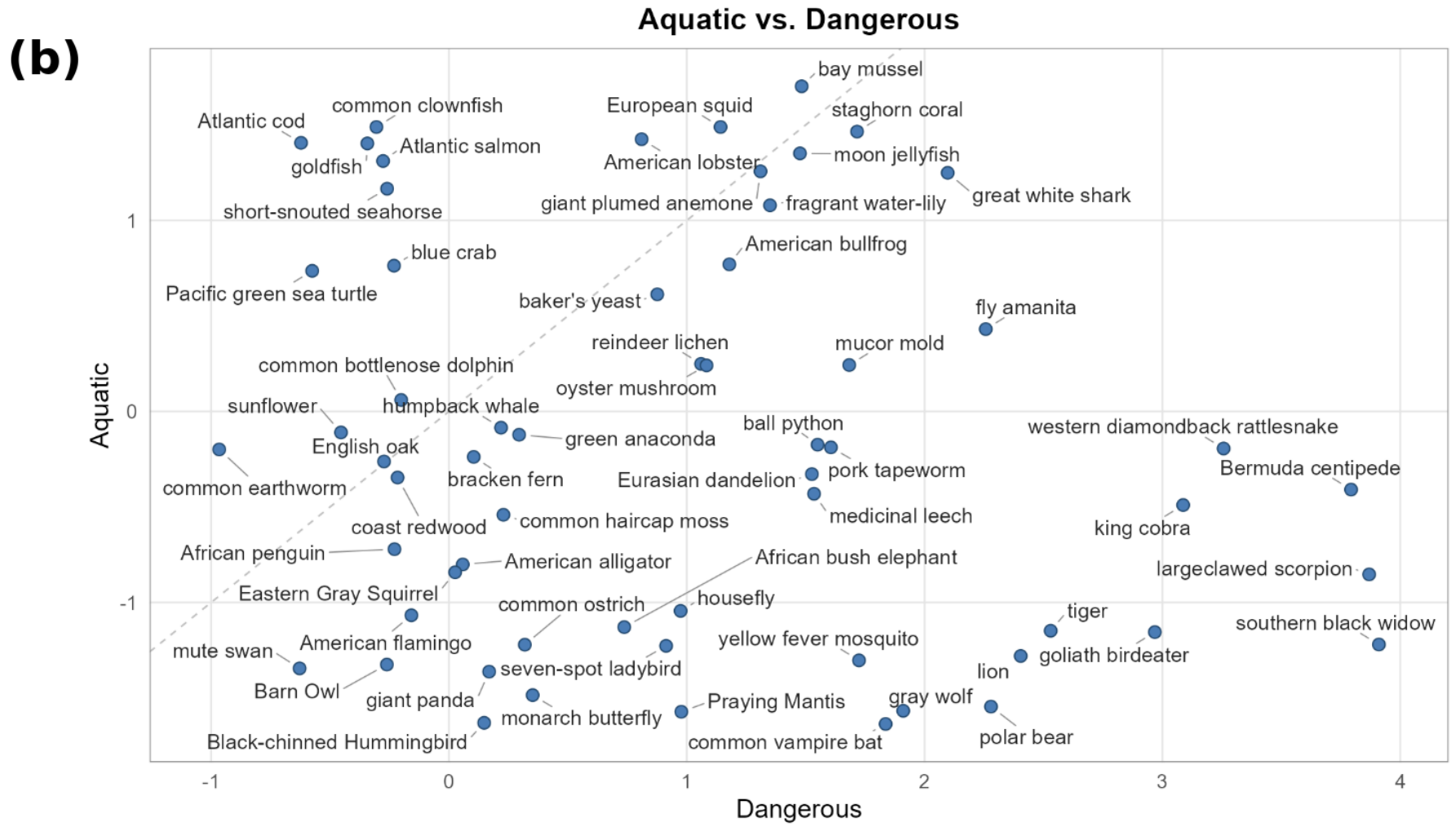
In the Aquatic-versus-Dangerous panel, many fishes and marine invertebrates have elevated Aquatic scores, and several venomous or predatory species have elevated Dangerous scores. The great white shark combines both profiles. The common bottlenose dolphin and humpback whale, however, lie near zero on the standardized Aquatic axis despite being marine mammals. This illustrates that a factor score summarizes the full pattern of included traits rather than literal habitat membership. These selected examples aid interpretation; they are not a held-out validation of the dimensions or a quantitative measure of risk to humans.

Figure 1a. Familiar species in phenotype space. Durable versus Plant-like. Axes show standardized, winsorized unit-weighted scores with small plotting jitter.



The selected species illustrate relative profiles. The dashed y = x line is a visual reference, not a fitted model or a biologically calibrated equality. Labels and point positions are retained from the analysis figure.

Figure 1b. Familiar species in phenotype space. Aquatic versus Dangerous. Continuation of Figure 1.



The axes summarize multivariate profiles. Marine mammals near zero on Aquatic illustrate that this dimension is not a direct habitat classification; Dangerous is not a calibrated measure of risk to humans.

3.4 External benchmark agreement

Of the 35,856 species, 15,525 (43.3%) matched at least one external resource, with coverage strongest for vertebrates and weaker for plants, fungi, and most invertebrates. Table 3 presents selected direct correspondences; Table S3 retains all five reported matches for each of 44 selected criteria. The displayed subset is not a comprehensive summary of all 168 criteria screened.

Broad taxonomic associations were often strong. Animalia was negatively associated with cellulose production (r = -0.99), cell walls (r = -0.97), and photosynthesis (r = -0.97), and positively associated with centrioles (r = 0.98). Plantae showed the reverse pattern. Mammalia was associated with hair, fur, or trichomes (r = 0.99) and negatively associated with the item about seeds, eggs outside the body, or spores (r = -0.98). Arthropoda and Insecta were associated with chitin (both r = 0.97), and fish-class indicators were associated with gill-like gas exchange (r = 0.96), underwater survival (r = 0.95), and aquatic lifestyle (r = 0.93). These are contrasts among broad groups, not evidence that every item is accurate within a group.

Selected ecological associations also matched their intended content: bird nocturnality with nighttime activity (r = 0.92), marine-mammal foraging with underwater survival (r = 0.97), marine-invertebrate host association with parasitism (r = 0.89), and marine-invertebrate sessility with the sessile-life item (r = 0.84). Other large correlations had no clear biological interpretation, including some matches involving lignin degradation within mammals or birds. Sparse cross-tabulations, source coding, annotation errors, and selecting extrema may all contribute. Large absolute correlations alone are therefore insufficient evidence of validity. AmphiBIO-derived associations are not used here as confirmation of biological absences because of the missing-value recoding described above.

Table 3. Selected direct correspondences with external resources. Biologically interpretable examples from the 44 selected criteria reported in Table S3.

External criterion	LifeDive question	n*	r
Animalia (taxonomy)	Does this organism produce cellulose?	8,427	-.99
Plantae (taxonomy)	Does this organism produce cellulose?	8,427	.99
Mammalia (taxonomy)	Does this organism have hair (fur) or trichomes?	8,427	.99
Arthropoda (taxonomy)	Does this organism use chitin as a structural material?	8,427	.97
Fish classes (taxonomy)	Does this organism use gills or gill like structures for gas exchange?	8,427	.96
Bird nocturnality (EltonTraits)	Does this organism do most of its activities or growth at night?	2,283	.92
Mammal nocturnality (EltonTraits)	Does this organism do most of its activities or growth at night?	1,464	.75
Marine-mammal foraging (EltonTraits)	Can this organism survive long periods under water?	1,464	.97
Mammal predation on mammals (EltonTraits)	Is this organism capable of trapping or capturing prey?	1,415	.59
Marine-invertebrate sessility (SeaLifeBase)	Is this organism sessile fixed in place for most of its adult life?	1,792	.84
Host-associated marine invertebrates (SeaLifeBase)	Is this organism parasitic on other living organisms?	1,792	.89

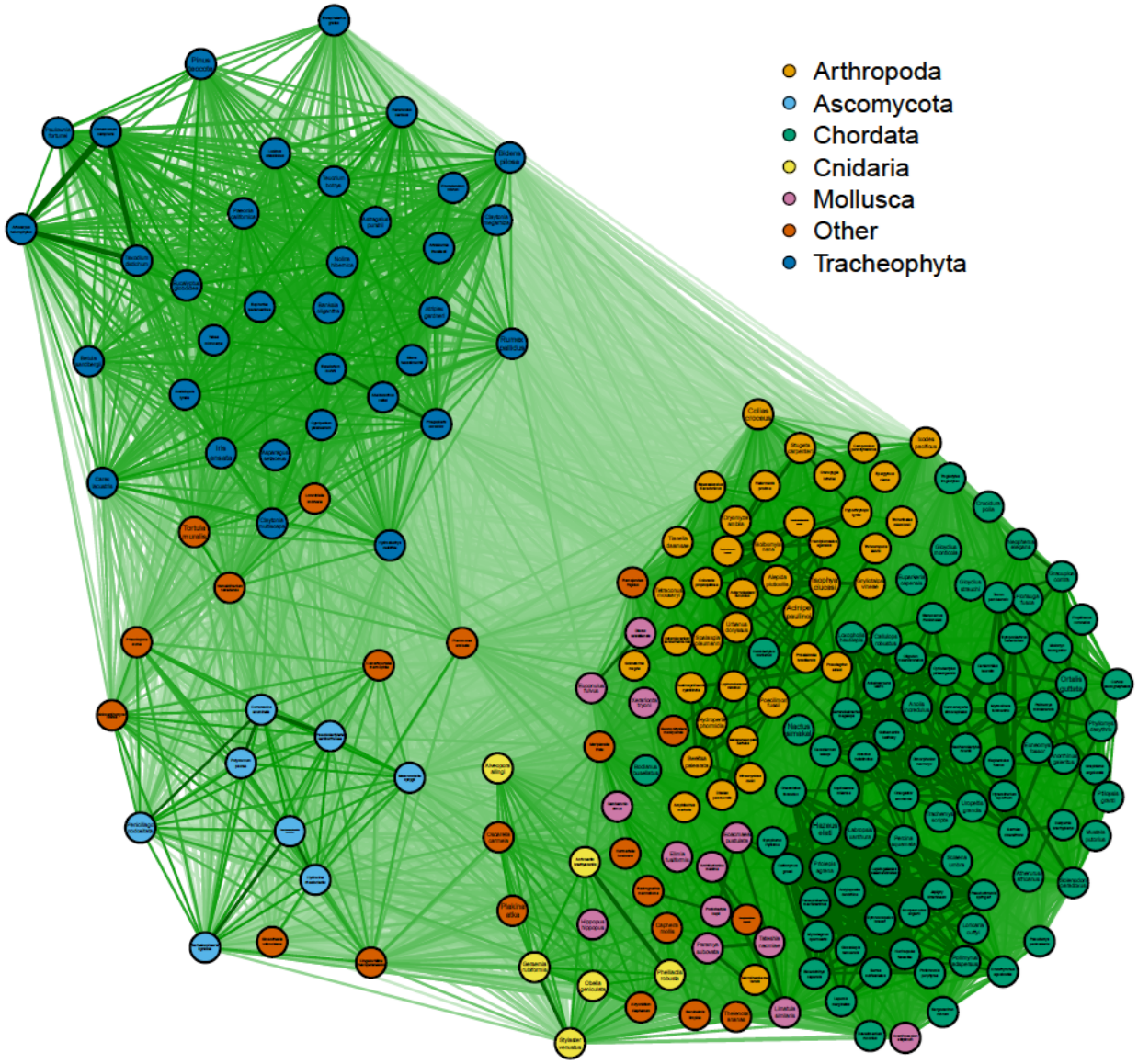
* n is the number of species with a nonmissing criterion, not a reported pair-specific complete-case denominator. Correlations were selected from comparisons with 196 LifeDive items and are exploratory associations, not accuracy percentages. AmphiBIO-derived results are flagged separately in Table S3 because unrecorded categories cannot establish absence.

3.5 Model-assisted audit and phenotype similarity

The audit reported 96.5% directional agreement for a taxonomy-rule subset with a final stated denominator of 7,641 cells, and approximately 237 agreements among 240 selected species-specific checks. The original confidence-stratum counts total 7,643 rather than 7,641; this discrepancy is retained explicitly in the audit summary. The audited matrix contained no Unknown codes, but that does not establish the raw model's abstention rate because cell-level preprocessing and imputation provenance were not resolved in the report. These selected, model-adjudicated checks provide exploratory evidence about the examined cells; they do not establish a 95-97% accuracy rate for the complete dataset or calibrated probabilities for individual response codes.

The phenotype-similarity network showed recognizable taxonomic and functional organization (Figure 2). Mean within-group polychoric similarities were 0.877 for Ascomycota, 0.853 for Cnidaria, 0.823 for Arthropoda, 0.810 for Chordata, 0.785 for Mollusca, and 0.766 for Tracheophyta, compared with 0.510 for the heterogeneous Other category. Examples of strong similarities included pairs of trees, frogs, flatfishes, and sea anemones. Some cross-group similarities also had plausible functional interpretations, such as an active predatory cuttlefish near fishes. These patterns describe organization in the generated trait matrix; they do not independently verify its entries.

Figure 2. Phenotype-similarity network of 200 species. Node colors indicate phylum-level plotting groups; Other combines taxa outside the six named phyla.



Edges represent positive polychoric profile similarities across the 196 questions; negative correlations were set to zero for plotting. The force-directed layout reflects the network's overall structure, and physical distance is not a calibrated similarity or evolutionary distance. Dense labels limit individual-node reading; Table S1 lists all 200 species.

4 Discussion

LifeDive demonstrates a practical way to assemble a broad computational phenotype matrix across animals, plants, and fungi. Its 7,027,776 completed entries contain recognizable biological contrasts, and selected external associations agree with established trait resources. The combination of broad coverage and a common question framework can support exploratory comparisons, hypothesis generation, and prioritization of expert review. The evidence is strongest for these descriptive uses; it does not establish general cell-level accuracy or predictive performance across all traits and taxa.

The reported expenditure makes the scale of the workflow concrete, but it should not be read as an estimate of the cost of producing an equivalently accurate expert-curated resource. Likewise, format constraints, retry logic, and web search were components of the workflow rather than experimentally isolated causes of performance. Web search was enabled, but its contribution was not separately tested. The roles of schema design, parsing, provenance, and expert verification are consistent with broader work on scientific use of language models (Dagdelen et al., 2024; Farrell et al., 2024; Polak & Morgan, 2024).

4.1 Guidance for users

Users should distinguish raw annotations from completed data. The raw response categories retain Unknown/not applicable values and are preferable for reviewing individual traits. The completed matrix is convenient for methods requiring complete input, but its imputed cells are predictions from other machine-generated annotations. Analyses should preserve the original category and an imputation indicator wherever these are available, and verify influential species-by-trait values against appropriate biological sources. A definite response is a model judgment, not a guarantee of correctness.

The substantive categories are ordered, but equal spacing between their numerical codes is an analytical assumption. Aggregate scores depend on the included questions and score-construction rules, and should be named and interpreted as relative profiles. For example, the Aquatic factor is not a direct habitat classification, Dangerous is not a calibrated risk measure, and Simple does not imply evolutionary primitiveness. Human-salience items describe model- and source-mediated human judgments. Their usefulness and potential biases differ from those of anatomical or physiological questions.

The species sample also reflects deliberate selection. It emphasizes broad family coverage, accepted names, common names, and familiar organisms, with differing priorities among kingdoms. Analyses of prevalence, diversity, or global biodiversity should not treat it as a representative random draw. Trait meanings may depend on life stage, sex, location, environmental conditions, or the definition of the organism; the shared question framework does not eliminate these sources of variation.

4.2 Limits of the present evidence

Several features can make apparent agreement stronger than generalizable accuracy. The factor analysis and network summarize the same generated matrix. Benchmark resources were curated outside this study but were not shown to be absent from the annotation model's training data or web retrieval. Different language models can share source material and systematic errors. Taxonomic contrasts may dominate both internal structure and external correlations, while selecting the largest correlations among many candidates favors extreme values. The 44 reported criteria do not establish performance over the entire benchmark, and some large coefficients are biologically indirect or implausible.

The model-assisted audit is limited by selected questions, selected spot-check species, fallible reference judgments, and unresolved count and preprocessing details. It cannot supply an overall accuracy estimate without a representative sampling design and independently adjudicated reference labels. Confidence-stratified agreement in the original report likewise does not establish probabilistic calibration. A future audit should sample cells across taxa and trait domains, distinguish generated and imputed values, preserve source citations and adjudication decisions, and account for clustering by species and question.

Missingness creates a separate limitation. Unknown and inapplicable responses were combined before imputation, and imputed entries may strengthen apparent covariance. The stability of factor structure and scores under alternative missing-data treatments was not established. Fourteen factors were selected for interpretability despite much larger estimates from the retention criteria. The present representation is therefore one useful descriptive reduction, not a claim that organismal phenotypes have exactly fourteen underlying dimensions. Sensitivity analyses should also examine within-clade structure and alternative similarity estimators.

The AmphiBIO recoding illustrates why benchmark provenance matters as much as model output. An undocumented category cannot automatically be treated as an observed absence. A corrected analysis should retain uncertainty or establish explicit negatives from suitable source records before estimating presence/absence performance. The original AmphiBIO correlations are preserved as flagged outputs for transparency, rather than silently relabeled as corrected results.

4.3 Reproducibility and further development

Reproducibility is best understood at the level of a versioned data product. The workflow can be rerun, but exact annotations from a commercial model with web retrieval may change with model updates, prompts, API behavior, and source availability. Reuse should identify the dataset version, model identifier, collection dates, prompt and question versions, processing code, response scale, and validation materials. Cell-level source and imputation metadata are needed to distinguish generated, predicted, checked, and unchecked entries. General concerns about factuality, source bias, and responsible open-data practice apply to this resource (Chafetz et al., 2024; Nasr & El-Deeb, 2025; Wysocka et al., 2024).

Future development can combine expert adjudication with targeted model annotation. Priorities include checking influential cells and poorly covered taxa, resolving ambiguous questions, separating unknown from inapplicable responses, verifying taxonomic identity, and correcting benchmark coding. Cross-model agreement can help triage review but should itself be validated. Predefined trait-to-criterion mappings, held-out evaluations, and comparison with simple

taxonomic prediction rules would help establish what information the annotations add beyond broad group membership.

5 Conclusion

A constrained language-model workflow can assemble a cross-kingdom phenotype resource with broad coverage and recognizable biological structure. LifeDive provides such a resource for 35,856 species, together with descriptive dimensions and exploratory comparisons. Its principal contribution is a documented computational trait layer and a concrete basis for further evaluation. The resource is most useful when its estimates, imputation, selective validation, and sampling design remain visible to users and when conclusions that depend on specific traits are checked against biological evidence.

Transparency and availability

Data and code. The LifeDive project is available through the Open Science Framework at https://osf.io/aqkhn, with DOI 10.17605/OSF.IO/AQKHN. The author's website provides an exploration interface and a download of the completed analysis matrix at https://www.mooremetrics.com/lifedive/. Reuse should cite the specific files and versions obtained. Third-party trait resources retain their providers' terms and licenses. This PDF contains the main manuscript, supplementary methods and results, an edited audit account, and Tables S1-S3; it does not substitute for the underlying data and executable code.

Version history. This permanent preprint incorporates the manuscript and supplementary material in one document. An earlier version was posted in part on Research Square at https://www.researchsquare.com/article/rs-10220777/v1. The present text narrows the interpretation of the original exploratory validation outputs; it does not report a new analysis of the source matrices.

Acknowledgments. The author thanks the curators and contributors of the public trait resources, Catalogue of Life, and GBIF for the data and taxonomy on which this work depends.

Funding. No external funding supported this work. The author bore the reported API and computing costs of approximately US$10,000.

Author contributions. Tyler M. Moore conceived the study, built the annotation pipeline, performed the analyses, prepared the figures and tables, and wrote the manuscript.

Competing interests. The author declares no competing interests and no relationship with the commercial model providers beyond use of their services as a paying customer.

Ethics. This study involved no human participants, live-animal experimentation, or field collection. It used published resources and computational annotations.

Use of generative AI. Grok generated the phenotype annotations and Claude Opus 4.7 extended performed the model-assisted audit described here. Claude Opus 4.8, GPT-5.5 Pro Extended, and Grok 4.3 were used in earlier manuscript drafting and proofreading; ChatGPT assisted with editorial review and preparation of this preprint. Responsibility for the study and its published claims rests with the author.

References

References cited in the main text and supplementary information are consolidated here.

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature, 387, 489-493. https://doi.org/10.1038/387489a0

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Castro, A., Pinto, J., Reino, L., Pipek, P., & Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. Ecological Informatics, 82, 102742. https://doi.org/10.1016/j.ecoinf.2024.102742

Catalogue of Life. (2026). Catalogue of Life checklist (Version 2026-04-18 XR) [Data set]. https://doi.org/10.48580/dgxjw

Chafetz, H., Saxena, S., & Verhulst, S. G. (2024). A fourth wave of open data? Exploring the spectrum of scenarios for open data and generative AI. New York, NY: The GovLab and Open Data Policy Lab. https://arxiv.org/abs/2405.04333

Cid del Prado-Vera, I., Ferris, H., & Subbotin, S. A. (2018). Six new and one known species of Geomonhystera (Nematoda, Monhysteridae) from moss, an epiphytic plant and soil in Mexico and Ecuador. Zootaxa, 4471, 76-96. https://doi.org/10.11646/zootaxa.4471.1.3

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., … Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15, 1418. https://doi.org/10.1038/s41467-024-45563-x

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, e70011. https://doi.org/10.1002/aps3.70011

Endara, L., Cui, H., & Burleigh, J. G. (2018). Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. Applications in Plant Sciences, 6, e1035. https://doi.org/10.1002/aps3.1035

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software, 48, 1-18. https://doi.org/10.18637/jss.v048.i04

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4, 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., … Mideo, N. (2024). The changing landscape of text mining: A review of approaches for ecology and evolution. Proceedings of the Royal Society B: Biological Sciences, 291, 20240423. https://doi.org/10.1098/rspb.2024.0423

Folk, R. A., Guralnick, R. P., & LaFrance, R. T. (2024). FloraTraiter: Automated parsing of traits from descriptive biodiversity literature. Applications in Plant Sciences, 12, e11563. https://doi.org/10.1002/aps3.11563

Froese, R., & Pauly, D. (Eds.). (2024). FishBase (Version 2024) [Data set]. World Wide Web electronic publication. https://www.fishbase.org

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21, 1129-1164. https://doi.org/10.1002/spe.4380211102

GBIF Secretariat. (2023). GBIF backbone taxonomy [Data set]. Global Biodiversity Information Facility. https://doi.org/10.15468/39omei

Gougherty, A. V., & Clipp, H. L. (2024). Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. npj Biodiversity, 3, 13. https://doi.org/10.1038/s44185-024-00043-9

H2O.ai. (2026). Build a random forest model: h2o.randomForest [Computer software documentation]. Retrieved May 21, 2026, from https://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/reference/h2o.randomForest.html

Harris, B. J., Clark, J. W., Schrempf, D., Szöllősi, G. J., Donoghue, P. C. J., Hetherington, A. M., & Williams, T. A. (2022). Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. Nature Ecology and Evolution, 6, 1634-1643. https://doi.org/10.1038/s41559-022-01885-x

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology, 17, 65-70. https://doi.org/10.1111/j.2044-8317.1964.tb00244.x

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185. https://doi.org/10.1007/BF02289447

Jackson, C. J., & Reyes-Prieto, A. (2014). The mitochondrial genomes of the glaucophytes Gloeochaete wittrockiana and Cyanoptyche gloeocystis: Multilocus phylogenetics suggests a monophyletic Archaeplastida. Genome Biology and Evolution, 6, 2774-2785. https://doi.org/10.1093/gbe/evu218

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., ... Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology, 90, 2648. https://doi.org/10.1890/08-1494.1

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., ... Wirth, C. (2020). TRY plant trait database: Enhanced coverage and open access. Global Change Biology, 26, 119-188. https://doi.org/10.1111/gcb.14904

Keck, F., Broadbent, H., & Altermatt, F. (2026). A scalable workflow for extracting species interactions from ecological literature using large language models. Ecology and Evolution, 16(11), e74626. https://doi.org/10.1002/ece3.74626

Kissling, W. D., Dalby, L., Fløjgaard, C., Lenoir, J., Sandel, B., Sandom, C., ... Svenning, J.-C. (2014). Establishing macroecological trait datasets: Digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution, 4, 2913-2930. https://doi.org/10.1002/ece3.1136

Lai, H., Liu, J., Bai, C., Liu, H., Pan, B., Luo, X., ... Ge, L. (2025). Language models for data extraction and risk of bias assessment in complementary medicine. npj Digital Medicine, 8, 74. https://doi.org/10.1038/s41746-025-01457-w

Larraín, J., Fife, A., & Atala, C. (2017). Lorentziella (Gigaspermaceae, Bryophyta) new for Chile, and lectotypification of the genus. Boletin de la Sociedad Argentina de Botanica, 52, 351-359.

Lee, S.-H., Kim, D.-S., Park, S.-H., & Park, H. (2022). Phytochemistry and applications of Cinnamomum camphora essential oils. Molecules, 27, 2695. https://doi.org/10.3390/molecules27092695

Mammides, C., Gu, H., Nimalrathna, T. S., Avicena, N. R., Papadopoulos, H., & Campos-Arceiz, A. (2026). Emerging applications of large language models in ecology and conservation science. Conservation Biology, e70287. https://doi.org/10.1111/cobi.70287

Marcos, D., van de Vlasakker, R., Athanasiadis, I. N., Bonnet, P., Goëau, H., Joly, A., ... Panousis, K. P. (2025). Fully automatic extraction of morphological traits from the web: Utopia or reality? Applications in Plant Sciences, 13, e70005. https://doi.org/10.1002/aps3.70005

Moore, T. M. (2025). Using large language models to obtain quantitative data. PsyArXiv. https://doi.org/10.31234/osf.io/aqu9x

Myhrvold, N. P., Baldridge, E., Chan, B., Sivam, D., Freeman, D. L., & Ernest, S. K. M. (2015). An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. Ecology, 96, 3109. https://doi.org/10.1890/15-0846R.1

Nasr, R. S., & El-Deeb, S. (2025). Generative AI and its potential impact on humanity. International Journal of Consumer Studies, 49, e70024. https://doi.org/10.1111/ijcs.70024

Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C., & Costa, G. C. (2017). AmphiBIO, a global database for amphibian ecological traits. Scientific Data, 4, 170123. https://doi.org/10.1038/sdata.2017.123

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44, 443-460. https://doi.org/10.1007/BF02296207

Palomares, M. L. D., & Pauly, D. (Eds.). (2024). SeaLifeBase (Version 2024) [Data set]. World Wide Web electronic publication. https://www.sealifebase.org

Pedersen, T. L. (2025). patchwork: The composer of plots (R package version 1.3.2) [Computer software]. https://patchwork.data-imaginist.com

Pérez-Harguindeguy, N., Díaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P., ... Cornelissen, J. H. C. (2013). New handbook for standardised measurement of plant functional traits worldwide. Australian Journal of Botany, 61, 167-234. https://doi.org/10.1071/BT12225

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (pp. 2463-2473). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15, 1569. https://doi.org/10.1038/s41467-024-45914-8

R Core Team. (2026). R: A language and environment for statistical computing (Version 4.6.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ranasinghe, R. A. S. N., Maduwanthi, S. D. T., & Marapana, R. A. U. J. (2019). Nutritional and health benefits of jackfruit (Artocarpus heterophyllus Lam.): A review. International Journal of Food Science, 2019, 4327183. https://doi.org/10.1155/2019/4327183

Revelle, W. (2026). psych: Procedures for psychological, psychometric, and personality research (R package version 2.6.5) [Computer software]. Evanston, IL: Northwestern University. https://CRAN.R-project.org/package=psych

RS-eco. (2023). traitdata: R package for easy access to various ecological trait data [Computer software]. https://github.com/RS-eco/traitdata

Slowikowski, K. (2026). ggrepel: Automatically position non-overlapping text labels with ggplot2 (R package version 0.9.8) [Computer software]. https://ggrepel.slowkow.com/

Soria, C. D., Pacifici, M., Di Marco, M., Stephen, S. M., & Rondinini, C. (2021). COMBINE: A coalesced mammal database of intrinsic and extrinsic traits. Ecology, 102, e03344. https://doi.org/10.1002/ecy.3344

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., Diana, E., Lehmann, G., Toren, D., Wang, J., Fraifeld, V. E., & de Magalhães, J. P. (2018). Human Ageing Genomic Resources: new and updated databases. Nucleic Acids Research, 46, D1083-D1090. https://doi.org/10.1093/nar/gkx1042

Tobias, J. A., Sheard, C., Pigot, A. L., Devenish, A. J. M., Yang, J., Sayol, F., … Schleuning, M. (2022). AVONET: Morphological, ecological and geographical data for all birds. Ecology Letters, 25, 581-597. https://doi.org/10.1111/ele.13898

Varnum, M. E. W., Baumard, N., Atari, M., & Gray, K. (2024). Large language models based on historical text could offer informative tools for behavioral science. Proceedings of the National Academy of Sciences of the United States of America, 121, e2407639121. https://doi.org/10.1073/pnas.2407639121

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41, 321-327. https://doi.org/10.1007/BF02293557

Villanueva, R., Perricone, V., & Fiorito, G. (2017). Cephalopods as predators: A short journey among behavioral flexibilities, adaptations, and feeding habits. Frontiers in Physiology, 8, 598. https://doi.org/10.3389/fphys.2017.00598

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (R package version 4.0.3). New York, NY: Springer-Verlag. https://ggplot2.tidyverse.org

Wilhite, L. P., & Toliver, J. R. (1990). Taxodium distichum (L.) Rich. In R. M. Burns & B. H. Honkala (Tech. Coords.), Silvics of North America: Vol. 1. Conifers (Agriculture Handbook 654). Washington, DC: U.S. Department of Agriculture, Forest Service.

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M. M., & Jetz, W. (2014). EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. Ecology, 95, 2027. https://doi.org/10.1890/13-1917.1

Wysocka, M., Wysocki, O., Delmas, M., Mutel, V., & Freitas, A. (2024). Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. Journal of Biomedical Informatics, 158, 104724. https://doi.org/10.1016/j.jbi.2024.104724

xAI. (2026). Batch API [Computer software documentation]. Retrieved May 21, 2026, from https://docs.x.ai/developers/advanced-api-usage/batch-api

Supplementary information

S1 Supplementary methods

Organism selection and taxonomic source

The taxon-selection workflow aimed to represent every eligible family while moderately preferring species with common names. It used a Catalogue of Life Data Package (ColDP), read NameUsage.tsv, normalized prefixed column names, and retained species-rank records with accepted or provisionally accepted status (Catalogue of Life, 2026). Taxonomic identifiers and available kingdom, phylum, class, order, family, and genus fields were retained. Binomials were constructed from genericName and specificEpithet when available, with scientificName as a fallback. English vernacular names were merged from VernacularName.tsv, using the first English record when several were available.

The representative organism selection focused on Animalia, Plantae, and Fungi, including approximately 5% extinct species. The script then selected a target set by first choosing the highest-priority species in each family, thereby preserving broad family-level taxonomic coverage, and then filling remaining slots from the highest-priority species not already selected. The priority score favored fully accepted names over provisionally accepted names, species with an English common name, and Animalia/Plantae over Fungi; a small random component was added only as a tie-breaker, with a fixed random seed for reproducibility. This resulted in the initial set of 35,000 organisms. Further inspection of the list revealed that many commonly known organisms (e.g., barn owl, Brazilian wandering spider) were missing. I therefore asked an LLM (GPT 5.5 Pro) to compile a list of well-known organisms that were not included in the list of 35,000. This resulted in a final list of 35,856 organisms.

LLM-based phenotype annotation

Phenotype annotations were collected with the xAI Grok API using batch processing. Batch processing was selected because the task consisted of many independent organism-by-question inference requests, a use case for which asynchronous batch APIs are designed (xAI, 2026). In the R workflow, API credentials were loaded from an environment file, requests were submitted to the xAI /v1/batches endpoint, and individual inference jobs targeted the /v1/responses endpoint. The scripted model name was grok-4-1-fast-reasoning, and each request enabled the web_search tool so that the model could use web search for obscure taxa.

Each species was evaluated using substantive response codes 0 = Definitely No, 1 = Probably No, 3 = Probably Yes, and 4 = Definitely Yes, with code 2 reserved for Unknown or not applicable. The final analysis retained 196 phenotype questions. Names were sanitized for JSON and comma-separated output. Each request listed three organisms, the current question subset, and the response definitions, and required one row per organism with one numeric column per question. Every cell was to be populated, using code 2 when an answer could not be determined. This is a four-category substantive response scale with a separate missing/inapplicable category, rather than a five-level trait-intensity continuum.

The batch workflow included several quality-control steps. JSONL requests were generated and validated before upload. After a batch was created, the script saved the batch identifier, polled batch status, resumed from saved identifiers when possible, retrieved completed results using

pagination, and checkpointed parsed data. The parser stripped code-fence formatting, read the CSV response, aligned question columns using Q-number headers when available, padded or trimmed rows if a response returned too few or too many organism rows, replaced nonnumeric or missing cells with the Unknown code, and logged problematic requests. Responses with unreliable column shifts were discarded and replaced with Unknown values rather than silently treated as valid data. A retry JSONL file was generated for requests with extraction errors, missing response text, assignment errors, or unreliable column alignment. Because the values are model-generated annotations, they should be interpreted as computational phenotype estimates rather than direct measurements or expert-curated observations.

Data assembly and imputation

The phenotype collection was conducted in multiple item subsets to avoid overloading individual LLM requests. The script read 11 organism-by-item CSV matrices, assigned the original phenotype-question text as column names, and merged the subsets by organism. The Unknown/not applicable response code was then treated as missing. Values outside the response scale were also set to missing. Items with very high missingness were excluded by retaining columns with fewer than 25,000 missing values; the retained set was renamed Q1 through Q196 for imputation and subsequent factor analysis.

Missing item responses were imputed using random forest classification, an ensemble method based on aggregating many decision trees (Breiman, 2001). The imputation code initialized an H2O cluster with 14 threads and iterated through the 196 items in random order with a fixed seed. For each target item, rows with observed values on that item were used as the training set; the target item was converted to a categorical outcome; all other retained phenotype items were used as predictors; and h2o.randomForest was used to train a multinomial classifier (H2O.ai, 2026). Predicted classes were then inserted only for cases missing on the target item. This process yielded a complete organism-by-phenotype matrix for the retained 196 items.

Factor analysis and score construction

After imputation, original codes 0, 1, 3, and 4 were recoded to 0, 1, 2, and 3. The matrix was analyzed with psych::irt.fa, which performs exploratory factor analysis of tetrachoric or polychoric correlations and expresses associated parameters in IRT terms (R Core Team, 2026; Revelle, 2026). A 14-factor solution with promax rotation was selected for interpretability (Fabrigar et al., 1999; Hendrickson & White, 1964). Parallel analysis suggested 58 factors and the minimum-average-partial criterion suggested 41 (Horn, 1965; Velicer, 1976), so 14 should be understood as a descriptive reduction. Unit-weighted sum scores based on the loading patterns were used instead of theta scores to reduce domination by a small number of highly discriminating items. The selectively displayed, rounded loading table is insufficient to reproduce every scoring detail; exact item selection, reverse scoring, and transformations require the analysis code.

The 14 score dimensions were labeled on the basis of high-loading items as Plant-like, Fungus-like, Aquatic, Insect-like, Simple, Robust, Durable, Bird-like, Dangerous, Shelled, Drought-tolerant, Parasitic, Cold-tolerant, and Environmentally-sensitive. The signs of the Parasitic and Cold-tolerant scores were reversed to make higher scores correspond to higher levels of the named construct. For visualization only, a small uniform random jitter was added to the factor scores to reduce overplotting; the script notes that these jittered scores were not

intended for release. Scores were then winsorized with a trim of 0.0005 and standardized before graphing.

Graphing of factor scores

The plotting workflow read a separate file of familiar organisms and their scores, then produced two scatterplots with ggplot2, ggrepel, and patchwork (Pedersen, 2025; Slowikowski, 2026; Wickham, 2016). Panel (a) plots Durable against Plant-like, and panel (b) plots Aquatic against Dangerous. Common names were placed with repelled labels. The dashed y = x line is a visual reference, not a fitted model. Scores were winsorized and standardized with small jitter for display only. The selected examples illustrate score interpretation rather than validate the scores against new outcomes.

Validation

The validation strategy examined several aspects of the LifeDive resource. External comparisons assessed agreement with existing resources; a model-assisted audit examined selected cells among 200 randomly sampled species; and factor analysis and network visualization summarized internal structure. These assessments answer different questions. In particular, selection of random species does not make the subsequently chosen audit questions or spot-check cells a representative sample of all 7,027,776 entries.

External benchmark criteria

The external benchmark used public trait resources and a taxonomic backbone. EltonTraits supplies bird and mammal diet, foraging, mass, and activity information (Wilman et al., 2014). Other resources were PanTHERIA (Jones et al., 2009), MammalDIET (Kissling et al., 2014), AmphiBIO (Oliveira et al., 2017), SeaLifeBase (Palomares & Pauly, 2024), FishBase (Froese & Pauly, 2024), and the traitdata taxonomy table incorporating GBIF Backbone Taxonomy (GBIF Secretariat, 2023; RS-eco, 2023). The workflow also describes overlapping body-mass sources from the Amniote Life-History Database and AnAge (Myhrvold et al., 2015; Tacutu et al., 2018). Separate source tables should not be assumed to represent independent observations, and the benchmark was not established as absent from model training or retrieval sources.

Data preparation

Species were matched by normalized genus-species binomials. Overlapping source variables were retained separately. Categorical traits were expanded into indicators, keeping species not represented by a source missing. FishBase indicators coded -1 = yes and 0 = no were converted to 1/0. In the original AmphiBIO processing, unflagged categories among matched species were set to 0. The source documentation defines such NA entries as unrecorded categories, not verified absence (Oliveira et al., 2017). Associations from that processing therefore compare documented and undocumented occurrences and can be affected by unequal recording completeness. The three displayed AmphiBIO criteria are flagged in Table S3 and are not used as confirmation of presence/absence validity; their coding must be verified before any corrected analysis. The benchmark contained 168 numeric columns, and 15,525 species (43.3%) matched at least one external resource.

Correlation methods

The correlation workflow classified variables by observed values: two distinct nonmissing values were treated as dichotomous; integer columns with three to seven distinct values and range no greater than 7 were treated as polytomous; other numeric columns were treated as continuous. It then selected Pearson for continuous-continuous pairs, biserial for continuous-dichotomous pairs, polyserial for continuous-polytomous pairs, tetrachoric for dichotomous-dichotomous pairs, and polychoric for two polytomous variables or a polytomous-dichotomous pair (Olsson, 1979; Revelle, 2026). Pairwise complete cases were used, with a minimum overlap of 30. Constant variables and estimation failures were assigned NA. The five largest absolute item correlations were retained per criterion, and 44 selected criteria are reproduced in Table S3. These choices describe the implemented screen; automatic type classification does not establish biological ordering or latent-normal assumptions, and selected maxima are not unbiased performance estimates.

Random n = 200 subsample approach

Two hundred distinct species were randomly sampled from LifeDive for more detailed exploratory assessment. Table S1 lists their phylum, class, and order as retained in the analysis. The audit and similarity network used this same species subset.

Model-assisted audit

Claude Opus 4.7 extended, which had not generated the LifeDive values, used its knowledge and web searches to construct a taxonomy-rule comparison and a smaller species-specific spot check. The report states 7,641 cells in the final rule evaluation and 240 cells in the spot check, while some intermediate and confidence-stratum counts differ. Section S3 retains the prompt and gives an edited summary with explicit interpretive limits. Neither procedure was an independent expert adjudication or a uniform probability sample of all cells.

Organism phenotype-similarity network

Using the n=200 deeper-validation sub-sample - i.e., from an organism-by-phenotype matrix containing 200 organisms scored across the 196 ordinal phenotype questions - an organism phenotype-similarity network was constructed. Taxonomic metadata were used to assign organisms to seven plotting groups: Arthropoda, Ascomycota, Chordata, Cnidaria, Mollusca, Tracheophyta, and "Other." The "Other" category included organisms outside the six focal phyla and was therefore expected to be biologically heterogeneous. Organism labels were shortened to binomial scientific names for readability, and common names were excluded from node labels to reduce visual crowding.

Pairwise species-profile similarities were computed as polychoric correlations across the 196 questions (Olsson, 1979; Revelle, 2026). Although responses are ordered, questions are heterogeneous and often related to one another; using them as paired observations does not automatically meet the estimator's assumptions. Shared negative responses and repeated trait content can contribute to high similarity. The coefficients are therefore treated as descriptive summaries. Negative correlations were retained in numerical results and set to zero in the plotted edge matrix.

The resulting nonnegative weighted matrix was drawn with qgraph using a force-directed spring layout (Epskamp et al., 2012; Fruchterman & Reingold, 1991). Edge weights are proportional to

positive similarities. Layout depends on the complete network and is not a calibrated mapping of pairwise similarities to physical distances.

Node colors show the phylum-level plotting groups. Because questions mix biological traits and human-associated descriptions, the graph summarizes the chosen phenotype profiles; it should not be read as a phylogenetic tree or as independent validation of its individual entries.

S2 Supplementary results

LifeDive phenotype matrix

The completed matrix contains 35,856 species by 196 retained questions, or 7,027,776 entries. Sampling aimed for family-level breadth and favored accepted names and common names. The questions span cellular, physiological, anatomical, behavioral, ecological, life-history, and human-salience domains, applying the same response framework across the sampled kingdoms.

Unknown/not applicable and out-of-range responses were treated as missing. Random-forest classification then completed the retained matrix; 16.6% of its entries were imputed. The completed matrix supports the reported descriptive analyses but combines original annotations and values predicted from other annotations.

Factor-analysis results

The 14-factor promax-rotated solution was readily interpretable from the displayed pattern coefficients (Table S2). Assigning each item to the factor on which it had the largest absolute loading yielded the following primary-loading counts: 82 Plant-like items, 12 Fungus-like items, 19 Aquatic items, 15 Insect-like items, 5 Simple items, 10 Robust items, 12 Durable items, 10 Bird-like items, 5 Dangerous items, 5 Shelled items, 5 Drought-tolerant items, 7 Parasitic items, 6 Cold-tolerant items, and 3 Environmentally-sensitive items.

The Plant-like factor was the broadest and strongest dimension. High positive loadings included cellulose production (.97), etiolation or comparable low-light responses (.96), large fluid-filled vacuoles (.94), mycorrhizal symbiosis (.92), production of fruit or fruiting bodies (.91), phenolic compounds (.91), photosynthesis (.91), boron use (.91), and cell walls (.91). Strong negative loadings on centrioles (-.94), glycogen storage (-.93), rapid response to mechanical contact (-.92), sodium physiology (-.92), externally acquired vitamins or cofactors (-.85), and motile cells (-.85) reinforced the interpretation that this dimension separates plant-like trait profiles from animal-like or otherwise nonplant physiological profiles.

The Fungus-like factor was also coherent. Its strongest positive indicators were lignin-degrading enzymes (.93), siderophore production (.84), multinucleate cells (.80), decomposer nutrition (.69), a predominantly haploid life stage (.63), filamentous growth (.58), and extracellular digestion (.58). Negative loading on multiple distinct tissue types (-.64) further supported the interpretation of this factor as a fungal/decomposer morphology and physiology dimension rather than simply a second plant axis.

The Aquatic factor captured aquatic habitat and physiology. The strongest positive indicators were survival under water (.86), primarily aquatic habitat (.82), gill or gill-like gas exchange (.82), salinity tolerance (.65), and indeterminate growth (.63). Negative loadings on periodic shedding of dead mass (-.69), rapid post-fire recovery (-.65), and metabolic heat generation (-.55) helped

differentiate the aquatic profile from terrestrial plant regrowth and endothermic animal profiles. The Insect-like factor loaded highly on chitin use (.92), chitinase or chitin-degrading enzymes (.69), very small body mass (.65), short lifespan (.64), and specialized surface openings for gas exchange (.62), with strong negative loadings on closed circulatory or vascular systems (-.83), rigid internal skeletons (-.73), and hemoglobin-like physiology (-.73). These loadings are consistent with an arthropod-like or small invertebrate-like structural dimension.

Several narrower factors represented interpretable ecological or morphological dimensions. The Simple factor loaded on budding reproduction (.85), gelatinous or jelly-like body composition (.67), and colony or network resource sharing (.46), while negatively loading on intercellular cytoplasmic connections (-.55). The Robust factor was defined by nuisance or opportunistic ecological status, including being a weed, crop pest, or agricultural nuisance (.89), invasive in some areas (.88), annoying to many humans (.67), urban survival (.61), rapid population growth (.54), and allergenicity (.53). The Durable factor was marked by hard or durable tissues sought by humans (.75), large body size (.63), protective outer covering (.61), longevity greater than 100 years (.61), fossil-preserving structures (.57), and ecosystem engineering (.46), with negative loadings on hamster-scale smallness (-.59) and fragility (-.49). The Bird-like factor combined appearance and aerial/reproductive features, including structural coloration (.70), perceived beauty (.68), external eggs/seeds/spores (.59), blue coloration (.56), gliding (.49), orange coloration (.45), and internal gas-filled cavities (.44).

The remaining factors were more compact but substantively clear. The Dangerous factor loaded on serious threat to human life (.73), venom (.65), defensive toxins (.59), mammalian poison (.57), and ferocity (.46). The Shelled factor loaded most strongly on calcium carbonate structural material (.78) and tentacles (.71), with additional negative loadings on elongated body form (-.38) and scales (-.34). The Drought-tolerant factor loaded on long periods without water (.75), arid or desert thriving (.72), survival at 50 degrees Celsius for 100 hours (.64), desiccation and rehydration (.45), and water-storage structures (.40). The Parasitic factor loaded on parasitism (.76), dependence on close symbiosis (.59), human-body colonization (.41), and living on the surface of another organism (.35). The Cold-tolerant factor was indicated by polar or Arctic/Antarctic occurrence (.56), freezing tolerance (.47), seasonal long-distance movement (.45), and temperature-triggered developmental transitions (.31), with a negative loading on tropical occurrence (-.59). The Environmentally-sensitive factor loaded on sensitivity to pollution or environmental change (.76), endangered status (.64), and freshwater habitat (.62). Overall, the solution recovered broad biological dimensions that are plausible across kingdoms while also identifying narrower, interpretable ecological and structural trait clusters.

Graphing of factor scores

The factor-score scatterplots provided an intuitive check on the interpretability of selected dimensions (Figure 1). In the Durable versus Plant-like panel, woody plants occupied the high Plant-like and high Durable region: English oak and coast redwood had among the highest scores on both dimensions. Other plants, such as sunflower, Eurasian dandelion, fragrant water-lily, bracken fern, and common haircap moss, also scored high on Plant-like but lower on Durable, consistent with their nonwoody or less massive forms. Most animals had negative Plant-like scores. Large or structurally robust animals and animal-like taxa, including American alligator, African bush elephant, American lobster, polar bear, Pacific green sea turtle, humpback whale, and staghorn coral, occupied higher positions on Durable than smaller or more delicate organisms such as moon jellyfish, housefly, monarch butterfly, and European squid. Fungal examples tended to cluster at positive Plant-like but low Durable scores, reflecting shared nonanimal properties without wood-like durability.

The Aquatic versus Dangerous panel similarly separated familiar trait profiles. Fish and marine organisms, including Atlantic cod, Atlantic salmon, goldfish, common clownfish, short-snouted seahorse, blue crab, European squid, bay mussel, staghorn coral, giant plumed anemone, and moon jellyfish, had high Aquatic scores. Venomous or predatory organisms such as southern black widow, largeclawed scorpion, Bermuda centipede, king cobra, tiger, lion, goliath birdeater, gray wolf, polar bear, and common vampire bat had high Dangerous scores and generally lower Aquatic scores. The great white shark occupied the high-Aquatic/high-Dangerous region, demonstrating that the factors can jointly represent organisms that combine aquatic life with high hazard. Conversely, aquatic but comparatively low-danger organisms such as common bottlenose dolphin and humpback whale were located away from the high-Dangerous cluster. The plotted examples illustrate the relative profiles, but they are not a held-out validation. In particular, dolphins and whales have Aquatic scores near zero, so that axis cannot be interpreted as a direct habitat classifier.

Table S3 reports the five largest absolute item correlations for each of 44 selected criteria. These cover taxonomy and selected ecological traits from EltonTraits, PanTHERIA, MammalDIET, AmphiBIO, and SeaLifeBase. Broad taxonomic indicators often have large associations with recognizable trait contrasts: Animalia with cellulose (-0.99), cell walls (-0.97), and centrioles (0.98); Plantae with cellulose (0.99), photosynthesis (0.97), and cell walls (0.97); and Mammalia with hair, fur, or trichomes (0.99). The mammal indicator is negatively associated (-0.98) with the question about seeds, eggs outside the body, or spores. The qualifier "outside the body" is material to interpreting this item. These are selected latent-correlation estimates, not percentages of correctly classified cells.

Among selected within-taxon comparisons, nighttime activity correlated with bird nocturnality at 0.92, with the EltonTraits mammal-nocturnality indicator at 0.75, and with the corresponding diurnal indicator at -0.73. Marine-mammal foraging correlated with underwater survival at 0.97. A mammal-eats-mammals indicator correlated with prey capture at 0.59, perceived ferocity at 0.58, and black coloration at 0.55; the last is less directly interpretable. Marine-invertebrate sessility correlated with the sessile-life item at 0.84. AmphiBIO-derived associations remain visible in Table S3 but are excluded from affirmative presence/absence validation because of their source-coding uncertainty.

Some large correlations are biologically implausible or indirect. Examples include lignin-degradation items among birds or mammals and a genus-age item associated with marine

sessility. Sparse cross-tabulations, boundary estimates, annotation mistakes, criterion coding, and selection of extrema may all contribute; the retained results do not isolate a single cause. The appropriate conclusion is that selected associations recover recognizable contrasts alongside problematic matches, rather than that every large coefficient validates a trait.

Organism phenotype-similarity network

The network recovered a strong phylum-level signal. Mean within-phylum polychoric correlations were high for Ascomycota (mean ρ = .877), Cnidaria (ρ = .853), Arthropoda (ρ = .823), Chordata (ρ = .810), Mollusca (ρ = .785), and Tracheophyta (ρ = .766). In contrast, the "Other" category had a much lower mean within-group correlation (ρ = .510), consistent with its composition as a residual group containing algae, bryophytes, nematodes, sponges, echinoderms, annelids, and other non-focal lineages. The strongest between-phylum similarities were among animal groups, particularly Chordata-Arthropoda (mean ρ = .689), Chordata-Mollusca (ρ = .680), and Arthropoda-Mollusca (ρ = .668). Tracheophytes were much less similar to animal groups, including Chordata (ρ = .171), Arthropoda (ρ = .233), and Mollusca (ρ = .205), explaining the visual separation of the plant cluster from the animal-dominated portion of the network.

At the organism level, the strongest positive correlations were generally interpretable as either close taxonomic similarity or broad functional similarity. Examples of very strong associations occurred among trees: jackfruit, Artocarpus heterophyllus, and camphor tree, Cinnamomum camphora (ρ = .997); jackfruit and bald cypress, Taxodium distichum (ρ = .989); and camphor tree and bald cypress (ρ = .985). Other strong positive examples included frog-frog similarity between Boulenger's callulops frog, Callulops robustus, and the San Lorenzo robber frog, Serranobatrachus megalops (ρ = .974); flatfish similarity between the threespot sand flounder, Ancylopsetta dendritica, and the patchtail tonguefish, Symphurus rhytisma (ρ = .973); reef-fish similarity between the yellowtail tubelip wrasse, Labropsis xanthura, and the network reefgoby, Priolepis agrena (ρ = .972); and sea-anemone similarity between Acricoactis brachyacontis and Phelliactis robusta (ρ = .967). The strongest negative correlations were much weaker in magnitude and were not displayed in the graph because negative values were set to zero prior to plotting. Examples included jackfruit and the loricariid catfish Loricaria cuffyi (ρ = -.164), jackfruit and the glass knifefish Japigny kirschbaum (ρ = -.116), and common lake sedge, Carex lacustris, and the Haitian solenodon, Solenodon paradoxus (ρ = -.115).

The positions of the major clusters are biologically plausible. Tracheophytes formed a dense and spatially distinct cluster, consistent with the shared plant functional traits represented in the phenotype set, such as photosynthesis, cell walls, cellulose-rich and lignified tissues, perennial growth, vascular or structural support, seeds, secondary compounds, and plant-specific ecological interactions. Plant functional traits are commonly interpreted as integrated morphological, physiological, and phenological features that reflect ecological strategy, which helps explain why functionally similar plants can cluster together even when they are not close relatives (Pérez-Harguindeguy et al., 2013).

The animal groups occupied the opposite side of the network. Chordates and arthropods were adjacent because both groups include many active, bilaterally organized animals with complex sensory responses, external or internal support structures, motility, feeding on other organisms, and egg- or gamete-based reproduction. Molluscs occupied a boundary region between chordates, arthropods, and cnidarians, which is also reasonable: the molluscan sample includes active cephalopods, terrestrial gastropods, marine bivalves, limpets, and other forms that share

different subsets of traits with different animal groups. Cnidarians clustered near other aquatic invertebrates, reflecting traits such as radial organization, tentacles, sessility or limited adult movement in many taxa, gelatinous or colonial forms, and calcium-carbonate-associated structures in some species.

Several boundary placements were especially informative. The mollusc Acanthosepion ellipticum, a cuttlefish, was positioned near chordates rather than deeply within the molluscan region. This is biologically sensible in a phenotype network. Cuttlefishes are cephalopod molluscs, but cephalopods are active marine predators with sophisticated sensory systems and behavioral flexibility; in broad phenotype space, an active, visually oriented, predatory, bilaterian, swimming cuttlefish can resemble fishes more than it resembles slow terrestrial snails or sessile bivalves (Villanueva et al., 2017). The correlations support this interpretation: Acanthosepion ellipticum correlated strongly with multiple chordate fishes, including Gross's dragonet, Callionymus grossi (ρ = .866), the common silver-biddy, Gerres subfasciatus (ρ = .842), Englert's scorpionfish, Scorpaenodes englerti (ρ = .836), and the freetail brotula, Bidenichthys capensis (ρ = .833).

By contrast, the mollusc Discus catskillensis, the angular disc land snail, fell closer to arthropods. This placement is also plausible. Discus catskillensis is a small terrestrial gastropod with a protective shell, and in the phenotype matrix it shares a "small terrestrial invertebrate" syndrome with many arthropods: small body size, external protective structures, egg production, terrestrial microhabitat use, vulnerability to desiccation and predation, and relatively limited locomotion compared with vertebrates. Its strongest correlations still included other molluscs, such as the tawny glass snail Euconulus fulvus (ρ = .913) and the bicolor cactussnail Xerarionta tryoni (ρ = .876), but it also correlated strongly with arthropods, including the oribatid mite Bipassalozetes macedonicus (ρ = .821) and the mite Asternolaelaps fecundus (ρ = .802). Thus, its position reflects phenotypic overlap between terrestrial gastropods and small arthropods rather than a taxonomic ambiguity.

The exceptionally strong links among jackfruit, camphor tree, and bald cypress are also interpretable, but they illustrate an important distinction between phenotypic and phylogenetic similarity. Jackfruit is a tropical angiosperm tree in Moraceae, camphor tree is an aromatic evergreen tree in Lauraceae, and bald cypress is a long-lived deciduous conifer. These taxa are not close phylogenetic relatives; however, all are large, perennial, woody, seed-producing plants with substantial standing biomass, durable structural tissues, long life spans, secondary chemistry, and human-recognized uses or cultivation histories. These shared functional and architectural traits can generate very high phenotype correlations even across deep plant lineages (Lee et al., 2022; Ranasinghe et al., 2019; Wilhite & Toliver, 1990). Therefore, the strong bonds among these three trees are biologically plausible as functional-trait similarities, but they should not be interpreted as evidence of close evolutionary relationship.

The "Other" nodes generally occupied biologically interpretable positions rather than forming a coherent cluster. Gloeochaete wittrockiana, a glaucophyte alga, was positioned near Ascomycota. This placement makes sense only in phenotype-space terms: Gloeochaete is not a fungus, but as a small, cell-walled, non-animal, aquatic/microbial photosynthetic eukaryote it shares some coarse traits with fungi and other non-animal organisms. Its strongest correlation was with the green alga Pteromonas aculeata (ρ = .853), followed by more moderate relationships with fungi and plant-like organisms. Thus, proximity to Ascomycota reflects overlapping non-animal phenotype states, not fungal identity (Jackson & Reyes-Prieto, 2014).

The bryophytes Lorentziella imbricata and Tortula muralis were located near Tracheophyta, which is one of the clearest biologically expected placements among the "Other" nodes. Bryophytes are nonvascular land plants, whereas tracheophytes are vascular land plants; despite major differences in vascular tissue, sporophyte dominance, and body architecture, they share many land-plant traits, including photosynthesis, cell walls, spores or gametes, terrestrial growth, and responses to water and light. Lorentziella imbricata correlated strongly with the wall screw-moss Tortula muralis (ρ = .843) and with several tracheophytes, including Mohr's thoroughwort, Eupatorium mohrii (ρ = .838), and Rydberg's spring-beauty, Claytonia multiscapa (ρ = .835). Tortula muralis showed a similar pattern, correlating strongly with the moss Homalothecium nevadense (ρ = .890), Lorentziella imbricata (ρ = .843), and several tracheophytes. This supports the interpretation that "Other" is a plotting category, not a biological group (Harris et al., 2022; Larraín et al., 2017).

The nematode Geomonhystera mexiquense was located near arthropods, and this is also biologically plausible. Nematodes and arthropods are not the same phylum, but both belong to the molting-animal clade Ecdysozoa, and the phenotype set captures broad features of small bilaterian invertebrates. Geomonhystera mexiquense correlated most strongly with arthropods, including the mite Lophonotacarus minutus (ρ = .911), the bush-cricket Isophya ciucasi (ρ = .874), the mite Asternolaelaps fecundus (ρ = .868), and the mite Hemipteroseius ageneius (ρ = .861). Thus, its position near arthropods is one of the more biologically coherent "Other" placements (Aguinaldo et al., 1997; Cid del Prado-Vera et al., 2018).

Finally, Pteromonas aculeata occupied a central, weakly assigned position. This is consistent with its biology as a green alga: it shares photosynthetic and cell-wall traits with plants and other algae, but lacks vascular tissue, woody architecture, seeds, and most macroscopic tracheophyte traits. At the same time, its small size, aquatic habit, motile-cell features, and microbial-like body organization distinguish it from most land plants and allow moderate similarity to several non-plant groups. Its group-level affinities were diffuse rather than concentrated, with mean correlations ranging from ρ = .358 to ρ = .498 across the plotted phylum categories. The strongest organism-level association was with Gloeochaete wittrockiana (ρ = .853), another algal "Other" taxon. Therefore, its central location is best interpreted as a mixed-trait or bridging position, not as membership in any of the focal phylum clusters.

The network organizes many broad trait profiles in recognizable ways, including woody plants, fungi, vertebrates, and several invertebrate groups. These interpretations are descriptive and often post hoc. Apparent agreement with taxonomy can partly follow from questions about taxonomically characteristic properties and shared absences. Negative correlations were omitted from the plotted edge matrix, and spring-layout distance is not a calibrated evolutionary or functional distance.

S3 Model-assisted audit: methods and reported results

This section is an edited account of the audit performed by Claude Opus 4.7 extended. The prompt is reproduced below. Repeated activity messages and unsupported extrapolations from the original response have been omitted. The numerical results remain those reported by the auditing model; the cell-level evaluation has not been rerun for this preprint. Agreement with the model-generated reference judgments should not be equated with independently adjudicated biological accuracy.

S3.1 Original audit prompt

> I have a large table of animal, plant, and fungus characteristics that I've obtained using Grok's batch API. I now want to get an idea of how accurate these characteristics are. Each characteristic is in the form of a question, and the possible response options are:
> 0 = Definitely No
> 1 = Probably No
> 2 = Unknown (or not applicable)
> 3 = Probably Yes
> 4 = Definitely Yes
> I've selected a random sample of 200 of these organisms (out of over 35,000) and am attaching their data. I would like you to do the absolute best you can, given your existing knowledge and all tools available to you, to check what percentage of these responses are correct. Use whatever approach you think will yield the most useful approximation of accuracy.

S3.2 Procedures and aggregate results

The audit started with 200 species by 196 questions, totaling 39,200 cells. For the taxonomy-rule comparison, the auditing model assigned broad organism groups and selected approximately 50 questions for which it expected broad taxonomic rules to be informative. It compared codes 0/1 with expected No and codes 3/4 with expected Yes. The model revised some rules and taxonomic assignments after examining disagreements, so the reference rules were not fixed independently of the observed data. This can correct mistaken reference judgments, but it can also make the resulting agreement optimistic.

For the species-specific check, the model selected 20 organisms and 12 questions per organism and used its knowledge and targeted searches where it lacked confidence. The report names sources such as IUCN, Animal Diversity Web, BirdLife, taxonomic literature, and Wikipedia, but does not supply a complete cell-by-cell source and adjudication ledger in the manuscript. It is therefore more accurate to call this a model-assisted spot check than a fully independently web-verified evaluation of all 240 cells.

Procedure	Reported cells	Reported directional agreement
Selected taxonomy-rule comparison	7,641 in final summary	96.5%
Selected species-specific spot check	240	Approximately 237/240 (98.8%)

The rule-based rate is reported as 96.5% for a final total of 7,641 evaluable cells (about 19.5% of the 39,200-cell matrix). The spot-check result is approximately 237/240, or 98.8%. The questions were selected, the smaller species sample was purposive, and the reference labels were fallible model judgments. Neither result can be extrapolated to the full LifeDive matrix. No

population-accuracy interval is presented because a simple binomial interval would ignore selection, repeated cells within species and questions, reference errors, and unresolved preprocessing provenance.

S3.3 Response distribution and denominator discrepancy

Code	Meaning	Count	%
0	Definitely No	24,607	62.8%
1	Probably No	1,839	4.7%
2	Unknown / N/A	0	0.0%
3	Probably Yes	3,332	8.5%
4	Definitely Yes	9,422	24.0%

The response counts sum to 39,200. Extreme substantive codes (0 and 4) account for 34,029 cells (86.8%), and intermediate substantive codes (1 and 3) for 5,171 (13.2%). These percentages describe the entire audited matrix, not the smaller rule-evaluable subset. The lack of code 2 in this matrix cannot establish the raw annotation model's abstention rate. Preprocessing and imputation can remove Unknown responses, and the audit report does not resolve the provenance of each examined cell.

Within the separate taxonomy-rule subset, the original report gives 97.1% agreement for 7,440 extreme-code cells and 73.4% for 203 intermediate-code cells. These counts total 7,643, whereas the final total is 7,641; an intermediate stage of the transcript also used 7,643. The discrepancy cannot be resolved from the summary. These strata are retained here as approximate reported outputs, not as calibrated probabilities or as a basis for weighting the entire dataset. Even reconciled category differences would show confidence ordering, not full probabilistic calibration.

S3.4 Reported subgroup and question summaries

The following subgroup summaries are reproduced from the audit. The groups shown are selected groups, not an exhaustive partition of all 200 species. These are the auditing model's reported group assignments and counts; group membership was not reconciled with the taxonomy in Table S1. Differences should not be interpreted as precise comparisons of population accuracy.

Group	Species	Evaluated cells	Reported agreement
Bony fish	28	1,288	97.0%
Insects	26	988	98.2%
Angiosperms	30	810	93.6%
Birds	14	700	97.0%
Reptiles	14	616	96.4%
Mammals	11	517	97.7%

Group	Species	Evaluated cells	Reported agreement
Arachnids	10	410	95.4%
Fungi	10	340	96.5%
Amphibians	6	270	96.3%
Gastropod mollusks	7	259	98.5%
Cnidarians	6	234	95.3%
Crustaceans	4	144	99.3%

The report identified the following lower-agreement questions. Their wording is abbreviated here; identifiers are those used in the audit and are not the row order of the sorted pattern table. Because taxonomic rules may omit species exceptions or qualifiers in the actual question, disagreement can reflect a mistaken reference rule as well as an annotation error.

Audit Q	Question (abbreviated)	Reported agreement
190	Structures commonly preserved in fossil record	78.5%
75	Has scales or squama	83.3%
53	Alternation of multicellular haploid/diploid generations	88.7%
26	Can survive long periods underwater	89.2%
157	Primarily aquatic	90.4%
195	Specialized gas-exchange pores (stomata/spiracles/lenticels)	91.8%
165	Uses hemoglobin or hemoglobin-like molecules	93.5%

S3.5 Reference errors and implications

The original report itself documents errors in the auditor's organism classification and notes that some initial rules were too rigid. In addition, its summary rule for reproduction omitted the phrase "eggs (outside the body)" from the actual item, which matters for interpreting mammalian responses. A cycad was also invoked in a discussion of angiosperm errors even though it is a gymnosperm. These examples show why model-generated reference standards need explicit, independently checked definitions.

The audit supports a limited conclusion: high agreement was reported for particular selected cells, and some ambiguous questions and questionable annotations were identified for review. It does not establish overall 95-97% accuracy, the number of correct entries in the full dataset, or the effectiveness of re-prompting. A stronger evaluation would use a prespecified sample across species and questions, separate generated from imputed values, retain source citations and explicit negative evidence, adjudicate disagreements with relevant expertise, and report uncertainty that respects the sampling and clustering structure.

Table S1. Species in the 200-species subset

Scientific names and higher taxonomy are retained from the original analysis. The table contains 200 distinct species; Other in Figure 2 is a plotting category rather than a taxonomic rank.

Organism	Phylum	Class	Order
Abelmoschus caillei (West African okra)	Tracheophyta	Magnoliopsida	Malvales
Acanthosepion ellipticum (Ovalbone cuttlefish)	Mollusca	Cephalopoda	Sepiida
Acinipe paulinoi (Paulino's Stone Grasshopper)	Arthropoda	Insecta	Orthoptera
Acricoactis brachyacontis	Cnidaria	Anthozoa	Actiniaria
Alcyonidium diaphanum (sea chervil)	Bryozoa	Gymnolaemata	Ctenostomatida
Alepida picticollis (Canegrub)	Arthropoda	Insecta	Coleoptera
Alveopora allingi (net coral)	Cnidaria	Anthozoa	Scleractinia
Amphilochus marionis	Arthropoda	Malacostraca	Amphipoda
Ancylopsetta dendritica (threespot sand flounder)	Chordata	Actinopterygii	Pleuronectiformes
Anolis incredulus (Turquino Emerald Anole)	Chordata	Reptilia	Squamata
Anorrhinus galeritus (Bushy-crested Hornbill)	Chordata	Aves	Bucerotiformes
Antaioserpens warro (Warrego Burrowing Snake)	Chordata	Reptilia	Squamata
Aquilolamna milarcae	Chordata	Chondrichthyes	Unknown
Arabidopsis lyrata (lyrate rockcress)	Tracheophyta	Magnoliopsida	Brassicales
Architectonica maxima (giant sundial)	Mollusca	Gastropoda	Heterostropha
Aristolochia thwaitesii (Seaside Dutchman's Pipe)	Tracheophyta	Magnoliopsida	Piperales
Artocarpus heterophyllus (jackfruit)	Tracheophyta	Magnoliopsida	Rosales
Asparagus setaceus (Asparagus Fern)	Tracheophyta	Liliopsida	Asparagales
Asternolaelaps fecundus	Arthropoda	Arachnida	Mesostigmata
Asterolecanium sanbernardense (Sanbernard pit scale)	Arthropoda	Insecta	Hemiptera
Astragalus purshii (Pursh milk-vetch)	Tracheophyta	Magnoliopsida	Fabales
Atherurus africanus (African Brush-tailed Porcupine)	Chordata	Mammalia	Rodentia
Atractus indistinctus (Indistinct Ground Snake)	Chordata	Reptilia	Squamata
Atriplex gardneri (Gardner's orache)	Tracheophyta	Magnoliopsida	Caryophyllales
Austrosynthemis cyanitincta (Turquoise Tigertail)	Arthropoda	Insecta	Odonata
Banksia oligantha (Wagin Banksia)	Tracheophyta	Magnoliopsida	Proteales
Barbatosphaeria lignicola	Ascomycota	Sordariomycetes	Unknown
Betula sandbergii (Sandberg's birch)	Tracheophyta	Magnoliopsida	Fagales
Bidenichthys capensis (freetail brotula)	Chordata	Actinopterygii	Ophidiiformes
Bidens pilosa (Spanish-needles)	Tracheophyta	Magnoliopsida	Asterales
Bipassalozetes macedonicus	Arthropoda	Arachnida	Sarcoptiformes
Bodianus busellatus (Blackspot Hogfish)	Chordata	Actinopterygii	Labriformes
Bolbomyia nana	Arthropoda	Insecta	Diptera
Cacosternum leleupi (Katanga Caco)	Chordata	Amphibia	Anura

Organism	Phylum	Class	Order
Calcarisporiella thermophila	Calcarisporiellomycota	Calcarisporiellomycetes	Calcarisporiellales
Callionymus grossi (Gross's dragonet)	Chordata	Actinopterygii	Syngnathiformes
Callulops robustus (Boulenger's Callulops Frog)	Chordata	Amphibia	Anura
Cambarunio simus (Caney fork rainbow)	Mollusca	Bivalvia	Unionida
Camponotus pennsylvanicus (black carpenter ant)	Arthropoda	Insecta	Hymenoptera
Cantorchilus leucotis (Buff-breasted Wren)	Chordata	Aves	Passeriformes
Capheira mollis (gentle sea cucumber)	Echinodermata	Holothuroidea	Apodida
Carex lacustris (common lake sedge)	Tracheophyta	Liliopsida	Poales
Chasmistes fecundus (Webug Sucker)	Chordata	Actinopterygii	Cypriniformes
Cingulochitina chacoparanaense	Unknown	Chitinozoa	Operculatifera
Cinnamomum camphora (camphor tree)	Tracheophyta	Magnoliopsida	Laurales
Claytonia megarhiza (alpine spring-beauty)	Tracheophyta	Magnoliopsida	Caryophyllales
Claytonia multiscapa (Rydberg's spring-beauty)	Tracheophyta	Magnoliopsida	Caryophyllales
Colias croceus (Clouded Yellow)	Arthropoda	Insecta	Lepidoptera
Colocasia propinquilinea (Close-banded Demas)	Arthropoda	Insecta	Lepidoptera
Cornuvesica acuminata	Ascomycota	Sordariomycetes	Microascales
Corvus leucognaphalus (White-necked Crow)	Chordata	Aves	Passeriformes
Cranopygia brindlei	Arthropoda	Insecta	Dermaptera
Craugastor amniscola (Rivulet Rainfrog)	Chordata	Amphibia	Anura
Crocidura polia (Fuscous Shrew)	Chordata	Mammalia	Eulipotyphla
Cypripedium yatabeanum (Spotted Lady's Slipper)	Tracheophyta	Liliopsida	Asparagales
Cyrtodactylus pinlaungensis (Pinlaung Bent-toed Gecko)	Chordata	Reptilia	Squamata
Dasyornis brachypterus (Eastern Bristlebird)	Chordata	Aves	Passeriformes
Deveximentum ruconius (deep pugnose ponyfish)	Chordata	Actinopterygii	Perciformes
Dimorphodon macronyx (Dimorphodon)	Chordata	Reptilia	Pterosauria
Discias pascuensis	Arthropoda	Malacostraca	Decapoda
Discus catskillensis (angular disc)	Mollusca	Gastropoda	Stylommatophora
Dryomyza amblia	Arthropoda	Insecta	Diptera
Elephantulus fuscus (Dusky Elephant Shrew)	Chordata	Mammalia	Macroscelidea
Eleutherodactylus ricordii (Oriente Yellow-mottled Frog)	Chordata	Amphibia	Anura
Elimia fusiformis (Fusiform Elimia)	Mollusca	Gastropoda	Unknown
Encephalartos gratus (Mulanje Cycad)	Tracheophyta	Cycadopsida	Cycadales
Eoacmaea pustulata (Spotted Limpet)	Mollusca	Gastropoda	Patellogastropoda
Epargyreus clarus (Silver-spotted Skipper)	Arthropoda	Insecta	Lepidoptera
Eucalyptus globoidea (white stringybark)	Tracheophyta	Magnoliopsida	Myrtales
Euconulus fulvus (tawny glass snail)	Mollusca	Gastropoda	Stylommatophora

Organism	Phylum	Class	Order
Euneomys fossor (Burrowing Euneomys)	Chordata	Mammalia	Rodentia
Euparkeria capensis	Chordata	Reptilia	Unknown
Eupatorium mohrii (Mohr's thoroughwort)	Tracheophyta	Magnoliopsida	Asterales
Euphorbia ipecacuanhae (American ipecac)	Tracheophyta	Magnoliopsida	Malpighiales
Flatormenis proxima (northern flatid planthopper)	Arthropoda	Insecta	Hemiptera
Florisuga fusca (Black Jacobin)	Chordata	Aves	Apodiformes
Garrulax cinereifrons (Ashy-headed Laughingthrush)	Chordata	Aves	Passeriformes
Geomonhystera mexiquense	Nematoda	Chromadorea	Monhysterida
Gerres subfasciatus (common silver belly)	Chordata	Actinopterygii	Perciformes
Gersemia rubiformis (red soft coral)	Cnidaria	Anthozoa	Alcyonacea
Gloeochaete wittrockiana	Glaucophyta	Glaucophyceae	Gloeochaetales
Gloioxanthomyces nitidus	Basidiomycota	Agaricomycetes	Agaricales
Glossolepis ramuensis (Ramu Rainbowfish)	Chordata	Actinopterygii	Atheriniformes
Gloydius monticola (Likiang pitviper)	Chordata	Reptilia	Squamata
Gloydius strauchi (Strauch's pitviper)	Chordata	Reptilia	Squamata
Golovatchia magda	Arthropoda	Diplopoda	Chordeumatida
Gracupica contra (Pied Myna)	Chordata	Aves	Passeriformes
Graphiurus angolensis (Angolan African Dormouse)	Chordata	Mammalia	Rodentia
Gryllotalpa vineae (Vineyard Mole-cricket)	Arthropoda	Insecta	Orthoptera
Guibemantis tasifotsy (White-flanked Malagasy Tree Frog)	Chordata	Amphibia	Anura
Gymnoscopelus braueri (Brauer's lanternfish)	Chordata	Actinopterygii	Myctophiformes
Harmeriella terebrans	Bryozoa	Gymnolaemata	Ctenostomatida
Hazeus elati (Eilat Sandgoby)	Chordata	Actinopterygii	Gobiiformes
Hemidactylus montanus (Mountain Leaf-toed Gecko)	Chordata	Reptilia	Squamata
Hemipteroseius ageneius	Arthropoda	Arachnida	Mesostigmata
Hippopus hippopus (bear paw clam)	Mollusca	Bivalvia	Cardiida
Homalothecium nevadense (Nevada homalothecium moss)	Bryophyta	Bryopsida	Hypnales
Hydrocina chaetocladia	Ascomycota	Leotiomycetes	Helotiales
Hydroperla phormidia (Brownwater Springfly)	Arthropoda	Insecta	Plecoptera
Hydrostachys multifida	Tracheophyta	Magnoliopsida	Cornales
Hypochrysops ignita (Fiery Jewel)	Arthropoda	Insecta	Lepidoptera
Hyracotherium leporinum	Chordata	Mammalia	Perissodactyla
Iris ensata (Japanese iris)	Tracheophyta	Liliopsida	Asparagales
Isophya ciucasi (Ciucas Plump Bush-cricket)	Arthropoda	Insecta	Orthoptera
Ixodes pacificus (California black-legged tick)	Arthropoda	Arachnida	Ixodida
Japigny kirschbaum (Kirschbaum knifefish)	Chordata	Actinopterygii	Gymnotiformes

Organism	Phylum	Class	Order
Kannemeyeria simocephalus	Chordata	Synapsida	Therapsida
Labropsis xanthura (yellowtail tubelip)	Chordata	Actinopterygii	Labriformes
Lepidogalaxias salamandroides (mud minnow)	Chordata	Actinopterygii	Lepidogalaxiiformes
Lepomis marginatus (Dollar Sunfish)	Chordata	Actinopterygii	Centrarchiformes
Limatula similaris (furrowless fileclam)	Mollusca	Bivalvia	Limida
Lophonotacarus minutus	Arthropoda	Arachnida	Mesostigmata
Lorentziella imbricata (imbricate lorentziella moss)	Bryophyta	Bryopsida	Funariales
Loricaria cuffyi (Cuffy's Whiptail Pleco)	Chordata	Actinopterygii	Siluriformes
Loxopholis hexalepis (Six-scaled Tegu)	Chordata	Reptilia	Squamata
Lupinus obtusilobus (Blunt-lobe Lupine)	Tracheophyta	Magnoliopsida	Fabales
Mariplanella frisia	Platyhelminthes	Rhabditophora	Rhabdocoela
Melanconiella syzygii	Ascomycota	Sordariomycetes	Diaporthales
Melomys leucogaster (White-bellied Melomys)	Chordata	Mammalia	Rodentia
Metapenaeopsis barbata (red rice prawn)	Arthropoda	Malacostraca	Decapoda
Microthambema tenuis	Arthropoda	Malacostraca	Isopoda
Mustela putorius (European Polecat)	Chordata	Mammalia	Carnivora
Myrmothera fulviventris (White-lored Antpitta)	Chordata	Aves	Passeriformes
Myxodagnus opercularis (dart stargazer)	Chordata	Actinopterygii	Blenniiformes
Nactus simakal (Dauan Island Gecko)	Chordata	Reptilia	Squamata
Neophema elegans (Elegant Parrot)	Chordata	Aves	Psittaciformes
Nolina hibernica (Green Beargrass Tree)	Tracheophyta	Liliopsida	Asparagales
Nuchequula flavaxilla (Yellowspotted Ponyfish)	Chordata	Actinopterygii	Perciformes
Obelia geniculata (bell hydroid)	Cnidaria	Hydrozoa	Leptothecata
Oligodon melanozonatus (Abor Hills Kukri Snake)	Chordata	Reptilia	Squamata
Oncorhynchus aguabonita (California golden trout)	Chordata	Actinopterygii	Salmoniformes
Ortalis guttata (Speckled Chachalaca)	Chordata	Aves	Galliformes
Oscarella carmela (slime membrane sponge)	Porifera	Homoscleromorpha	Homosclerophorida
Paeonia californica (California peony)	Tracheophyta	Magnoliopsida	Saxifragales
Paramya subovata (subovate softshell)	Mollusca	Bivalvia	Myida
Parasphaerolaimus stagnalis	Nematoda	Chromadorea	Monhysterida
Paraspinibarbus macracanthus (Pocket-like lip barbel)	Chordata	Actinopterygii	Cypriniformes
Paulownia fortunei (dragontree)	Tracheophyta	Magnoliopsida	Lamiales
Penicillago nodositata	Ascomycota	Eurotiomycetes	Eurotiales
Percina squamata (olive darter)	Chordata	Actinopterygii	Perciformes
Petinomys mindanensis (Mindanao Flying Squirrel)	Chordata	Mammalia	Rodentia
Phaeolepiota aurea (Golden Bootleg)	Basidiomycota	Agaricomycetes	Agaricales
Phegopteris excelsior (tall beech fern)	Tracheophyta	Polypodiopsida	Polypodiales

Organism	Phylum	Class	Order
Phelliactis robusta (robust sea anemone)	Cnidaria	Anthozoa	Actiniaria
Phoradendron rubrum (narrow-leaf mistletoe)	Tracheophyta	Magnoliopsida	Santalales
Phyllomys dasythrix (Drab Atlantic Tree-rat)	Chordata	Mammalia	Rodentia
Pictichromis porphyrea (Magneta Dottyback)	Chordata	Actinopterygii	Perciformes
Pinus teocote (Aztec Pine)	Tracheophyta	Pinopsida	Pinales
Plakina atka (Atka membrane sponge)	Porifera	Homoscleromorpha	Homosclerophorida
Poecilimon fussii (Fuss' Bright Bush-cricket)	Arthropoda	Insecta	Orthoptera
Pollimyrus adspersus (Adspersus Mormyrid)	Chordata	Actinopterygii	Osteoglossiformes
Polycoccum psorae	Ascomycota	Dothideomycetes	Pleosporales
Pontohedyle kepii	Mollusca	Gastropoda	Acochlidiimorpha
Priolepis agrena (Network Reefgoby)	Chordata	Actinopterygii	Gobiiformes
Propithecus coronatus (Crowned Sifaka)	Chordata	Mammalia	Primates
Protalebrella brasiliensis (Brasilian leafhopper)	Arthropoda	Insecta	Hemiptera
Pseudagrion estesi (Estes's Sprite)	Arthropoda	Insecta	Odonata
Pseudemys peninsularis (Peninsula Cooter)	Chordata	Reptilia	Testudines
Pseudoastrosphaeriella bambusae	Ascomycota	Dothideomycetes	Pleosporales
Pseudochromis springeri (bluestriped dottyback)	Chordata	Actinopterygii	Perciformes
Pseudodactylaria xanthorrhoeae	Ascomycota	Sordariomycetes	Pseudodactylariales
Pseudoschulthessiella cerciata (Mweru Wantipa Bush Grasshopper)	Arthropoda	Insecta	Orthoptera
Pteromonas aculeata	Chlorophyta	Chlorophyceae	Chlamydomonadales
Ptilopsis granti (Southern White-faced Owl)	Chordata	Aves	Strigiformes
Raatzbrockmannia boulardi	Arthropoda	Insecta	Hemiptera
Ramajendas frigidus	Tardigrada	Eutardigrada	Parachela
Ranunculus sardous (hairy buttercup)	Tracheophyta	Magnoliopsida	Ranunculales
Rastrognathia macrostoma	Gnathostomulida	Unknown	Bursovaginoidea
Rumex pallidus (large-tuberculed dock)	Tracheophyta	Magnoliopsida	Caryophyllales
Sargocentron rubrum (redcoat)	Chordata	Actinopterygii	Holocentriformes
Sciaena umbra (brown meagre)	Chordata	Actinopterygii	Perciformes
Scorpaenodes englerti (Englert's Scorpionfish)	Chordata	Actinopterygii	Scorpaeniformes
Secubelmis limpida (aquatic oligochaete worm)	Annelida	Clitellata	Unknown
Serranobatrachus megalops (San Lorenzo Robber Frog)	Chordata	Amphibia	Anura
Silene haussknechtii (Haussknecht's Catchfly)	Tracheophyta	Magnoliopsida	Caryophyllales
Solenodon paradoxus (Haitian solenodon)	Chordata	Mammalia	Eulipotyphla
Spalangia plaumanni	Arthropoda	Insecta	Hymenoptera
Stenocercus rhodomelas (Red-black Whorltail Iguana)	Chordata	Reptilia	Squamata
Stugeta carpenteri (Carpenter's Sapphire)	Arthropoda	Insecta	Lepidoptera
Stylaster venustus (fuchsia lace hydrocoral)	Cnidaria	Hydrozoa	Anthoathecata
Sweltsa palearata (Shenandoah Sallfly)	Arthropoda	Insecta	Plecoptera

Organism	Phylum	Class	Order
Symphurus rhytisma (patchtail tonguefish)	Chordata	Actinopterygii	Pleuronectiformes
Symposiachrus sacerdotum (Flores Monarch)	Chordata	Aves	Passeriformes
Tateshia naomiae	Mollusca	Gastropoda	Neogastropoda
Taxodium distichum (bald cypress)	Tracheophyta	Pinopsida	Cupressales
Tetraconus mocsaryi	Arthropoda	Insecta	Hymenoptera
Teucrium botrys (cut-leaf germander)	Tracheophyta	Magnoliopsida	Lamiales
Thelenota ananas (Pineapple Sea Cucumber)	Echinodermata	Holothuroidea	Synallactida
Tianella daamsae	Arthropoda	Diplopoda	Chordeumatida
Tortula muralis (Wall Screw-moss)	Bryophyta	Bryopsida	Pottiales
Trachemys scripta (red-eared slider)	Chordata	Reptilia	Testudines
Treron pembaensis (Pemba Green Pigeon)	Chordata	Aves	Columbiformes
Trichoribates columbreti	Arthropoda	Arachnida	Sarcoptiformes
Trichouropoda laevis	Arthropoda	Arachnida	Mesostigmata
Troglodytes troglodytes (Winter Wren)	Chordata	Aves	Passeriformes
Urbanus doryssus (White-tailed Longtail)	Arthropoda	Insecta	Lepidoptera
Uropeltis grandis (Smith's Earth Snake)	Chordata	Reptilia	Squamata
Wheenyoides cooki	Arthropoda	Arachnida	Trombidiformes
Xerarionta tryoni (bicolor cactussnail)	Mollusca	Gastropoda	Stylommatophora
Yabea microcarpa (California-hedge-parsley)	Tracheophyta	Magnoliopsida	Apiales

Table S2. Retained questions and displayed pattern coefficients

Promax-rotated 14-factor solution. Questions and available coefficients are retained from the sorted analysis table. Bold coefficients have absolute magnitude at least 0.50. Blanks denote coefficients not displayed in that output, not zero. NR marks three Bird-like coefficients that were unavailable in the retained table; no values have been inferred.

PL = Plant-like; FL = Fungus-like; Aq = Aquatic; In = Insect-like; Sm = Simple; Rb = Robust; Du = Durable; Bd = Bird-like; Dg = Dangerous; Sh = Shelled; Dr = Drought-tolerant; Pa = Parasitic; Co = Cold-tolerant; En = Environmentally-sensitive. Question order is sorted by the source output and does not correspond to the audit Q identifiers.

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism produce cellulose?	**0.97**													
Does this organism exhibit etiolation or similar responses to low light conditions?	**0.96**													
Does this organism have centrioles or centriole like structures in its cells?	**-0.94**													
Do this organism's cells contain large fluid filled vacuoles as a major structural component?	**0.94**													
Does this organism store energy primarily as glycogen?	**-0.93**													
Can this organism detect and rapidly respond to physical touch or mechanical contact?	**-0.92**													
Does this organism form symbiotic relationships with mycorrhizal fungi?	**0.92**													
Is sodium (Na) used in any way in this organism's physiology?	**-0.92**													
Does this organism produce fruit or fruiting bodies?	**0.91**													
Does this organism produce phenolic compounds (e.g., tannins, flavonoids)?	**0.91**													
Can this organism photosynthesize?	**0.91**													
Is boron (B) used in any way in this organism's physiology?	**0.91**													
Do this organism's cells have cell walls?	**0.91**													
Can this organism direct its growth toward or away from chemical or nutrient gradients (chemotropism)?	**0.88**													
Does this organism alter its growth form in response to sustained mechanical stimulation such as wind or physical pressure (thigmomorphogenesis)?	**0.87**													
Does this organism have any noticeable green coloring?	**0.86**													

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism exhibit primarily apical or tip based growth?	**0.86**													
Does this organism require externally acquired vitamins or cofactors it cannot synthesize?	**-0.85**													
Does this organism have motile cells at any point in its life cycle (e.g., sperm, zoospores)?	**-0.85**	-0.32												
Can parts of this organism move fast enough to see in real time without time lapse?	**-0.85**													
Does this organism have separate male and female individuals?	**-0.84**													
Does this organism produce terpenes or terpenoid compounds?	**0.84**													
Does this organism produce oxalate crystals as a normal part of its physiology?	**0.83**													
Is this organism hermaphroditic (having both male and female reproductive organs simultaneously)?	**0.83**	-0.33												
Can this organism detect and respond to vibrations or sound transmitted through air, water, or substrate?	**-0.81**													
Does this organism have a life cycle involving an alternation between distinct multicellular haploid and diploid generations?	**0.81**													
Can this organism transmit electrical signals between its cells?	**-0.80**	-0.34												
Is this organism sessile (fixed in place) for most of its adult life?	**0.80**													
Does this organism produce pheromones or chemical signals to coordinate behavior or development among members of its own species?	**-0.80**													
Does this organism produce compounds that specifically attract insects (e.g., for pollination, spore dispersal, prey capture, or symbiotic benefit)?	**0.79**													
Does this organism produce structures specifically adapted to attract animals for seed or spore dispersal?	**0.78**													
Does this organism primarily store energy as oils or lipids?	**-0.77**													
Does this organism consume other organisms?	**-0.77**													
Does this organism have structures arranged in a spiral or helical pattern (e.g., phyllotaxis, shell coiling, hyphal spiraling)?	**0.75**									0.45				
Does this organism have bilateral symmetry?	**-0.75**													

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism produce melanin or melanin like pigments?	**-0.74**													
Does this organism form perennial structures that persist and regrow across multiple years or seasons?	**0.74**													
Does this organism produce alkaloid compounds?	**0.72**													
Does this organism produce carotenoid pigments?	**0.70**							0.41						
Does this organism produce notable chemical compounds used by humans (medicines, dyes, fragrances, etc.)?	**0.69**								0.32					
Does this organism have camouflage that helps it blend into its surroundings?	**-0.68**													
Does this organism have claws?	**-0.68**		**-0.64**											
Does this organism have a modular body plan (built from repeated, semi-autonomous units)?	**0.67**													
Does this organism have a body plan that includes a pseudocoelom or coelom like cavity?	**-0.67**			-0.41										
Could this organism survive its entire life cycle without sunlight?	**-0.66**													
Does this organism produce β glucan as a significant structural or functional component?	**0.65**	**0.55**												
Does this organism rely on wind, water, or other organisms for reproduction?	**0.65**		**0.58**											
Can this organism survive and regrow from a small severed fragment into a complete new organism?	**0.64**				0.39									
Does this organism produce antibiotic or antimicrobial compounds?	**0.63**													
Does this organism do most of its activities (or growth) at night?	**-0.62**							**NR**						
Can part of this organism survive passage through another organism's digestive tract as a normal part of its dispersal or life cycle?	**0.61**													
Is this organism commonly polyploid (naturally having more than two complete sets of chromosomes)?	**0.60**													
Can this organism grow taller than 10 meters?	**0.59**						**0.50**						-0.31	
Does this organism absorb nutrients primarily by direct uptake across its body surface rather than via an internal digestive cavity?	**0.58**	0.44												

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Is this organism grown or raised intentionally by humans?	**0.56**													
Does this organism have specialized storage organs for energy or nutrient reserves (e.g., tubers, bulbs, fat bodies, sclerotia)?	**-0.56**													
Does this organism contain or produce highly complex molecules used in pharmaceutical research?	**0.55**				0.30				**0.54**					
Is this organism capable of trapping or capturing prey?	**-0.54**				0.35				0.32					
Does this organism build a nest or similar structure for its offspring?	**-0.53**		-0.34	-0.32										
Does this organism produce cyanogenic compounds (capable of releasing hydrogen cyanide)?	**0.52**													
Is silicon (Si) or silica an important structural component of this organism?	**0.51**													
Is this organism capable of asexual reproduction?	0.49				0.43									
Can this organism change its appearance seasonally?	0.47													
Does this organism have any noticeable black coloring?	-0.46							0.37						
Does this organism produce psychoactive or hallucinogenic compounds?	0.46													
Does this organism exhibit radial symmetry?	0.45				0.41									
Is copper (Cu) important in this organism's physiology?	0.45													
Is this organism commonly kept by humans for ornamental or aesthetic purposes?	0.45							0.34						
Can this organism produce complex vocalizations or sounds?	-0.45		-0.42	-0.45										
Can this organism survive in highly acidic soils or waters?	0.44													
When mature, is a significant portion of this organism's body composed of dead cells?	0.43													
Has this organism been significantly altered through selective breeding or domestication?	0.42					0.39								
Can this organism detect gravity and orient its growth or posture accordingly?	0.41			-0.39										
Is being eaten by another organism one of this organism's primary dangers?	-0.40	-0.32										-0.35		

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Can this organism detect and orient its growth or movement toward or away from light?	0.39	-0.39												
Does this organism require a mate (or pollinator partner organism) to reproduce?	-0.39													
Does this organism have arms?	-0.38			-0.38										
Does this organism exhibit mimicry of another species?	-0.36													
Does this organism show significant developmental plasticity (dramatically different growth forms depending on environmental conditions)?	0.33													
Does this organism produce latex, resin, or sticky secretions?	0.29													
Is nickel (Ni) used in any way in this organism's physiology?	0.27													
Does this organism have a notable ecological association with dung or animal waste?	-0.24													
Does this organism produce enzymes capable of degrading lignin?		**0.93**												
Does this organism produce siderophores or specialized iron chelating compounds?		**0.84**												
Does this organism commonly have multinucleate cells as a normal part of its biology?		**0.80**												
Does this organism obtain nutrients primarily by decomposing dead organic matter?		**0.69**		0.44										
Does this organism have multiple distinct tissue types with specialized functions?		**-0.64**			-0.32									
Does this organism have a life stage that is predominantly haploid?		**0.63**												
Does this organism have a filamentous or thread like growth form?		**0.58**												
Does this organism rely on extracellular digestion (secreting enzymes outside its body to break down nutrients before absorbing them)?		**0.57**							0.32					
Is this organism commonly found in close association with decaying wood?		**0.53**	-0.35											
Is this organism always multicellular (never existing as a single cell)?		**-0.53**												
Does this organism produce enzymes capable of degrading cellulose?		0.49		0.35										
Does this organism produce ethanol as a metabolic byproduct?		0.35												

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Can this organism survive long periods under water?			**0.86**											
Is this organism primarily aquatic?			**0.82**											
Does this organism use gills or gill like structures for gas exchange?			**0.82**		-0.31									
Does this organism "shed" dead parts of its mass periodically?		-0.32	**-0.69**											
Can this organism tolerate high levels of salinity in its environment?			**0.65**											-0.42
Can this organism quickly recolonize or recover after fire?			**-0.65**								0.31			
Does this organism continue growing throughout its life (indeterminate growth)?	0.42		**0.63**											
Does this organism generate significant metabolic heat (thermogenesis)?	-0.35		**-0.55**	**-0.55**										
Can this organism change its sex or mating type during its lifetime?			**0.53**											
Is wind important in this organism's life?	0.43		-0.46					0.36						
Is this organism commonly eaten by humans?			0.46		-0.31	0.37								
Can this organism survive in very low oxygen or anaerobic environments?			0.44											
Does this organism have a waxy or waterproof outer coating?			-0.44	0.36										
Does this organism have hair, fur, or trichomes?			-0.44					**NR**						
Does this organism prefer very high altitudes?			-0.38											
Does this organism exhibit developmental responses to changes in day length (photoperiodism)?		-0.32	-0.38											
Can this organism regenerate lost or damaged body parts (e.g., lost limb, lost tail, etc.)?			0.35											
Does this organism exhibit circadian "daily" rhythms in its physiology or activity?			-0.31											
Does this organism have sharp thorns, quills, prickles, spines, or spikes?			0.29											
Does this organism use chitin as a structural material?	-0.44			**0.92**										
Does this organism have a closed circulatory or vascular system?				**-0.83**										

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism have a rigid internal skeleton or structural support system?	-0.31			**-0.73**										
Does this organism use hemoglobin or hemoglobin like molecules in its physiology?	-0.42			**-0.72**										
Does this organism produce chitinase or chitin degrading enzymes?				**0.69**	-0.31									
Does this organism weigh less than 0.1 grams?				**0.65**										
Does this organism typically live less than one year?				**0.64**			-0.40							
Does this organism have specialized pores or openings in its outer surface for gas exchange (e.g., stomata, spiracles, lenticels)?	0.44	-0.35		**0.62**										
Does this organism undergo dramatic morphological changes between life stages?	-0.41			**0.58**										
Does this organism produce vitamin D or its precursors when exposed to UV light?	-0.47	0.46		**-0.54**										
Does this organism produce or secrete mucus, mucilage, or gel like substances?	-0.40		**0.50**	**-0.53**										
Is cobalt (Co) used in any way in this organism's physiology?	-0.31			-0.48										
Does this organism reproduce only once in its lifetime and then die (semelparity or monocarpy)?				0.38			-0.31							
Does this organism enter a dormant state during adverse environmental conditions?				0.32										
Is iron (Fe) used in any way in this organism's physiology?				-0.28										
Can this organism reproduce by budding?					**0.85**									
Is this organism's body (or the majority of its tissue) gelatinous or jelly like?					**0.67**									
Does this organism have intercellular cytoplasmic connections between adjacent cells (e.g., plasmodesmata, gap junctions, septal pores)?					**-0.55**									
Can this organism form a network or colony that shares resources across individuals?		0.41			0.46									
Does this organism form a symbiotic relationship with a photosynthetic organism?					0.37									

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Is this organism commonly considered a weed, crop pest, or agricultural nuisance?						**0.89**								
Is this organism considered "invasive" in any areas?						**0.88**								
Do many humans consider this organism "annoying"?						**0.67**			0.33					
Can this organism easily survive in cities						**0.61**								
Does this organism exhibit rapid population growth under favorable conditions?						**0.54**	-0.39							
Is this organism commonly known to cause allergic reactions in humans?						**0.53**								
Is this organism endemic to a very small geographic area (e.g., a single island, mountain, or lake)?						**-0.50**								0.42
Does this organism live in large colonies or groups of its own species?		0.35				0.45								
Does this organism harbor symbiotic bacteria within its cells or tissues as a normal part of its biology?					0.32	0.44								
Does this organism produce a strong or distinctive smell?						0.28								
Does this organism have tissues that are hard OR durable enough to be sought by humans?							**0.75**			0.34				
Is this organism larger (in total mass) than an adult grizzly bear?	0.46						**0.63**							
Does this organism have protective outer covering (e.g., bark, armor, durable scales, thick exoskeleton, etc.)?	-0.42			**0.53**			**0.61**							
Can this organism live longer than 100 years?	0.44				0.33		**0.61**							
Is this organism smaller (in total mass) than an adult hamster?	-0.31			0.34			**-0.59**							
Does this organism produce structures commonly preserved in the fossil record?							**0.57**							
Might some describe this organism as "fragile"?							-0.49							
Does this organism actively create or significantly modify its physical habitat (ecosystem engineering)?							0.46							
Does this organism (fully grown) have a mass between 150 and 250 kilograms?							0.46							
Is this organism grown raised by humans for its fibers (including keratin fibers, cellulose fibers, collagen fibers, etc.)?	0.38						0.44							

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism exhibit synchronized mass reproduction at irregular multi year intervals (masting)?	0.30						0.41							
Could this organism be described as a "living fossil"?							0.29							
Does this organism produce structural coloration (color generated by physical nanostructure rather than pigments)?	-0.34							**0.70**						
Do many humans consider this organism "beautiful"?								**0.68**						
Are seeds, eggs (outside the body), or spores involved in this organism's reproductive process?	0.41			**0.50**				**0.59**						
Does this organism have any noticeable blue coloring?								**0.56**						
Does this organism ever glide through the air?			-0.42					0.49						
Does this organism have any noticeable orange coloring?								0.45						
Does this organism have internal gas filled cavities or chambers?				-0.34				0.44						
Does this organism have any noticeable yellow coloring?								0.37						
Does this organism have any noticeable red coloring?								0.35						
Considering this organism's total mass, is most of it usually (or always) hidden from view?								**NR**						
Can this organism pose a serious threat to human life?						0.32			**0.73**					
Is this organism venomous?	-0.34				**0.50**				**0.65**					
Does this organism produce defensive toxins or chemicals?	0.37				0.33				**0.59**					
Is this organism poisonous to most mammals if eaten?	0.48								**0.57**					
Might some people call this organism "ferocious"?	-0.43								0.46					
Is calcium carbonate a significant structural material in this organism?	-0.32		0.40				0.35			**0.78**				
Does this organism have tentacles or tentacle like appendages?			0.40							**0.71**				
Considering this organism's length, width, and height, is one of those much greater than the other two?										-0.38				
Does this organism have scales or squama?	-0.34							0.32		-0.34				
Does this organism have any noticeable white or very pale coloring?										0.22				
Can this organism survive long periods without water?			-0.38								**0.75**			

Question	PL	FL	Aq	In	Sm	Rb	Du	Bd	Dg	Sh	Dr	Pa	Co	En
Does this organism thrive in arid or desert like conditions?			-0.36								**0.72**			
Could this organism survive at 50 degrees Celsius for 100 hours or more?											**0.64**			
Can this organism survive being completely dried out and later rehydrate?		0.44									0.45			
Does this organism have specialized structures for storing water?											0.40			
Is this organism parasitic on other living organisms?												**0.76**		
Does this organism depend on a close symbiotic relationship with another species to survive?	0.49											**0.59**		
Does this organism commonly colonize or inhabit the human body?												0.41		
Can this organism produce light (bioluminescence)?												-0.38		
Does this organism live on the surface of another living organism (epiphytic, epibiotic)?					0.31							0.35		
Is magnesium (Mg) used in any way in this organism's physiology?												0.22		
Does this organism produce volatile sulfur compounds?												-0.18		
Is this organism primarily found in tropical regions?				-0.36									**-0.59**	
Is this organism commonly found in polar or Arctic Antarctic regions?													**0.56**	
Can this organism survive long periods in freezing temperatures?													0.47	
Does this organism migrate (or otherwise move) long distances seasonally?													0.45	
Does this organism respond to seasonal temperature changes to trigger a major developmental transition (e.g., vernalization, hibernation, cold-induced fruiting)?													0.31	0.30
Did this become a distinct genus only in the last 1 million years?													-0.17	
Is this organism sensitive to pollution or environmental changes?														**0.76**
Is this organism endangered?						-0.42								**0.64**
Is this organism commonly found specifically in freshwater (non marine) aquatic habitats?			0.42											**0.62**

NR occurs only in the Bird-like column for the nighttime-activity, hair/fur/trichomes, and mostly-hidden-mass questions. The primary-loading counts in Table 2 are source-reported; they cannot be uniquely reconstructed from rounded, selectively displayed coefficients.

Table S3. Selected external-criterion associations

Five largest absolute LifeDive-item correlations for each of 44 selected criteria from a 168-criterion screen. The 220 reported correlations are preserved. The sources represented here are EltonTraits, PanTHERIA, MammalDIET, AmphiBIO, SeaLifeBase, and the taxonomic backbone.

Criterion n is the number of species with a nonmissing criterion, not a reported pair-specific complete-case n. Estimators were selected according to variable type as described in Methods. Selection by absolute magnitude can favor unstable or biologically indirect associations. Asterisks mark AmphiBIO criteria whose coding requires review: unrecorded categories do not establish absence. These entries are not used as affirmative presence/absence validation.

Is animal (Kingdom Animalia) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism produce cellulose?	-.99
Do this organism's cells contain large fluid filled vacuoles as a major structural component?	-.98
Does this organism have centrioles or centriole like structures in its cells?	.98
Do this organism's cells have cell walls?	-.97
Can this organism photosynthesize?	-.97

Is plant (Kingdom Plantae) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism produce cellulose?	.99
Do this organism's cells contain large fluid filled vacuoles as a major structural component?	.98
Does this organism have centrioles or centriole like structures in its cells?	-.98
Can this organism photosynthesize?	.97
Do this organism's cells have cell walls?	.97

Is fungus (Kingdom Fungi) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Can this organism change its sex or mating type during its lifetime?	.81
Is magnesium (Mg) used in any way in this organism's physiology?	-.81
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.80
Does this organism have multiple distinct tissue types with specialized functions?	-.79
Does this organism obtain nutrients primarily by decomposing dead organic matter?	.77

Is chordate (Phylum Chordata) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have a rigid internal skeleton or structural support system?	.93
Does this organism use hemoglobin or hemoglobin like molecules in its physiology?	.91
Does this organism have specialized pores or openings in its outer surface for gas exchange e.g. stomata spiracles lenticels?	-.89
Does this organism use chitin as a structural material?	-.88
Does this organism have a body plan that includes a pseudocoelom or coelom like cavity?	.88

Is arthropod (Phylum Arthropoda) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism use chitin as a structural material?	.97
Does this organism have a closed circulatory or vascular system?	-.95
Does this organism have a rigid internal skeleton or structural support system?	-.95
Does this organism have specialized pores or openings in its outer surface for gas exchange e.g. stomata spiracles lenticels?	.90
Does this organism use hemoglobin or hemoglobin like molecules in its physiology?	-.89

Is mollusk (Phylum Mollusca) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is calcium carbonate a significant structural material in this organism?	.76
Does this organism have a closed circulatory or vascular system?	-.73
Is magnesium (Mg) used in any way in this organism's physiology?	-.71
Does this organism produce vitamin D or its precursors when exposed to UV light?	-.71
Does this organism have a rigid internal skeleton or structural support system?	-.70

Is vascular plant (Phylum Tracheophyta) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism produce cellulose?	.99
Does this organism have centrioles or centriole like structures in its cells?	-.97
Can this organism photosynthesize?	.97
Do this organism's cells contain large fluid filled vacuoles as a major structural component?	.97
Does this organism form symbiotic relationships with mycorrhizal fungi?	.97

Is moss (Phylum Bryophyta) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have a life stage that is predominantly haploid?	.84
Can this organism survive being completely dried out and later rehydrate?	.80
Can this organism photosynthesize?	.80
Does this organism have a life cycle involving an alternation between distinct multicellular haploid and diploid generations?	.80
Can this organism transmit electrical signals between its cells?	-.79

Is red alga (Phylum Rhodophyta) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have a life stage that is predominantly haploid?	.86
Does this organism have a life cycle involving an alternation between distinct multicellular haploid and diploid generations?	.85
Can this organism transmit electrical signals between its cells?	-.85
Does this organism have centrioles or centriole like structures in its cells?	-.84
Do this organism's cells have cell walls?	.82

Is green alga (Phylum Chlorophyta) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is magnesium (Mg) used in any way in this organism's physiology?	-.81
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.77
Does this organism require a mate or pollinator partner organism to reproduce?	-.76
Does this organism have multiple distinct tissue types with specialized functions?	-.76
Does this organism have a life stage that is predominantly haploid?	.76

Is annelid (Phylum Annelida) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is magnesium (Mg) used in any way in this organism's physiology?	-.84
Does this organism produce enzymes capable of degrading lignin?	.74
Is silicon (Si) or silica an important structural component of this organism?	.73
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.72
Does this organism obtain nutrients primarily by decomposing dead organic matter?	.70

Is bird (Class Aves) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism produce pheromones or chemical signals to coordinate behavior or development among members of its own species?	-.90
Does this organism generate significant metabolic heat thermogenesis?	.86
Does this organism build a nest or similar structure for its offspring?	.84
Does this organism continue growing throughout its life indeterminate growth?	-.84
Is wind important in this organism's life?	.83

Is mammal (Class Mammalia) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have hair (fur) or trichomes?	.99
Are seeds (eggs) outside the body or spores involved in this organism's reproductive process?	-.98
Does this organism have scales or squama?	-.82
Does this organism generate significant metabolic heat thermogenesis?	.69
Does this organism have internal gas filled cavities or chambers?	-.68

Is reptile (Class Reptilia) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism generate significant metabolic heat thermogenesis?	-.89
Does this organism have protective outer covering e.g. (bark) armor durable scales thick exoskeleton (etc)?	.83
Does this organism have scales or squama?	.79
Does this organism continue growing throughout its life indeterminate growth?	.78
Can this organism produce complex vocalizations or sounds?	-.75

Is amphibian (Class Amphibia) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have scales or squama?	-.84
Does this organism have claws?	-.83
Does this organism generate significant metabolic heat thermogenesis?	-.83
Does this organism produce antibiotic or antimicrobial compounds?	.80
Does this organism undergo dramatic morphological changes between life stages?	.73

Is fish (any of 4 fish classes) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism use gills or gill like structures for gas exchange?	.96
Does this organism (shed) dead parts of its mass periodically?	-.96
Can this organism survive long periods under water?	.95
Is this organism primarily aquatic?	.93
Does this organism have claws?	-.89

Is cartilaginous fish (Class Elasmobranchii) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Can this organism tolerate high levels of salinity in its environment?	.77
Can this organism survive long periods under water?	.72
Does this organism use gills or gill like structures for gas exchange?	.66
Does this organism have claws?	-.62
Does this organism have specialized pores or openings in its outer surface for gas exchange e.g. stomata spiracles lenticels?	.61

Is insect (Class Insecta) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism use chitin as a structural material?	.97
Does this organism have a rigid internal skeleton or structural support system?	-.95
Does this organism have a closed circulatory or vascular system?	-.94
Does this organism have specialized pores or openings in its outer surface for gas exchange e.g. stomata spiracles lenticels?	.94
Does this organism use hemoglobin or hemoglobin like molecules in its physiology?	-.89

Is arachnid (Class Arachnida) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism rely on extracellular digestion secreting enzymes outside its body to break down nutrients before absorbing them?	.78
Is magnesium (Mg) used in any way in this organism's physiology?	-.75
Does this organism use chitin as a structural material?	.72
Is this organism venomous?	.68
Does this organism have a closed circulatory or vascular system?	-.67

Is crustacean (Class Malacostraca) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism have a closed circulatory or vascular system?	-.74
Is magnesium (Mg) used in any way in this organism's physiology?	-.70
Does this organism use chitin as a structural material?	.70
Does this organism have a rigid internal skeleton or structural support system?	-.68
Is calcium carbonate a significant structural material in this organism?	.65

Is bivalve (Class Bivalvia) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is magnesium (Mg) used in any way in this organism's physiology?	-.75
Is calcium carbonate a significant structural material in this organism?	.69
Does this organism form perennial structures that persist and regrow across multiple years or seasons?	.68
Does this organism produce vitamin D or its precursors when exposed to UV light?	-.68
Does this organism have a closed circulatory or vascular system?	-.67

Is gastropod (Class Gastropoda) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is magnesium (Mg) used in any way in this organism's physiology?	-.79
Does this organism have tentacles or tentacle like appendages?	.74
Is calcium carbonate a significant structural material in this organism?	.68
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.67
Does this organism produce enzymes capable of degrading lignin?	.65

Is eudicot (Class Magnoliopsida) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism produce cellulose?	.96
Can this organism photosynthesize?	.96
Does this organism form symbiotic relationships with mycorrhizal fungi?	.94
Does this organism have centrioles or centriole like structures in its cells?	-.93
Do this organism's cells contain large fluid filled vacuoles as a major structural component?	.93

Is monocot (Class Liliopsida) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Does this organism form perennial structures that persist and regrow across multiple years or seasons?	1.00
Does this organism produce cellulose?	.93
Does this organism have centrioles or centriole like structures in its cells?	-.93
Does this organism produce fruit or fruiting bodies?	.91
Does this organism exhibit etiolation or similar responses to low light conditions?	.91

Is fern (Class Polypodiopsida) GBIF/CoL backbone
Criterion n = 8,427

LifeDive question	r
Is silicon (Si) or silica an important structural component of this organism?	.82
Can this organism photosynthesize?	.82
Does this organism have a life cycle involving an alternation between distinct multicellular haploid and diploid generations?	.82
Does this organism require a mate or pollinator partner organism to reproduce?	-.82
Does this organism produce cellulose?	.81

Amphibian has larval stage AmphiBIO *
Criterion n = 957

LifeDive question	r
Does this organism undergo dramatic morphological changes between life stages?	.90
Does this organism rely on wind water or other organisms for reproduction?	.87
Is this organism commonly found specifically in freshwater non marine aquatic habitats?	.75
Does this organism use gills or gill like structures for gas exchange?	.74
Is this organism primarily aquatic?	.74

* Coding-sensitive AmphiBIO output; see Methods and the note at the start of Table S3.

Amphibian is viviparous AmphiBIO *
Criterion n = 957

LifeDive question	r
Are seeds (eggs) outside the body or spores involved in this organism's reproductive process?	-.95
Can this organism live longer than 100 years?	-.82
Does this organism have tentacles or tentacle like appendages?a	.66
Does this organism have multiple distinct tissue types with specialized functions?	-.56
Is this organism always multicellular never existing as a single cell?	-.56

* Coding-sensitive AmphiBIO output; see Methods and the note at the start of Table S3.

Amphibian eats vertebrates AmphiBIO *
Criterion n = 1,051

LifeDive question	r
Can this organism live longer than 100 years?	-.89
Might some people call this organism ferocious?	.61
Is this organism smaller in total mass than an adult hamster?	-.59
Does this organism produce siderophores or specialized iron chelating compounds?	.51
Is this organism grown raised by humans for its fibers including keratin fibers cellulose fibers collagen fibers (etc)?	.50

* Coding-sensitive AmphiBIO output; see Methods and the note at the start of Table S3.

Bird is nocturnal EltonTraits
Criterion n = 2,283

LifeDive question	r
Does this organism do most of its activities or growth at night?	.92
Does this organism produce enzymes capable of degrading lignin?	-.90
Can this organism change its sex or mating type during its lifetime?	-.83
Does this organism build a nest or similar structure for its offspring?	-.75
Does this organism have camouflage that helps it blend into its surroundings?	.69

Bird is pelagic specialist EltonTraits
Criterion n = 2,283

LifeDive question	r
Can this organism tolerate high levels of salinity in its environment?	.91
Can this organism change its sex or mating type during its lifetime?	-.83
Does this organism produce enzymes capable of degrading lignin?	-.82
Does this organism produce a strong or distinctive smell?	.67
Can this organism live longer than 100 years?	.65

Mammal is nocturnal EltonTraits
Criterion n = 1,464

LifeDive question	r
Does this organism do most of its activities or growth at night?	.75
Does this organism ever glide through the air?	.53
Is this organism smaller in total mass than an adult hamster?	.48
Do many humans consider this organism beautiful?	-.47
Does this organism have claws?	.46

Mammal is diurnal EltonTraits
Criterion n = 1,464

LifeDive question	r
Does this organism do most of its activities or growth at night?	-.73
Does this organism ever glide through the air?	-.57
Does this organism have claws?	-.53
Does this organism have any noticeable blue coloring?	.51
Is this organism commonly found in polar or Arctic Antarctic regions?	.47

Mammal foraging stratum: aerial EltonTraits
Criterion n = 1,464

LifeDive question	r
Is this organism smaller in total mass than an adult hamster?	.74
Does this organism build a nest or similar structure for its offspring?	-.66
Does this organism enter a dormant state during adverse environmental conditions?	.66
Is wind important in this organism's life?	.65
Does this organism do most of its activities or growth at night?	.61

Mammal foraging stratum: arboreal EltonTraits
Criterion n = 1,464

LifeDive question	r
Does this organism have arms?	.46
Is this organism primarily found in tropical regions?	.46
Does this organism typically live less than one year?	-.41
Does this organism ever glide through the air?	.39
Is this organism larger in total mass than an adult grizzly bear?	-.38

Mammal foraging stratum: ground EltonTraits
Criterion n = 1,464

LifeDive question	r
Does this organism ever glide through the air?	-.78
Does this organism build a nest or similar structure for its offspring?	.58
Does this organism have arms?	-.51
Does this organism typically live less than one year?	.50
Is wind important in this organism's life?	-.46

Mammal foraging stratum: marine EltonTraits
Criterion n = 1,464

LifeDive question	r
Can this organism survive long periods under water?	.97
Is this organism primarily aquatic?	.97
Can this organism tolerate high levels of salinity in its environment?	.92
Does this organism have claws?	-.81
Is this organism larger in total mass than an adult grizzly bear?	.79

Mammal eats other mammals EltonTraits
Criterion n = 1,415

LifeDive question	r
Is this organism capable of trapping or capturing prey?	.59
Might some people call this organism ferocious?	.58
Does this organism have any noticeable black coloring?	.55
Can this organism easily survive in cities?	.46
Does this organism have any noticeable yellow coloring?	.46

Mammal diet score: mammals (0-3) MammalDIET
Criterion n = 1,414

LifeDive question	r
Does this organism produce enzymes capable of degrading lignin?	-.90
Can this organism live longer than 100 years?	-.90
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.52
Does this organism have multiple distinct tissue types with specialized functions?	-.51
Is this organism capable of trapping or capturing prey?	.44

Mammal diet score: vertebrates (0-3) MammalDIET
Criterion n = 1,415

LifeDive question	r
Does this organism produce enzymes capable of degrading lignin?	-.95
Does this organism ever glide through the air?	-.40
Is this organism venomous?	.40
Might some people call this organism ferocious?	.38
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.34

Mammal diet score: plants (0-3) MammalDIET
Criterion n = 1,414

LifeDive question	r
Is this organism capable of trapping or capturing prey?	-.37
Does this organism enter a dormant state during adverse environmental conditions?	-.36
Does this organism produce chitinase or chitin degrading enzymes?	-.33
Is this organism venomous?	-.32
Does this organism have claws?	-.30

Mammal trophic level (1-3; MammalDIET) MammalDIET
Criterion n = 1,411

LifeDive question	r
Is this organism capable of trapping or capturing prey?	.64
Is this organism venomous?	.59
Does this organism produce chitinase or chitin degrading enzymes?	.50
Does this organism consume other organisms?	.49
Is this organism smaller in total mass than an adult hamster?	.46

Mammal activity cycle (1-3) PanTHERIA
Criterion n = 535

LifeDive question	r
Does this organism do most of its activities or growth at night?	-.76
Is this organism capable of asexual reproduction?	.53
Does this organism have any noticeable blue coloring?	.52
Does this organism have claws?	-.48
Does this organism live in large colonies or groups of its own species?	.40

Marine invertebrate is sessile SeaLifeBase
Criterion n = 1,792

LifeDive question	r
Did this become a distinct genus only in the last 1 million years?	-.90
Is this organism sessile fixed in place for most of its adult life?	.84
Does this organism have intercellular cytoplasmic connections between adjacent cells e.g. plasmodesmata gap junctions septal pores?	-.78
Can this organism transmit electrical signals between its cells?	-.77
Is this organism capable of asexual reproduction?	.76

Marine invertebrate is host-associated SeaLifeBase
Criterion n = 1,792

LifeDive question	r
Is this organism parasitic on other living organisms?	.89
Did this become a distinct genus only in the last 1 million years?	-.82
Can part of this organism survive passage through another organism's digestive tract as a normal part of its dispersal or life cycle?	.70
Does this organism depend on a close symbiotic relationship with another species to survive?	.68
Does this organism have tentacles or tentacle like appendages?	-.58