

A language model generates seven million phenotype estimates across the tree of life

One-Sentence Summary: Embedded in a validated pipeline, a general-purpose language model produces a dense cross-kingdom organism-by-phenotype matrix for 35,856 species—work once requiring person-years of expert curation—for roughly US\$10,000.

Tyler M. Moore^{1,2*}

¹Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

²Lifespan Brain Institute (LiBI), Children's Hospital of Philadelphia and Penn Medicine, Philadelphia, PA 19104, USA.

*Corresponding author. Email: tymoore@pennmedicine.upenn.edu

Abstract

Dense, standardized trait matrices underpin comparative biology, macroecology, and conservation, yet assembling them is slow and taxonomically biased: expert curation scales poorly, and text mining cannot recover traits that source texts never state. I show that a general-purpose, web-search-enabled large language model, embedded in a scripted pipeline with a fixed ordinal scale, constrained output, quality control, and imputation, can act as a structured phenotype annotator. The pipeline rated 35,856 animal, plant, and fungal species on 196 cross-kingdom questions, yielding LifeDive—a matrix of more than seven million organism-by-phenotype values generated by a single investigator for roughly US\$10,000. Factor analysis recovers fourteen interpretable dimensions, items correlate strongly with independent expert databases, and an audit estimates ~95–97% directional accuracy, reconstructing major taxonomic groupings.

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 (1, 2, 3).

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 (4, 5, 6, 7). 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.

Automated text mining can accelerate trait mobilization by converting floras, monographs, web pages, and articles into structured records (8, 9, 10), and recent large language model (LLM) systems extract ecological records, plant traits, and species interactions from unstructured text with encouraging precision

(11, 12, 13). But extraction is bounded by source access, wording, controlled vocabularies, and the fact that many useful traits are not stated in any available description. A pipeline that reads text cannot recover a trait the text never states, and its coverage inherits the taxonomic, geographic, linguistic, and publication biases of the accessible corpus.

A complementary strategy is to query an LLM as a structured phenotype annotator rather than solely as a reader of a fixed corpus. Pretrained models encode substantial factual and relational information, but they also produce fluent errors and are sensitive to prompt wording, training distribution, and retrieval context (14, 15). The relevant question is therefore not whether any single generated value should be treated as an observation, but whether a constrained, versioned, and validated pipeline can convert model knowledge into an analyzable data product. Prior work indicates that model-generated quantitative data can recover multivariate biological or behavioral structure even when individual estimates are noisy (16, 17), and studies of LLM-based extraction emphasize explicit schemas, constrained outputs, parsers, missing-data rules, and validation against trusted sources (18, 19, 20, 21). Under this framing, a data product should be judged by the statistical properties of the whole matrix—whether errors are sparse, whether uncertainty is visible, whether downstream dimensions are interpretable, and whether independent benchmarks recover expected biology—rather than by the literal certainty of any one cell.

This shift has a practical consequence that is easy to understate. Assembling a dense, cross-kingdom trait matrix on the scale described here would, until recently, have required a consortium, years of expert curation, and funding on the order of millions of dollars. Here, the entire resource was produced by a single investigator for roughly US\$10,000 in commercial API fees. If the outputs preserve biological signal, this is not merely a cheaper way to build one dataset; it is a general and newly accessible route to converting the latent knowledge of frontier models into structured, validated scientific data.

I present LifeDive, a cross-kingdom organism-by-phenotype resource generated with a web-search-enabled LLM (Grok) embedded in a scripted batch pipeline. The workflow drew accepted or provisionally accepted species from the Catalogue of Life with a family-first sampling strategy and asked the model to rate 35,856 animals, plants, and fungi on 196 phenotype questions spanning cellular biology, anatomy, physiology, ecology, behavior, life history, reproduction, environmental tolerance, and human salience (Table 1), using a common ordinal scale from Definitely No to Definitely Yes. The resulting matrix contains 7,027,776 values. The design trades some trait-specific precision for coverage and comparability: the same question set applies to trees, birds, fungi, molluscs, and insects, even where conventional databases cover only one clade or one trait family. This resource has two linked aims - to document a reproducible approach for large-scale LLM-based phenotype annotation, and to test whether the resulting matrix preserves biological structure - which I assess by factor-analyzing the full matrix, comparing items against independent expert databases, and auditing a random 200-organism subsample. Detailed procedures, complete quality-control logic, prompts, and the full independent-audit report are provided in the Supplementary Information.

Methods

Study design and organism selection

Species candidates came from the Catalogue of Life checklist (version 2026-04-18 XR) (30). The compilation retained accepted or provisionally accepted records at the species rank, together 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 selected the highest-priority species in every family and then filled the remaining quota, favoring accepted names and species with common names; a fixed random seed was used for tie-breaking. This procedure returned 35,000 species. GPT 5.5 Pro was then used to identify well-known organisms absent from the initial list, producing the final sample of 35,856.

LLM phenotype annotation and quality control

Phenotype annotations were collected with the xAI Grok Batch API (31), using the scripted model grok-4-1-fast-reasoning with web search enabled. Each organism was evaluated on 196 retained phenotype questions using a five-point response scale (32): 0, Definitely No; 1, Probably No; 2, Unknown or not applicable; 3, Probably Yes; and 4, Definitely Yes. Organisms were processed in groups of three, and each prompt required a complete CSV row for every organism. A scripted quality-control pipeline validated requests, parsed responses, aligned columns by question identifier, converted missing, nonnumeric, or unreliable cells to Unknown, and generated retries for failed requests. The resulting values should be interpreted as computational phenotype estimates rather than direct measurements or expert-curated observations.

Data assembly, imputation, and factor analysis

Annotation was conducted in 11 question subsets and merged by organism. Unknown and out-of-range values were treated as missing, and items with very high missingness were excluded, leaving 196 questions. Missing responses - 16.6% of the organism-by-phenotype cells - were imputed item by item with multinomial random forests implemented in H2O (33, 34), using observed values for the target item and all other retained items as predictors under a fixed seed. The completed matrix was used for latent-structure analyses; the original, unimputed response categories remain the more appropriate source for cell-level interpretation.

For factor analysis, the response scale was recoded to four ordered levels after removal of the Unknown category: 0 and 1 retained their values, 3 was recoded to 2, and 4 was recoded to 3. A 14-factor item-response-theory factor model was estimated in R with `psych::irt.fa` and `promax` rotation (35-38). Fourteen factors were retained for interpretability; parallel analysis and the minimum-average-partial criterion suggested substantially higher-dimensional solutions, as described in the Supplementary Methods (39, 40). Unit-weighted sum scores based on the loading patterns were used in place of theta scores because a small number of highly discriminating items otherwise dominated the score levels. The signs of the Parasitic and Cold-tolerant dimensions were reversed so that higher values indicated more of the named construct. Scores used in Figure 1 were winsorized and standardized; a small random jitter was added only for plotting. Visualizations were created with `ggplot2`, `ggrepel`, and `patchwork` (41-43).

External benchmark validation

LifeDive was compared with 168 numeric criteria drawn from nine public, peer-reviewed trait databases and a standardized taxonomic backbone. These sources included EltonTraits for bird and mammal traits (5), PanTHERIA (4), MammalDIET (22), AmphiBIO (23), SeaLifeBase (24), FishBase (25), and the traitdata integration of the GBIF Backbone Taxonomy (26, 27). Organisms were matched by genus-species binomial; categorical traits were expanded to indicators while preserving missingness outside each source. Pairwise associations were estimated with Pearson, biserial, polyserial, tetrachoric, or polychoric correlations as appropriate to the variable types, using the `psych` package and a minimum overlap of 30 organisms (36, 44). For each criterion, the five LifeDive items with the largest absolute correlations were retained for inspection.

Independent audit and phenotype-similarity network

A random sample of 200 organisms was subjected to deeper validation. Claude Opus 4.7 extended, which was not used to generate LifeDive, constructed a taxonomy-based reference standard for approximately 50 questions whose answers were largely determined by broad group membership, yielding 7,641 evaluable cells. It also web-verified 240 species-specific cells spanning 20 organisms and 12 questions. The complete prompt, report, and caveats are reproduced in Note S1.

For the same 200 organisms, pairwise phenotype similarity was estimated with polychoric correlations across the 196 ordinal questions (36, 44). Negative correlations were retained for numerical interpretation

but set to zero for visualization. The nonnegative weighted matrix was plotted with qgraph using a force-directed spring layout (45, 46), and nodes were grouped as Arthropoda, Ascomycota, Chordata, Cnidaria, Mollusca, Tracheophyta, or Other. The network is therefore a visualization of positive phenotype similarity, not a phylogenetic tree.

Full request construction, API handling, parsing rules, software details, and validation procedures are provided in the Supplementary Methods.

Results

A cross-kingdom phenotype matrix

The workflow produced a matrix of 35,856 organisms scored on 196 retained questions, for 7,027,776 organism-by-phenotype values. The sampling design favored taxonomic breadth by first ensuring family-level representation and then filling remaining slots with high-priority taxa, preferentially those with common names. The questions cover anatomical, ecological, physiological, behavioral, life-history, and human-salience features across plants, animals, and fungi (Table 1). They include kingdom-level properties such as photosynthesis and cell-wall composition; animal body-plan variables such as gills, shells, limbs, and venom; ecological variables such as aquatic habitat, parasitism, and urban tolerance; and broader attributes such as human use, perceived beauty, fragility, and environmental sensitivity. Unknown or out-of-scale responses were treated as missing (16.6% of values), and random-forest imputation from the remaining phenotype profile produced a complete matrix for latent-structure analysis.

Table 1. Representative phenotype questions. Eleven of the 196 questions, spanning the trait domains covered by LifeDive. Every question was answered on the same ordinal scale: 0, Definitely No; 1, Probably No; 2, Unknown or not applicable; 3, Probably Yes; 4, Definitely Yes. The complete question list is given in Supplementary Table S2.

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?

Fourteen interpretable phenotype dimensions

A 14-factor promax-rotated solution was readily interpretable (Table 2; full pattern matrix in Supplementary Table S2). Assigning each item to its largest absolute loading yielded 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. Plant-like was the broadest dimension, with high positive loadings for cellulose, low-light responses, large vacuoles, mycorrhizal symbiosis, photosynthesis, cell walls, boron use, and plant secondary chemistry, and strong negative loadings for centrioles, glycogen storage, rapid mechanical response, sodium physiology, externally acquired cofactors, and motile cells—separating plant-like profiles from animal-like physiology.

Fungus-like captured a distinct nonanimal syndrome—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. 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. The narrower factors captured coherent clusters: Robust (weed or crop-pest status, invasiveness, urban survival, rapid population growth); Durable (hard tissues sought by humans, large mass, protective covering, longevity, fossil-preserving structures); Bird-like (structural coloration, perceived beauty, external eggs or seeds, blue and orange coloration, gliding); Dangerous (threat to human life, venom, defensive toxins, ferocity); Shelled (calcium carbonate, tentacles); Drought-tolerant (aridity, heat, desiccation tolerance); Parasitic (parasitism, host association); Cold-tolerant (polar occurrence, freezing tolerance, seasonal movement); and Environmentally-sensitive (pollution sensitivity, endangered status, freshwater association). These labels are empirical summaries of covariance among questions, not mutually exclusive categories: an organism can be simultaneously aquatic and dangerous, or plant-like and durable, and the score space retains these combinations.

Table 2. Summary of the fourteen LifeDive phenotype dimensions. Primary-loading counts are from the 14-factor promax-rotated item-factor solution. Loadings in parentheses are from the pattern matrix, rounded to two decimals; the full matrix is in Supplementary Table S2. Negative signs indicate inverse loading on the named dimension and are shown where they help define the contrast.

Dimension	Items (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 / 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 / internal skeleton / hemoglobin	Small arthropod-like or invertebrate body-plan syndrome.
Simple	5	+ budding (.85); + gelatinous tissue (.67); + colony / network resource sharing (.46); – intercellular cytoplasmic connections (–.55)	Simple, colonial, gelatinous, or asexually reproducing forms.
Robust	10	+ weed / 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 / durable tissues sought by humans (.75); + very large mass (.63); + protective covering (.61); + lifespan >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 / spores (.59); + blue coloring (.56); + gliding (.49)	Visually conspicuous, aerial / oviparous, and bird-associated traits.

Dimension	Items (n)	Defining LifeDive items (examples; loading)	Interpretation
Dangerous	5	+ serious threat to human life (.73); + venom (.65); + defensive toxins (.59); + poisonous to mammals (.57); + ferocity (.47)	Human-danger, venom, poison, and defensive / predatory hazard.
Shelled	5	+ calcium carbonate structure (.78); + tentacles (.71); + spiral / helical structures (.45); – elongated body form (–.38); – scales (–.34)	Shelled / calcified and tentaculate invertebrate-like structures.
Drought-tolerant	5	+ long periods without water (.75); + arid / desert thriving (.72); + 50 °C survival (.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 / seasonal physiological responses.
Environmentally-sensitive	3	+ pollution / environmental sensitivity (.76); + endangered status (.64); + freshwater habitat (.62)	Environmental sensitivity, conservation concern, and freshwater association.

Placement of familiar organisms in phenotype space

Selected factor-score plots gave a qualitative check on the labels (Figure 1). In Durable-versus-Plant-like space, woody plants such as English oak and coast redwood occupied the high-Plant-like, high-Durable region, whereas nonwoody plants such as sunflower, dandelion, water-lily, bracken fern, and haircap moss remained high on Plant-like but lower on Durable. Most animals had negative Plant-like scores; large or structurally robust taxa (American alligator, African bush elephant, American lobster, polar bear, green sea turtle, humpback whale, staghorn coral) scored higher on Durable than small or delicate ones (moon jellyfish, housefly, monarch butterfly, European squid). Fungi occupied an intermediate region—positive on Plant-like through shared nonanimal properties, but low on Durable. In Aquatic-versus-Dangerous space, fish and marine invertebrates scored high on Aquatic, venomous or predatory taxa scored high on Dangerous, and the great white shark occupied the joint high-Aquatic, high-Dangerous corner, while lower-danger aquatic taxa such as bottlenose dolphin and humpback whale separated from the Dangerous cluster. Because the dimensions were named before these organisms were plotted, the placements test the scores rather than restate them.



Figure 1. Selected organisms plotted on LifeDive factor scores. Axes show standardized, winsorized factor scores for selected common organisms; a small random jitter was added for plotting only. Top, Durable versus Plant-like; bottom, Aquatic versus Dangerous. The dashed line in each panel is a $y = x$ reference, not a fitted model. Because the dimensions were labeled before these organisms were plotted, their placements provide a qualitative check on the scores.

Agreement with independent expert databases

To test the matrix against independently curated data, I compiled a benchmark of 168 numeric criteria from nine public trait databases anchored to a GBIF/Catalogue of Life taxonomic backbone, drawing bird and mammal diet, foraging, mass, and activity from EltonTraits (5); mammal life-history and ecological

traits from PanTHERIA (4) and detailed diet from MammalDIET (22); amphibian ecology from AmphiBIO (23); marine-invertebrate habitat from SeaLifeBase (24); fish ecology from FishBase (25); and standardized higher taxonomy from a trait-data table integrating the GBIF Backbone (26, 27). Of the 35,856 organisms, 15,525 (43.3%) matched at least one database, with coverage strongest for vertebrates. Table 3 condenses the results by domain; the full top-five item correlations for 44 criteria appear in Supplementary Table S3.

The strongest relationships were near-ceiling and biologically direct. Animalia was negatively associated with cellulose production ($r = -0.99$), cell walls ($r = -0.97$), and photosynthesis ($r = -0.97$), and positively with centrioles ($r = 0.98$); Plantae showed the mirror pattern; Mammalia was associated with hair, fur, or trichomes ($r = 0.99$) and negatively with external eggs or spores ($r = -0.98$). Arthropoda and Insecta were associated with chitin (both $r = 0.97$), and fish-class indicators with gill-like gas exchange ($r = 0.96$), underwater survival ($r = 0.95$), and aquatic lifestyle ($r = 0.93$). More specific ecological criteria also recovered face-valid matches: amphibian larval stage with metamorphosis ($r = 0.90$) and water-based reproduction ($r = 0.87$); bird nocturnality with nighttime activity ($r = 0.92$); marine-mammal foraging with underwater survival ($r = 0.97$) and salinity tolerance ($r = 0.92$); and marine-invertebrate host association with parasitism ($r = 0.89$). No external resource has uniform cross-kingdom coverage, so this benchmark is imperfect; nonetheless, the strength of these correlations across thousands of organisms indicates that the generated matrix preserved major taxonomic, cellular, and ecological contrasts.

Table 3. External validation against independent trait resources, by domain. For each criterion, best $|r|$ is the largest absolute correlation between that criterion and any LifeDive item among the retained top-five matches; the table reports the strongest such value in each domain and the median of the criterion-level best values. Representative matches were chosen for interpretability and are not always the numerical maximum. Correlation signs follow the coding of the external criterion and the LifeDive item. Full item-level results for 44 criteria are in Supplementary Table S3.

Validation domain	External source(s)	Criteria (n)	Organisms with criterion	Best / median $ r $	Representative top matches	Main takeaway
Taxonomy: kingdoms	GBIF / CoL backbone	3	8,427	0.99 / 0.99	Animalia–cellulose (–.99); Plantae–cellulose (.99); Fungi–decomposer nutrition (.77)	Near-ceiling recovery of broad kingdom-level contrasts.
Taxonomy: phyla	GBIF / CoL backbone	8	8,427	0.99 / 0.85	Tracheophyta–cellulose (.99); Arthropoda–chitin (.97); Chordata–internal skeleton (.93); Mollusca–calcium carbonate (.76)	Major body-plan and tissue-composition contrasts recovered across phyla.
Taxonomy: classes	GBIF / CoL backbone	14	8,427	1.00 / 0.86	Mammalia–hair/fur/trichomes (.99); Insecta–chitin (.97); fish–gills (.96); Aves–thermogenesis (.86)	Class-level markers recovered with high correlations, especially for well-defined body-plan traits.
Amphibian ecology / reproduction	AmphiBIO	3	957–1,051	0.95 / 0.90	Larval stage–metamorphosis (.90); viviparity–external eggs/spores (–.95); vertebrate diet–ferocity (.61)	Life-stage and reproductive variables validate strongly; diet variables are more indirect.
Bird ecology	EltonTraits	2	2,283	0.92 / 0.92	Nocturnal–activity at night (.92); pelagic specialist–salinity tolerance (.91)	Direct behavioral and habitat correspondences.
Mammal ecology / foraging	EltonTraits, PanTHERIA	8	535–1,464	0.97 / 0.74	Marine foraging–underwater survival (.97); nocturnality–activity at night (.75); ground foraging–gliding (–.78)	Activity and habitat / foraging strata recovered, with signs depending on source coding.
Mammal diet / trophic level	MammalDIET	4	1,411–1,415	0.95 / 0.77	Trophic level–prey capture (.64); plant diet–prey capture (–.37); vertebrate diet–inverse decomposer/plant signal ($ r $ = .95)	Diet / trophic variables recover predation gradients but are less direct than taxonomic or habitat criteria.
Marine-invertebrate ecology	SeaLifeBase	2	1,792	0.90 / 0.90	Sessility–sessile adult life (.84); host-associated–parasitism (.89); host-associated–symbiosis (.68)	Sessility and host association are represented by direct phenotype items.

Independent audit and phenotype-similarity network

A deeper 200-organism sample provided an independent check on cell-level accuracy. A different model that was not used to generate LifeDive (Claude Opus 4.7) built a taxonomy-based reference standard for questions largely determined by group membership and web-verified a smaller set of species-specific cells. The taxonomy-rule evaluation yielded 96.5% directional accuracy across 7,641 evaluable cells, and the web-verified spot-check yielded roughly 97–99% across 240 cells; the two methods converged on an estimated 95–97% directional accuracy. Crucially, the errors were concentrated where the model signaled doubt: confident answers—Definitely Yes or Definitely No, 86.8% of audited cells—were 97.1% accurate, whereas hedged answers (Probably Yes or Probably No) were 73.4% accurate, and errors clustered in ambiguous or species-specific questions such as fossil-record representation and alternation of generations. The model also abstained rarely, returning no Unknown codes in the audited sample, so its four-level certainty spectrum, rather than an explicit abstention class, carries the usable uncertainty signal. This calibration is practically important: confident cells can be preferred when high-confidence values are needed, and hedged cells can be prioritized for expert review. Extrapolated to the full matrix, this range implies that roughly 6.7–6.8 million values are directionally correct, though the extrapolation is an approximation rather than a guarantee.

The phenotype-similarity network for the same 200 organisms recovered a strong taxonomic and functional signal (Figure 2). Mean within-phylum polychoric correlations were high for Ascomycota ($\rho = 0.877$), Cnidaria ($\rho = 0.853$), Arthropoda ($\rho = 0.823$), Chordata ($\rho = 0.810$), Mollusca ($\rho = 0.785$), and Tracheophyta ($\rho = 0.766$), whereas the deliberately heterogeneous "Other" group was lower ($\rho = 0.510$). Trees clustered with trees, frogs with frogs, flatfish with flatfish, and sea anemones with sea anemones. Boundary cases were informative rather than anomalous: a cuttlefish fell near chordate fishes, consistent with an active, visually oriented, predatory marine profile, and a small terrestrial snail fell near arthropods, sharing a small, shelled, terrestrial-invertebrate syndrome. Negative correlations were weak and were omitted before plotting, so the layout should be read as positive phenotype similarity, not phylogeny.

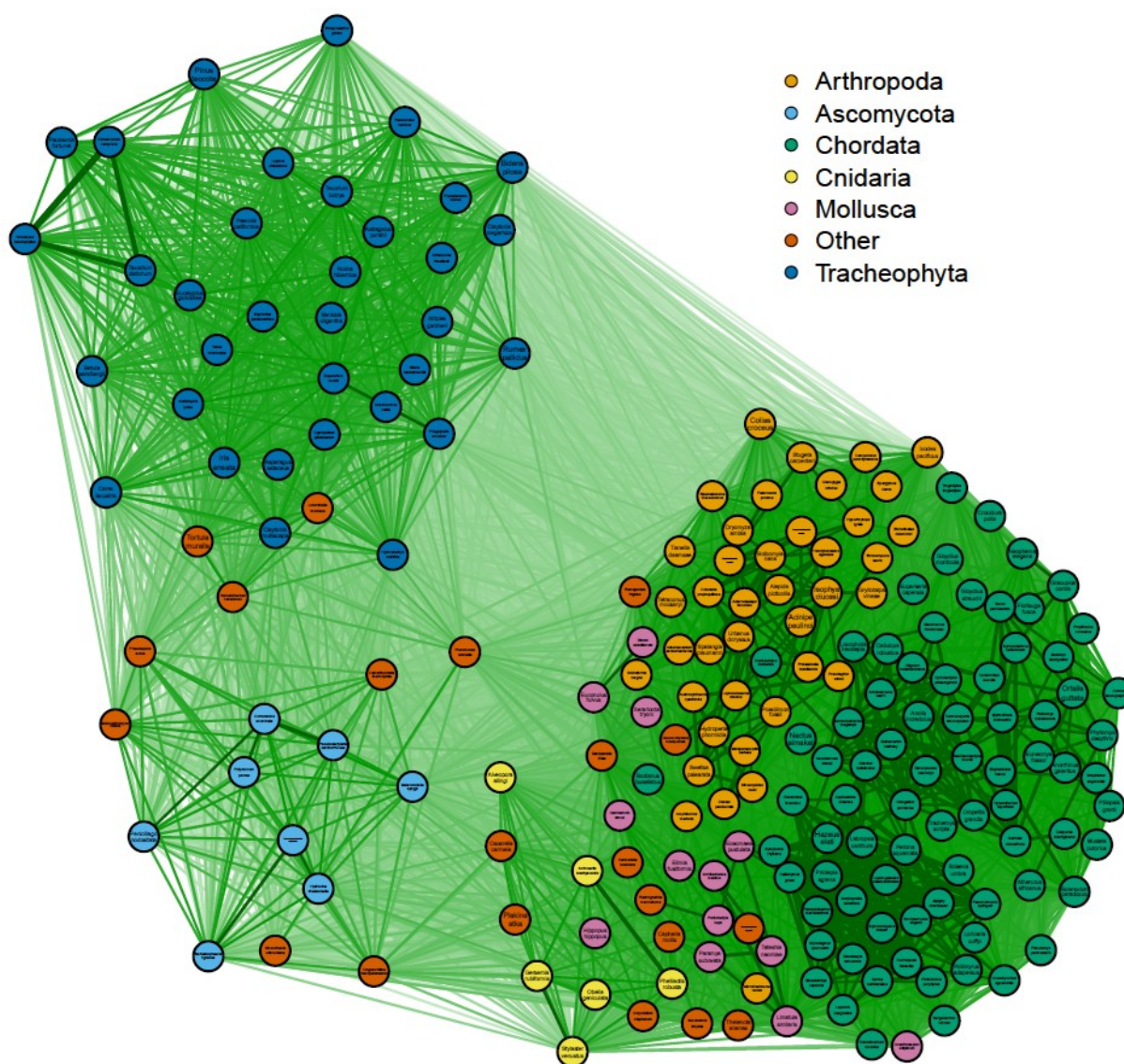


Figure 2. Phenotype-similarity network of 200 organisms, colored by phylum. Each node is one organism; color indicates phylum-level plotting group, with taxa outside the six focal phyla shown as "Other." Edges are positive pairwise polychoric correlations across the 196 questions; negative correlations were set to zero before plotting. Node positions come from a force-directed layout, so organisms with stronger positive phenotype similarity appear closer together. The layout is a qualitative visualization of phenotype similarity, not a calibrated phylogenetic or evolutionary distance.

Discussion

LifeDive demonstrates that a general-purpose, web-search-enabled LLM, embedded in a scripted pipeline, can generate a dense, uniform, cross-kingdom phenotype matrix at a scale far beyond ordinary manual curation—35,856 organisms scored on 196 questions, more than seven million values. Three independent lines of evidence support its usefulness: factor analysis recovered fourteen interpretable dimensions, items correlated strongly with independent trait-database criteria, and focused validation of a random subsample estimated high directional accuracy while reconstructing major organismal groupings. Together these indicate that LLM annotation can complement expert curation by filling a different niche—rapid, standardized, broad-coverage estimation across many taxa and traits.

The claim is not that every LifeDive cell is an authoritative fact; a single generated value may be wrong, oversimplified, or contingent on prompt wording and retrieval context. The claim is that a large matrix of structured estimates can preserve biological signal well enough to support exploratory, comparative, and

methodological analyses. Its strongest uses involve aggregate dimensions, broad contrasts, hypothesis generation, prioritization for expert review, and analyses that can model cell-level uncertainty; its weakest are species-level claims that hinge on a single unverified cell. Set against the cost and coordination of comparable curation efforts, the broader implication concerns who can build such resources and how quickly: a capability that recently required a consortium and years of funded effort is now within reach of an individual investigator for a few thousand dollars—provided that reliability is demonstrated rather than assumed, which is the burden this resource is designed to meet.

Several design choices likely contributed to the recovered structure: a fixed ordinal scale and CSV schema, scripted batching and parsing, retry-or-discard handling of malformed responses, explicit treatment of Unknown responses, and validation built into the workflow—mirroring broader recommendations for LLM-based scientific extraction (10, 18, 19). Web search probably helped for obscure taxa, though retrieval adds source-dependence and temporal instability. The validation also defines appropriate caution: accuracy was higher for confident than for hedged responses; external validation was strongest where expert databases are strongest (vertebrates, some marine organisms, selected ecological traits) and weaker for fungi, many invertebrates, and uncatalogued plant traits; and one component of the audit was itself LLM-assisted and should be read as convergent rather than definitive. Expert review remains necessary for rare taxa, ambiguous traits, and analyses that depend on specific cells.

Reproducibility here is best understood at the level of a versioned data product. The organism list, prompts, parsing code, imputation, and analyses can be rerun, but exact raw annotations from a commercial, web-search-enabled model may shift with model updates, safety tuning, API behavior, and the state of the web. LifeDive should therefore be cited with model name, collection date, prompt set, response scale, code version, and validation report, and per-cell metadata—original response category, imputation status, prompt version, and validation status—would let users distinguish generated, imputed, validated, and unchecked values. More generally, hallucination, training-data bias, and fluent but incorrect outputs are systemic risks of generative AI (15, 28), so responsible reuse requires data dictionaries, provenance, versioning, and explicit statements of intended use (29); without that context, dense AI-generated tables can appear more authoritative than their evidentiary status warrants.

Future development should be hybrid. Expert review can focus on hedged responses, low-validity traits, rare taxa, and high-leverage cells; cross-model ensembling can use agreement among independent systems as a noise filter; and retrieval-augmented designs can ground difficult taxa in curated floras, monographs, and trait databases while preserving a uniform schema. Because the same organism list, questions, and validation criteria can be applied to successive models, LifeDive can also serve as a benchmark for scientific data generation itself, making it possible to test whether newer models improve accuracy, reduce taxonomic bias, or change the latent structure of the matrix. In its present form, LifeDive is best viewed as a complement to expert curation: a broad computational phenotype layer that extends coverage, exposes candidate structure, and provides a concrete platform for evaluating LLM-assisted data generation in biodiversity science.

Data and materials availability

All data, metadata, prompts, validation materials, and analysis code are available through the Open Science Framework at <https://osf.io/aqkhn>. The imputed analysis matrix is also available as a [CSV](#) file.

Acknowledgments and declarations

Funding. This work received no external funding; the application programming interface (API) and computing costs (approximately US\$10,000) were borne by the author.

Author contributions. T.M.M. is the sole author; he conceived the study, built and ran the pipeline, performed all analyses, and wrote the manuscript.

Competing interests. The author declares no competing interests.

Use of generative AI. The phenotype annotations were generated by a large language model, and an independent audit was performed with a separate model, as described in Methods. In addition, I used Claude Opus 4.8, GPT-5.5 Pro Extended, and Grok 4.3 to prepare initial drafts and for proofreading. I reviewed, edited, and take responsibility for all content.

Supplementary materials

- Supplementary Methods
- Supplementary Results
- Note S1 (independent-audit report)
- Tables S1-S3
- Supplementary References (30-56)

References

1. J. Kattge et al., TRY plant trait database—enhanced coverage and open access. *Glob. Change Biol.* **26**, 119 (2020).
2. V. Domazetoski et al., Using large language models to extract plant functional traits from unstructured text. *Appl. Plant Sci.* **13**, e70011 (2025).
3. D. Marcos et al., Fully automatic extraction of morphological traits from the web: Utopia or reality? *Appl. Plant Sci.* **13**, e70005 (2025).
4. K. E. Jones et al., PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648 (2009).
5. H. Wilman et al., EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* **95**, 2027 (2014).
6. C. D. Soria, M. Pacifici, M. Di Marco, S. M. Stephen, C. Rondinini, COMBINE: A coalesced mammal database of intrinsic and extrinsic traits. *Ecology* **102**, e03344 (2021).
7. J. A. Tobias et al., AVONET: Morphological, ecological and geographical data for all birds. *Ecol. Lett.* **25**, 581 (2022).
8. L. Endara, H. Cui, J. G. Burleigh, Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Appl. Plant Sci.* **6**, e1035 (2018).
9. R. A. Folk, R. P. Guralnick, R. T. LaFrance, FloraTraiter: Automated parsing of traits from descriptive biodiversity literature. *Appl. Plant Sci.* **12**, e11563 (2024).
10. M. J. Farrell et al., The changing landscape of text mining: A review of approaches for ecology and evolution. *Proc. R. Soc. B* **291**, 20240423 (2024).
11. A. Castro, J. Pinto, L. Reino, P. Pipek, C. Capinha, Large language models overcome the challenges of unstructured text data in ecology. *Ecol. Inform.* **82**, 102742 (2024).
12. A. V. Gougherty, H. L. Clipp, Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *npj Biodivers.* **13**, 13 (2024).
13. F. Keck, H. Broadbent, F. Altermatt, Extracting massive ecological data on state and interactions of species using large language models. *bioRxiv* (2025); <https://doi.org/10.1101/2025.01.24.634685>.
14. F. Petroni et al., Language models as knowledge bases? in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, 2019), pp. 2463–2473.
15. M. Wysocka, O. Wysocki, M. Delmas, V. Mutel, A. Freitas, Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *J. Biomed. Inform.* **158**, 104724 (2024).
16. T. M. Moore, Using large language models to obtain quantitative data. *PsyArXiv* (2025); <https://psyarxiv.com>.
17. M. E. W. Varnum, N. Baumard, M. Atari, K. Gray, Large language models based on historical text could offer informative tools for behavioral science. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2407639121 (2024).
18. J. Dagdelen et al., Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
19. M. P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
20. H. Lai et al., Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digit. Med.* **8**, 74 (2025).
21. C. Mammides et al., Emerging applications of large language models in ecology and conservation science. *Conserv. Biol.* **40**, e14307 (2026).
22. W. D. Kissling et al., Establishing macroecological trait datasets: Digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. *Ecol. Evol.* **4**, 2913 (2014).

23. B. F. Oliveira, V. A. São-Pedro, G. Santos-Barrera, C. Penone, G. C. Costa, AmphibiO, a global database for amphibian ecological traits. *Sci. Data* **4**, 170123 (2017).
24. M. L. D. Palomares, D. Pauly, Eds., *SeaLifeBase* (2024); <https://www.sealifebase.org>.
25. R. Froese, D. Pauly, Eds., *FishBase* (2024); <https://www.fishbase.org>.
26. RS-eco, *traitdata: R Package for Easy Access to Various Ecological Trait Data* (2023); <https://github.com/RS-eco/traitdata>.
27. GBIF Secretariat, *GBIF Backbone Taxonomy*, dataset (2023); <https://doi.org/10.15468/39omei>.
28. R. S. Nasr, S. El-Deeb, Generative AI and its potential impact on humanity. *Int. J. Consum. Stud.* **49**, e70024 (2025).
29. H. Chafetz, S. Saxena, S. G. Verhulst, *A Fourth Wave of Open Data? Exploring the Spectrum of Scenarios for Open Data and Generative AI* (The GovLab and Open Data Policy Lab, 2024); <https://opendatapolicylab.org/>.
30. Catalogue of Life, *Catalogue of Life checklist (version 2026-04-18 XR)*, dataset (2026); <https://doi.org/10.48580/dgxjw>.
31. xAI, *Batch API*, xAI Docs (2026); <https://docs.x.ai/developers/advanced-api-usage/batch-api> [accessed 21 May 2026].
32. R. Likert, A technique for the measurement of attitudes. *Arch. Psychol.* **22**, 1 (1932).
33. L. Breiman, Random forests. *Mach. Learn.* **45**, 5 (2001).
34. H2O.ai, *h2o.randomForest*, H2O-R package documentation (2026); <https://docs.h2o.ai> [accessed 21 May 2026].
35. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2026); <https://www.R-project.org/>.
36. W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, R package version 2.6.5 (Northwestern Univ., 2026).
37. L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, E. J. Strahan, Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**, 272 (1999).
38. A. E. Hendrickson, P. O. White, Promax: A quick method for rotation to oblique simple structure. *Br. J. Stat. Psychol.* **17**, 65 (1964).
39. J. L. Horn, A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179 (1965).
40. W. F. Velicer, Determining the number of components from the matrix of partial correlations. *Psychometrika* **41**, 321 (1976).
41. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
42. K. Slowikowski, *ggrepel: Automatically Position Non-Overlapping Text Labels with ggplot2*, R package version 0.9.8 (2026).
43. T. L. Pedersen, *patchwork: The Composer of Plots*, R package version 1.3.2 (2025).
44. U. Olsson, Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443 (1979).
45. S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, D. Borsboom, qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, 1 (2012).
46. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129 (1991).
47. N. Pérez-Harguindeguy et al., New handbook for standardised measurement of plant functional traits worldwide. *Aust. J. Bot.* **61**, 167 (2013).
48. R. Villanueva, V. Perricone, G. Fiorito, Cephalopods as predators: A short journey among behavioral flexibilities, adaptations, and feeding habits. *Front. Physiol.* **8**, 598 (2017).
49. S.-H. Lee, D.-S. Kim, S.-H. Park, H. Park, Phytochemistry and applications of *Cinnamomum camphora* essential oils. *Molecules* **27**, 2695 (2022).
50. R. A. S. N. Ranasinghe, S. D. T. Maduwanthi, R. A. U. J. Marapana, Nutritional and health benefits of jackfruit (*Artocarpus heterophyllus* Lam.): A review. *Int. J. Food Sci.* **2019**, 4327183 (2019).
51. L. P. Wilhite, J. R. Toliver, in *Silvics of North America: Volume 1, Conifers*, R. M. Burns, B. H. Honkala, Tech. Coords. (Agriculture Handbook 654, U.S. Department of Agriculture, Forest Service, 1990).
52. C. J. Jackson, A. Reyes-Prieto, The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanopteryche gloeocystis*: Multilocus phylogenetics suggests a monophyletic Archaeplastida. *Genome Biol. Evol.* **6**, 2774 (2014).
53. B. J. Harris et al., Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat. Ecol. Evol.* **6**, 1634 (2022).
54. J. Larraín, A. Fife, C. Atala, *Lorentziella* (Gigaspermaceae, Bryophyta) new for Chile, and lectotypification of the genus. *Bol. Soc. Argent. Bot.* **52**, 351 (2017).
55. A. M. A. Aguinaldo et al., Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489 (1997).
56. I. Cid del Prado-Vera, H. Ferris, S. A. Subbotin, Six new and one known species of *Geomonhystera* (Nematoda, Monhysteridae) from moss, an epiphytic plant and soil in México and Ecuador. *Zootaxa* **4471**, 76 (2018).

Supplementary Information

A language model generates seven million phenotype estimates across the tree of life

Tyler M. Moore

Corresponding author. Email: tymoore@pennmedicine.upenn.edu

Contents

- Supplementary Methods
- Supplementary Results
- Note S1. Independent validation report
- Supplementary Tables S1-S3
- Supplementary References (30-56)

Supplementary Methods

Organism selection and taxonomic source

There were two primary considerations when deciding which organisms to include in the LifeDive database: 1) every taxonomic family must be represented in the final sample, and 2) organisms with common names should be moderately preferred over organisms without one. The list of all possible organisms to select from was a Catalogue of Life Data Package (ColDP) download, which provides a structured checklist for taxonomic names and related metadata (30). The compilation script read the ColDP NameUsage.tsv table, normalized prefixed column names, and retained records at the species rank with status coded as accepted or provisionally accepted. For each accepted species, the script retained taxonomic identifiers and the available higher taxonomy fields (kingdom, phylum, class, order, family, and genus). Binomial scientific names were constructed from genericName and specificEpithet when those fields were available, with scientificName used as a fallback. English vernacular names were then merged from VernacularName.tsv when present, using the first English record for each taxon when multiple vernacular names 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 (31). 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 organism was evaluated on a set of phenotype questions using a five-point Likert-type response scale (32): 0 = Definitely No, 1 = Probably No, 2 = Unknown or not applicable, 3 = Probably Yes, and 4 = Definitely Yes. The final analysis included 196 retained phenotype questions. To reduce parsing failures, organism names were sanitized for JSON by removing or replacing characters likely to interfere with prompt or CSV structure.

Organisms were processed in batches of three. Each prompt listed the organisms, listed the phenotype questions, defined the response scale, and required a CSV response containing one organism row and one numeric column per question. The prompt also required that every response cell be populated and instructed the model to use the Unknown code when the answer could not be determined.

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 (33). 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 (34). 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, the original response scale was recoded for the factor analysis by collapsing the two affirmative codes downward: 0 and 1 retained their original meanings, 3 was recoded to 2, and 4 was recoded to 3. This produced a four-level ordinal response scale that excluded the Unknown category. The retained item matrix was analyzed in R (35) using the `psych` package (36). Specifically, the script used `psych::irt.fa` to estimate a 14-factor item response theory factor-analysis model and applied promax rotation (37, 38). The number of factors was set to 14 for interpretability; parallel analysis and the minimum-average-partial criterion indicated substantially more factors (58 and 41, respectively) (39, 40). Promax rotation was appropriate because the expected phenotype dimensions were not assumed to be independent; oblique rotations are commonly used when interpretable factors may correlate. Total (sum) scores were then computed based on the loadings of the item-factor analysis. True IRT (theta) scores were not used because some items had such high discrimination parameters that they determined the theta levels almost entirely on their own. Sum scores ensured a more “floodlight” (as opposed to “spotlight”) approach to scoring.

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

To visualize the factor-score structure in familiar organisms, the script read a separate file of common organisms and their factor scores, then generated two scatterplots with `ggplot2` (41), `ggrepel` (42), and `patchwork` (43). The first panel plotted Durable scores against Plant-like scores, and the second plotted Aquatic scores against Dangerous scores. Points were labeled with common names using repelled labels to reduce overlap. Each

panel included a dashed $y = x$ reference line; this line was a visual reference and not a fitted model. The plots were saved as a 10 inch by 12 inch PNG at 200 dpi.

Validation

For LifeDive to be of any use, the accuracy of responses obtained from Grok must be established. This is a daunting task because $35,856 \times 196 = 7,027,776$ values were generated. Because validating each value is not feasible, I used a multi-method approach, attempting to assess validity from multiple angles. The first method involved comparing LifeDive values to previously established values from multiple public databases. The second method involved taking a random sample of 200 LifeDive organisms and more thoroughly establishing validity on this smaller subset, under the assumption that the validity level of the subset would give an approximation for the validity of the larger, complete set.

External validity criteria

To assess the external validity of the LifeDive phenotype set, I compiled a benchmark of organism-level traits from nine publicly available, peer-reviewed databases, supplemented with a standardized higher-level taxonomic backbone. For birds, I used EltonTraits 1.0 (5), which provides species-level diet composition, foraging strata, body mass, and activity-period attributes for ~10,000 bird species. The corresponding mammal traits — diet composition, foraging stratum, activity period, and body mass — were drawn from EltonTraits 1.0 for mammals (5). Additional mammal life-history and ecological traits, including activity cycle, trophic level, and habitat breadth, came from PanTHERIA (4). Detailed mammal diet categorisations (0–3 ordinal scores indicating the prominence of each food category in the species' diet) were obtained from MammalDIET (22). For amphibians, ecological traits (habitat, diet, activity, reproductive mode) were taken from AmphiBIO (23). For marine invertebrates, habitat-zone categorizations were drawn from SeaLifeBase (24). For fish, ecological traits were drawn from FishBase (25). To anchor each organism to a standardized higher-level taxonomy (kingdom through family), I used the taxonomy table assembled by the traitdata R package (26), which integrates the GBIF Backbone Taxonomy (27) with the trait databases above.

Data preparation

The 35,856 organisms in the LifeDive dataset were matched to each external database using their genus-species binomials, after normalizing whitespace and case. Where multiple sources covered the same trait (e.g., body mass appears in EltonTraits, PanTHERIA, the Amniote Life-History Database, and AnAge), each source was retained as a separate column to permit independent validation. Categorical variables (e.g., FishBase DemersPelag, AmphiBIO multi-membership habitat fields) were expanded into per-category 0/1 indicator variables, with missing values propagated so that organisms outside the source database remained missing rather than being miscoded as absent. FishBase's coded binary indicators (–1 = yes, 0 = no) were recoded to standard 0/1. AmphiBIO's 1/NA-coded multi-membership fields (e.g., habitat use) were recoded so that a row present in AmphiBIO but unflagged for a given category received 0 (i.e., the category does not apply), while rows absent from AmphiBIO remained NA. After cleaning, the validation table contained 168 numeric columns spanning taxonomic indicators (e.g., is_bird, is_mammal), continuous traits (body mass, longevity, depth ranges), ordinal traits (mammal diet scores, fish commercial importance), and binary traits (e.g., bird is nocturnal, amphibian uses aquatic habitat). Of the 35,856 organisms, 15,525 (43.3%) matched at least one external trait database, with coverage strongest for vertebrates and weaker for plants, fungi, and most invertebrates.

Correlation methods

Correlations between each LifeDive item and each validation criterion were computed pairwise using the appropriate method given the measurement type of each variable: Pearson product-moment for continuous $\times$ continuous pairs, biserial for continuous $\times$ dichotomous pairs, polyserial for continuous $\times$ polytomous pairs, tetrachoric for dichotomous $\times$ dichotomous pairs, and polychoric for any pair involving two polytomous variables or one polytomous and one dichotomous variable (36, 44). All correlation functions were drawn from the psych R package (36). Variable types were assigned automatically by a classifier: columns with two unique non-missing values were treated as dichotomous; integer-valued columns with three to seven unique values and range ≤ 7 were treated as polytomous; all other numeric columns were treated as continuous. Pairwise complete observations were used throughout, with a minimum overlap of 30 observations required to compute any single correlation. Pairs that produced numerical errors during estimation (e.g., empty cross-tabulation cells leading to

undefined polychoric thresholds) were assigned a value of NA rather than dropped from analysis. Pairs in which either variable was constant within the overlap were also assigned NA. For each validation criterion, the five LifeDive items most strongly correlated (by absolute magnitude) with that criterion were retained for inspection.

Random n = 200 subsample approach

The second validation approach taken here was to randomly sample 200 organisms for which rigorous validation could be more realistically attempted. These 200 organisms (including their taxonomic phylum, class, and order) are listed in Supplementary Table S1. All procedures and analyses below were conducted on these 200 organisms.

LLM-assisted audit in the n = 200 subsample

A first step in assessing the validity of data on the 200 random organisms was to ask an LLM that was not involved in collecting the data—namely, Claude Opus 4.7 extended—to compare the values of the individual organisms against its own knowledge and that obtained from its own rapid web search capabilities. Briefly, Claude used two complementary and independent approaches. The first built a taxonomy-based reference standard: each organism was classified into a broad taxonomic group, and expected answers were derived for roughly 50 questions whose answers are unambiguously determined by group membership (e.g., cellulose production in land plants, chitin use in arthropods and fungi), yielding 7,641 evaluable cells (~20% of the total). The second approach verified 240 species-specific cells (20 organisms × 12 questions) against authoritative sources to cover the harder, non-taxonomic questions. The full prompt and response are provided in the Supplement.

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 organism similarity was estimated using polychoric correlations across the phenotype questions. Polychoric correlations were used because the phenotype scores were ordinal rather than continuous, and this estimator models the correlation between latent continuous variables assumed to underlie ordered categorical observations (36, 44). Negative correlations were retained for numerical interpretation but were set to zero before plotting; consequently, the displayed graph represents positive phenotype similarity only, not both attraction and repulsion among organisms.

The resulting non-negative weighted correlation matrix was visualized as a network using qgraph, with organisms represented as nodes and edge weights proportional to positive polychoric correlations. A force-directed spring layout was used, such that nodes with stronger positive similarities were drawn closer together, whereas nodes with weaker positive similarities tended to occupy more peripheral or intermediate positions (45, 46).

Node colors indicate phylum-level plotting group. Because the phenotype questions span morphology, physiology, ecology, reproduction, cellular traits, and human-associated organismal attributes, the network should be interpreted as a phenotype-similarity graph rather than as a phylogenetic tree.

Supplementary Results

LifeDive phenotype matrix

The workflow produced a broad phenotype matrix containing 35,856 organisms scored on 196 retained phenotype questions, for 7,027,776 organism-by-phenotype values. The taxonomic selection strategy intentionally favored breadth: rather than sampling only well-known organisms, the script first guaranteed family-level representation before filling the remaining quota with high-priority taxa. The retained phenotype variables covered anatomical, ecological, physiological, behavioral, life-history, and human-salience features, allowing the same response framework to be applied across plants, animals, and fungi.

The data-processing workflow also produced a complete analysis matrix by treating Unknown/not applicable responses as missing and imputing missing item values from the remaining phenotype profile. This

imputation strategy made it possible to conduct a full-matrix factor analysis across all retained items. However, because imputed values were inferred from other LLM-derived annotations, the completed matrix should be used as a latent-structure and visualization resource rather than as a substitute for organism-specific primary biological evidence.

Factor-analysis results

The 14-factor promax-rotated solution was readily interpretable from the full pattern matrix (Supplementary 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 (.47). 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 therefore support the substantive labels assigned to the factors and illustrate how LifeDive scores can place very different organisms in a shared phenotypic space.

External validation

Supplementary Table S3 reports the five LifeDive items most strongly correlated with each of 44 selected validation criteria, spanning kingdom-, phylum-, and class-level taxonomic indicators as well as within-taxon ecological traits drawn from seven trait databases. The strongest correlations were observed for higher-level taxonomic indicators, where multiple LifeDive items correlated with face-valid trait analogs at $|r| > .95$. For *is_animal*, the strongest negative correlations were with items capturing plant- and protist-specific cellular features — production of cellulose ($r = -.99$), presence of cell walls ($r = -.97$), and photosynthetic capacity ($r = -.97$) — while presence of centrioles, a hallmark of animal cells, correlated positively at $r = .98$. The mirror pattern was observed for *is_plant* (cellulose $r = .99$, photosynthesis $r = .97$, cell walls $r = .97$). At the class level, *is_mammal* correlated nearly perfectly with the LifeDive item asking whether the organism has hair, fur, or trichomes ($r = .99$), and negatively with the item asking whether eggs are involved in reproduction ($r = -.98$). The item asking whether the organism does most of its activities at night correlated with *bird_is_nocturnal* at $r = .92$ and with *mammal_is_active_nocturnal_elton* at $r = .75$ (and negatively with the corresponding diurnal indicator at $r = -.73$). Items capturing aquatic specialization correlated with *mammal_foraging_marine* at $r > .90$.

Within-taxon ecological criteria likewise recovered face-valid associations. For *amphibian_has_larval_stage*, the strongest correlate was the LifeDive item asking whether the organism undergoes dramatic morphological changes between life stages ($r = .90$), followed by items reflecting reliance on water for reproduction ($r = .87$), occupancy of freshwater habitats ($r = .75$), use of gills for gas exchange ($r = .74$), and aquatic lifestyle ($r = .74$). For *mammal_is_mammal_eater_binary* (whether a mammal eats other mammals), the top correlates were capability of trapping or capturing prey ($r = .59$), descriptions as “ferocious” ($r = .58$), and noticeable black coloring ($r = .55$). For *marine_invert_is_sessile*, the top positive correlate was the LifeDive item asking whether the organism is sessile ($r = .84$).

A small number of correlations were less interpretable. Within-taxon criteria with low prevalence occasionally produced large-magnitude correlations with LifeDive items having no obvious biological connection to the criterion — for example, *bird_is_nocturnal* showed a moderately strong negative correlation with the lignin-degradation item, despite no bird being expected to degrade lignin. We interpret these as instabilities of polychoric estimation under sparse pairwise cross-tabulations rather than as substantive validity evidence, and we revisit them in the Discussion. The overall pattern across criteria, however, is one of strong agreement between LifeDive ratings and

independent expert-curated phenotypic data, particularly at higher taxonomic levels and for ecological criteria with well-balanced class distributions.

Organism phenotype-similarity network

The network recovered a strong phylum-level signal. Mean within-phylum polychoric correlations were high for Ascomycota (mean $\rho = .877$), Cnidaria ($\rho = .853$), Arthropoda ($\rho = .823$), Chordata ($\rho = .810$), Mollusca ($\rho = .785$), and Tracheophyta ($\rho = .766$). In contrast, the “Other” category had a much lower mean within-group correlation ($\rho = .510$), 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 $\rho = .689$), Chordata–Mollusca ($\rho = .680$), and Arthropoda–Mollusca ($\rho = .668$). Tracheophytes were much less similar to animal groups, including Chordata ($\rho = .171$), Arthropoda ($\rho = .233$), and Mollusca ($\rho = .205$), 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. The strongest associations were among trees: jackfruit, *Artocarpus heterophyllus*, and camphor tree, *Cinnamomum camphora* ($\rho = .997$); jackfruit and bald cypress, *Taxodium distichum* ($\rho = .989$); and camphor tree and bald cypress ($\rho = .985$). Other strong positive examples included frog–frog similarity between Boulenger’s callulops frog, *Callulops robustus*, and the San Lorenzo robber frog, *Serranobatrachus megalops* ($\rho = .974$); flatfish similarity between the threespot sand flounder, *Ancylopsetta dendritica*, and the patchtail tonguefish, *Symphurus rhytisma* ($\rho = .973$); reef-fish similarity between the yellowtail tubelip wrasse, *Labropsis xanthurus*, and the network reefgoby, *Priolepis agrena* ($\rho = .972$); and sea-anemone similarity between *Acricoactis brachyacontis* and *Phelliactis robusta* ($\rho = .967$). 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 loriciid catfish *Loricaria cuffyi* ($\rho = -.164$), jackfruit and the glass knifefish *Japigny kirschbaum* ($\rho = -.116$), and common lake sedge, *Carex lacustris*, and the Haitian solenodon, *Solenodon paradoxus* ($\rho = -.115$).

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 (47).

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 (48). The correlations support this interpretation: *Acanthosepion ellipticum* correlated strongly with multiple chordate fishes, including Gross’s dragonet, *Callionymus grossi* ($\rho = .866$), the common silver-biddy, *Gerres subfasciatus* ($\rho = .842$), Englert’s scorpionfish, *Scorpaenodes engleri* ($\rho = .836$), and the freetail brotula, *Bidenichthys capensis* ($\rho = .833$).

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* ($\rho = .913$) and the bicolor cactus snail *Xerarionta tryoni* ($\rho = .876$), but it also correlated strongly with arthropods, including the oribatid mite *Bipassalozetes macedonicus* ($\rho = .821$) and the mite *Asternolaelaps fecundus* ($\rho = .802$). 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 (49, 50, 51). 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* ($\rho = .853$), followed by more moderate relationships with fungi and plant-like organisms. Thus, proximity to Ascomycota reflects overlapping non-animal phenotype states, not fungal identity (52).

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* ($\rho = .843$) and with several tracheophytes, including Mohr’s thoroughwort, *Eupatorium mohrii* ($\rho = .838$), and Rydberg’s spring-beauty, *Claytonia multiscapa* ($\rho = .835$). *Tortula muralis* showed a similar pattern, correlating strongly with the moss *Homalothecium nevadense* ($\rho = .890$), *Lorentziella imbricata* ($\rho = .843$), and several tracheophytes. This supports the interpretation that “Other” is a plotting category, not a biological group (53, 54).

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* ($\rho = .911$), the bush-cricket *Isophya ciucasi* ($\rho = .874$), the mite *Asternolaelaps fecundus* ($\rho = .868$), and the mite *Hemipteroseius ageneius* ($\rho = .861$). Thus, its position near arthropods is one of the more biologically coherent “Other” placements (55, 56).

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 $\rho = .358$ to $\rho = .498$ across the plotted phylum categories. The strongest organism-level association was with *Gloeochaete wittrockiana* ($\rho = .853$), 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.

Overall, the network structure is biologically reasonable when interpreted as phenotype similarity. The graph recovers major organismal syndromes—woody plants, fungi, active vertebrates, small terrestrial invertebrates, marine invertebrates, bryophytes, algae, and other residual lineages—rather than a strict evolutionary hierarchy. The principal caution is that negative correlations were omitted from the plotted edge matrix and the spring layout is qualitative. Consequently, physical distance in the graph should be interpreted as a visual summary of positive phenotype similarity, not as a calibrated phylogenetic, evolutionary, or taxonomic distance.

Note S1. Independent validation report generated by Claude Opus 4.7

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.

RESPONSE: Intermediate analysis and duplicated tool-status messages have been omitted; the final audit summary and report follow.

Summary

Best estimate: ~95–97% of the 39,200 cell responses are directionally correct. Two independent methods converge here:

Method	Cells	Accuracy
Taxonomy-rule evaluation	7,641 (19.5%)	96.5% (95% CI 96.1–96.9%)
Web-verified spot-check	240 (0.6%)	97–99%

Approach. I built two complementary reference standards. For Method 1, I classified all 200 organisms into broad taxonomic groups and wrote rules for ~50 questions where group membership unambiguously determines the answer (e.g., “produces cellulose” → yes for all land plants; “uses chitin” → yes for arthropods and fungi). That covers roughly 1 in 5 cells. For Method 2, I picked 20 organisms across diverse groups and verified 12 species-specific questions each against authoritative sources, catching things taxonomy can't (geographic range, IUCN status, behavior).

Key findings:

- **Calibration is good.** Confident answers (0 or 4, 86.8% of cells) are 97.1% accurate. Hedging answers (1 or 3, 13.2%) are 73.4% — exactly the pattern you want from a well-calibrated system.
- **Weakest group is angiosperms** (93.6%), driven by a handful of organisms where Grok misses basic plant anatomy (e.g., saying a cycad has no cell walls or uses chitin). These look like attention failures rather than knowledge gaps and would clean up with targeted re-prompting.
- **Weakest questions** are ones with interpretive ambiguity (Q12 internal gas-filled cavities — Grok reads narrowly as air sacs/swim bladders, not lungs), or species-specific detail not captured by broad taxonomy (Q190 fossil record, Q53 alternation of generations).
- **Notable: one of my “expected” answers turned out to be Grok being more correct than me.** I had *Harmeriella terebrans* classified as a mite; it's actually a bryozoan that bores into other bryozoans, and Grok's answers were the correct ones. There are likely a few more like this — meaning the true error rate is probably a fraction of a point lower than the 96.5% number suggests.

The report includes worst-organism and worst-question breakdowns, error-pattern analysis, and three concrete suggestions for cleaning up the database further (targeted re-prompting of hedging cells, prompt rewording on the lowest-accuracy questions, and a continuous validation set to measure pipeline changes).

LifeDive Grok Response Accuracy Audit

Headline

Best estimate: approximately 95–97% of the 39,200 cell responses are directionally correct (i.e., a “yes” answer when the truth is yes, or a “no” when the truth is no). The two independent estimation methods I used converge on this range:

Method	Cells	Accuracy	95% CI
Taxonomy-rule evaluation	7,641 (19.5%)	96.5%	96.1–96.9%
Manual web-verified spot-check	240 (0.6%)	97–99%	95.0–99.2%

These estimates are concordant, but each has biases (see *Caveats* below). The honest reading is that the database is in the mid-to-high 90s for directional accuracy on yes/no questions, with a higher rate on confident answers (0 or 4) and a much lower rate on the small minority where Grok hedged (1 or 3).

Sample composition

200 organisms × 196 questions = 39,200 cells. The Grok response distribution:

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%

Two pipeline-level observations worth noting:

1. **Grok always commits.** Zero cells were coded “2” (Unknown). For a calibration-aware downstream model, that’s a property of the data: the system has no abstention class, only a 5-level certainty spectrum collapsed to four levels in practice.
2. **The distribution is heavily skewed toward “No”** (67.5% no/probably-no vs. 32.5% yes/probably-yes), which is expected when each organism is asked many questions most of which don’t apply to it. Most questions are asking about features only a small fraction of organisms have.

Method 1: taxonomy-based rules engine

I built a reference standard by classifying all 200 organisms into broad taxonomic groups (vertebrate classes, arthropod subgroups, mollusk types, fungi, lichen-fungi, angiosperms, gymnosperms, ferns, mosses, algae, sponges, cnidarians, etc.) and then writing rules for ~50 questions where broad group membership unambiguously determines the answer. Examples:

- “Does this organism produce cellulose?” → yes for all land plants and green algae; no for animals and fungi
- “Is iron used in this organism’s physiology?” → yes for all animals (hemoglobin / cytochromes)
- “Are seeds, eggs, or spores involved in reproduction?” → yes for animals, plants, fungi (almost universally for these groups)
- “Does this organism use chitin as a structural material?” → yes for arthropods and fungi; no for vertebrates and plants

I deliberately did **not** write rules for questions where biological variation within a broad group is large enough that the answer is genuinely species-specific (e.g., “primarily aquatic” for angiosperms — there ARE aquatic flowering plants, so no group-level rule is safe).

Scoring: Grok 0/1 vs. “no” expected → correct; Grok 3/4 vs. “yes” expected → correct.

Result: 96.5% on 7,641 evaluable cells (95% CI 96.1–96.9%).

Stratified breakdowns

By Grok confidence level (on the taxonomy-evaluable subset):

- Confident responses (0 or 4): 97.1% accurate on 7,440 cells
- Hedging responses (1 or 3): 73.4% accurate on 203 cells

This is exactly what you’d expect of a well-calibrated system: when Grok says “Definitely,” it’s almost always right; when Grok hedges with “Probably,” it’s right about 3 times out of 4. The hedging cells are 13.2% of the full dataset (5,171 of 39,200), so even if they sit closer to 70–75% accuracy across the board, they only dent the overall number modestly.

By taxonomic group (cells / accuracy, top groups by sample size):

Group	n_orgs	cells	accuracy
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%
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%

Angiosperms are the weakest group at 93.6%, driven primarily by basic-biology questions where Grok occasionally misses things (“Do cells have cell walls?” → 0 for a cycad; chitin → yes for a cycad). Most other groups are 95–99%.

Worst-performing questions (after removing rules that were too rigid to be fair references):

Q	Question (abbreviated)	Accuracy
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%

These are mostly genuinely difficult questions where the answer depends on species-specific detail not captured by taxonomy alone (e.g., “fossil record” depends on whether the *specific* species has hard parts and is well-represented in the record, not just whether it's a vertebrate).

Method 2: manual spot-check with web verification

To get a read on the ~80% of cells *not* covered by taxonomy rules — the species-specific questions where the answer depends on actual biological detail — I picked 20 organisms across diverse groups and 12 well-defined questions per organism (geographic distribution, behavior, ecology, morphology), then verified answers against authoritative sources (Wikipedia, IUCN, Animal Diversity Web, BirdLife, peer-reviewed taxonomic literature) where I wasn't confident.

Result: ~237/240 ≈ 98.8% correct.

Examples of the kind of error caught: *Camponotus pennsylvanicus* (black carpenter ant) was coded as “primarily tropical = 4” (it's a Nearctic temperate species, eastern N. America) and “no close symbiosis = 0” (it has the obligate intracellular endosymbiont *Blochmannia pennsylvanicus*, well-documented since 2005). On the other hand, several initially-suspect calls held up — e.g., *Pseudoastrosphaeriella bambusae* was coded as freshwater = 4, which seemed odd for a fungus, but it's correctly listed in freshwaterfungi.org as a freshwater Dothideomycete.

The spot-check sample is small but it confirms that on species-specific questions Grok is also in the high-90s. Its 95% binomial CI overlaps the taxonomy-rule estimate.

Caveats and biases on the headline number

The estimate is bounded by the methods. Things that bias upward:

- **The taxonomy-evaluable subset (~20% of cells) is “easier” than the average cell.** Questions like “does this organism produce cellulose” have group-level answers; questions like “is this species endangered” don't. The taxonomy approach reports on the easier slice.
- **The spot-check sample was biased toward more familiar organisms.** Even though I included some genuinely obscure ones, organisms with more web presence are easier to verify and probably have more accurate Grok answers (more training data).

- **My rules were sometimes too rigid and caught Grok being correctly nuanced.** A few examples: I had marked Q24 (asexual reproduction) as “yes for all angiosperms” — many flowering plants don't commonly reproduce asexually, and Grok's “no” was actually defensible. I removed that rule. Similarly, on Q12 (internal gas-filled cavities) Grok seems to interpret narrowly (air sacs, swim bladders) rather than broadly (any lung), and reasonable people disagree.

Things that bias downward:

- **Some of my “expected” answers are themselves wrong**, and so count Grok correct as Grok wrong. *Harmeriella terebrans* was classified in my pipeline as a mite, but it's actually a bryozoan that bores into other bryozoans — Grok's “has tentacles = 4,” “primarily aquatic = 4” were correct calls that my reference standard initially marked wrong. I fixed that one organism after spot-checking the worst-accuracy outliers, but there may be others I haven't caught.
- **Questions like Q190 (fossil record) and Q53 (alternation of generations) have multiple defensible interpretations.** Counting Grok wrong every time it differs from one reading inflates the error rate.
- **The 5,171 hedging responses (1 or 3) are intentionally lower-confidence.** A “Probably No” that turns out to be “Definitely No” should arguably be scored as correct directionally even though we lose information. I scored it as correct only if the direction matched, which is the more honest definition.

Both directions of bias are present and roughly comparable in magnitude. My read is that the true rate is probably very close to 95–97% — possibly a hair lower than the headline if you weight harder questions, possibly a hair higher if you're generous with interpretive cases.

Error patterns

The 266 documented errors break down as:

- **Direction:** 60% expected-yes-but-Grok-said-no; 40% expected-no-but-Grok-said-yes. There's a slight bias toward false negatives (saying no when the answer is yes).
- **Confidence:** 80% of errors come from confident answers (0 or 4), 20% from hedging answers (1 or 3). But on a *rate* basis, hedging responses are much error-prone (73% accuracy vs. 97%).
- **Question type:** Errors cluster on questions requiring interpretation or species-specific biology rather than basic biology. Easy questions (cellulose, iron, chitin, cell walls in general) are 99–100% accurate.

Two concrete systematic patterns worth noting if you can re-prompt:

1. **Basic plant anatomy occasionally fails.** Examples: Grok said the Mulanje cycad doesn't have cell walls (every plant cell has a cell wall); said the same cycad uses chitin (plants don't); said Banksia and the cycad don't have closed circulatory/vascular systems (vascular plants by definition do). These look like attention failures rather than knowledge gaps. A targeted re-prompt for the ~5% of plants showing this pattern would clean up a chunk of errors cheaply.
2. **Internal anatomical questions on vertebrates are interpreted inconsistently.** “Internal gas-filled cavities” gets a “no” for mammals/reptiles/most fish but a “yes” for birds. Lungs are gas-filled cavities. The narrower reading (only air sacs / swim bladders qualify) is defensible but it's not how the question is naturally read. Worth either rewording the prompt or adding a clarifying example in the system prompt.

Recommendations for the broader 35,000-organism database

If the 200-organism sample is representative — and there's no obvious reason to think it isn't — then of the 7,027,776 cells in the full LifeDive database, an estimated 6.7–6.8 million are directionally correct. That's a strong foundation for a recommendation engine, where directional correctness on yes/no features is what matters most.

Three suggestions if you want to push the number higher:

1. **Targeted re-prompting on hedging responses.** The 13% of cells coded 1 or 3 are where most marginal improvement lives. A second pass with a model asked specifically to resolve those, possibly with web search, would likely reclaim 5–10 percentage points on that subset (and 0.5–1.5pp on the overall).
2. **Question-level audit on the lowest-accuracy questions.** Q190 (fossil record), Q75 (scales), Q53 (alternation of generations), Q12 (internal gas cavities) all have wording that produces interpretive disagreement. Rewording the prompt or adding 1–2 worked examples would lift these noticeably without re-running everything.

3. **A small validation set with hand-labeled ground truth** for a sample similar to this one (~500 organisms, all 196 questions, sampled stratified by taxonomic group) would let you measure accuracy continuously as you change prompts, models, or pipeline steps. The current setup is fine for a snapshot but doesn't let you tell whether a tweak helps or hurts.

Supplementary Tables

Table S1. Organisms Used for Deeper (n = 200) Validation

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

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
<i>Bipassalozetes macedonicus</i>	Arthropoda	Arachnida	Sarcoptiformes
<i>Bodianus busellatus</i> (Blackspot Hogfish)	Chordata	Actinopterygii	Labriformes
<i>Bolbomyia nana</i>	Arthropoda	Insecta	Diptera
<i>Cacosternum leleupi</i> (Katanga Caco)	Chordata	Amphibia	Anura
<i>Calcarisporiella thermophila</i>	Calcarisporiellomycota	Calcarisporiellomycetes	Calcarisporiellales
<i>Callionymus grossi</i> (Gross's dragonet)	Chordata	Actinopterygii	Syngnathiformes
<i>Callulops robustus</i> (Boulenger's Callulops Frog)	Chordata	Amphibia	Anura
<i>Cambarunio simus</i> (Caney fork rainbow)	Mollusca	Bivalvia	Unionida
<i>Camponotus pennsylvanicus</i> (black carpenter ant)	Arthropoda	Insecta	Hymenoptera
<i>Cantorchilus leucotis</i> (Buff-breasted Wren)	Chordata	Aves	Passeriformes
<i>Capheira mollis</i> (gentle sea cucumber)	Echinodermata	Holothuroidea	Apodida
<i>Carex lacustris</i> (common lake sedge)	Tracheophyta	Liliopsida	Poales
<i>Chasmistes fecundus</i> (Webbug Sucker)	Chordata	Actinopterygii	Cypriniformes
<i>Cingulochitina chacoparanaense</i>	Unknown	Chitinozoa	Operculatifera
<i>Cinnamomum camphora</i> (camphor tree)	Tracheophyta	Magnoliopsida	Laurales
<i>Claytonia megarhiza</i> (alpine spring-beauty)	Tracheophyta	Magnoliopsida	Caryophyllales
<i>Claytonia multiscapa</i> (Rydberg's spring-beauty)	Tracheophyta	Magnoliopsida	Caryophyllales
<i>Colias croceus</i> (Clouded Yellow)	Arthropoda	Insecta	Lepidoptera
<i>Colocasia propinquilinea</i> (Close-banded Demas)	Arthropoda	Insecta	Lepidoptera
<i>Cornuvesica acuminata</i>	Ascomycota	Sordariomycetes	Microascales
<i>Corvus leucognaphalus</i> (White-necked Crow)	Chordata	Aves	Passeriformes
<i>Cranopygia brindlei</i>	Arthropoda	Insecta	Dermaptera
<i>Craugastor amniscola</i> (Rivulet Rainfrog)	Chordata	Amphibia	Anura
<i>Crocidura polia</i> (Fuscous Shrew)	Chordata	Mammalia	Eulipotyphla
<i>Cypripedium yatabeanum</i> (Spotted Lady's Slipper)	Tracheophyta	Liliopsida	Asparagales
<i>Cyrtodactylus pinlaungensis</i> (Pinlaung Bent-toed Gecko)	Chordata	Reptilia	Squamata
<i>Dasyornis brachypterus</i> (Eastern Bristlebird)	Chordata	Aves	Passeriformes
<i>Deveximentum ruconius</i> (deep pugnose ponyfish)	Chordata	Actinopterygii	Perciformes
<i>Dimorphodon macronyx</i> (Dimorphodon)	Chordata	Reptilia	Pterosauria
<i>Discias pascuensis</i>	Arthropoda	Malacostraca	Decapoda
<i>Discus catskillensis</i> (angular disc)	Mollusca	Gastropoda	Stylommatophora
<i>Dryomyza ambliia</i>	Arthropoda	Insecta	Diptera
<i>Elephantulus fuscus</i> (Dusky Elephant Shrew)	Chordata	Mammalia	Macroscelidea

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
<i>Eleutherodactylus ricordii</i> (Oriente Yellow-mottled Frog)	Chordata	Amphibia	Anura
<i>Elimia fusiformis</i> (Fusiform Elimia)	Mollusca	Gastropoda	Unknown
<i>Encephalartos gratus</i> (Mulanje Cycad)	Tracheophyta	Cycadopsida	Cycadales
<i>Eoacmaea pustulata</i> (Spotted Limpet)	Mollusca	Gastropoda	Patellogastropoda
<i>Epargyreus clarus</i> (Silver-spotted Skipper)	Arthropoda	Insecta	Lepidoptera
<i>Eucalyptus globoidea</i> (white stringybark)	Tracheophyta	Magnoliopsida	Myrtales
<i>Euconulus fulvus</i> (tawny glass snail)	Mollusca	Gastropoda	Stylommatophora
<i>Euneomys fossor</i> (Burrowing Euneomys)	Chordata	Mammalia	Rodentia
<i>Euparkeria capensis</i>	Chordata	Reptilia	Unknown
<i>Eupatorium mohrii</i> (Mohr's thoroughwort)	Tracheophyta	Magnoliopsida	Asterales
<i>Euphorbia ipecacuanhae</i> (American ipecac)	Tracheophyta	Magnoliopsida	Malpighiales
<i>Flatormenis proxima</i> (northern flatid planthopper)	Arthropoda	Insecta	Hemiptera
<i>Florisuga fusca</i> (Black Jacobin)	Chordata	Aves	Apodiformes
<i>Garrulax cinereifrons</i> (Ashy-headed Laughingthrush)	Chordata	Aves	Passeriformes
<i>Geomonhystera mexiquense</i>	Nematoda	Chromadorea	Monhysterida
<i>Gerres subfasciatus</i> (common silver belly)	Chordata	Actinopterygii	Perciformes
<i>Gersemia rubiformis</i> (red soft coral)	Cnidaria	Anthozoa	Alcyonacea
<i>Gloeochaete wittrockiana</i>	Glaucophyta	Glaucophyceae	Gloeochaetales
<i>Gloioxanthomyces nitidus</i>	Basidiomycota	Agaricomycetes	Agaricales
<i>Glossolepis ramuensis</i> (Ramu Rainbowfish)	Chordata	Actinopterygii	Atheriniformes
<i>Gloydus monticola</i> (Likiang pitviper)	Chordata	Reptilia	Squamata
<i>Gloydus strauchi</i> (Strauch's pitviper)	Chordata	Reptilia	Squamata
<i>Golovatchia magda</i>	Arthropoda	Diplopoda	Chordeumatida
<i>Gracupica contra</i> (Pied Myna)	Chordata	Aves	Passeriformes
<i>Graphiurus angolensis</i> (Angolan African Dormouse)	Chordata	Mammalia	Rodentia
<i>Gryllotalpa vineae</i> (Vineyard Mole-cricket)	Arthropoda	Insecta	Orthoptera
<i>Guibemantis tasifotsy</i> (White-flanked Malagasy Tree Frog)	Chordata	Amphibia	Anura
<i>Gymnoscopelus braueri</i> (Brauer's lanternfish)	Chordata	Actinopterygii	Myctophiformes
<i>Harmeriella terebrans</i>	Bryozoa	Gymnolaemata	Ctenostomatida
<i>Hazeus elati</i> (Eilat Sandgoby)	Chordata	Actinopterygii	Gobiiformes
<i>Hemidactylus montanus</i> (Mountain Leaf-toed Gecko)	Chordata	Reptilia	Squamata
<i>Hemipteroseius ageneius</i>	Arthropoda	Arachnida	Mesostigmata
<i>Hippopus hippopus</i> (bear paw clam)	Mollusca	Bivalvia	Cardiida

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
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
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 similis (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 fulviventeris (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
Nuquequula flavaxilla (Yellowspotted Ponyfish)	Chordata	Actinopterygii	Perciformes
Obelia geniculata (bell hydroid)	Cnidaria	Hydrozoa	Leptothecata

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
<i>Oligodon melanozonatus</i> (Abor Hills Kukri Snake)	Chordata	Reptilia	Squamata
<i>Oncorhynchus aguabonita</i> (California golden trout)	Chordata	Actinopterygii	Salmoniformes
<i>Ortalis guttata</i> (Speckled Chachalaca)	Chordata	Aves	Galliformes
<i>Oscarella carmela</i> (slime membrane sponge)	Porifera	Homoscleromorpha	Homosclerophorida
<i>Paeonia californica</i> (California peony)	Tracheophyta	Magnoliopsida	Saxifragales
<i>Paramya subovata</i> (subovate softshell)	Mollusca	Bivalvia	Myida
<i>Parasphaerolaimus stagnalis</i>	Nematoda	Chromadorea	Monhysterida
<i>Paraspinibarbus macracanthus</i> (Pocket-like lip barbel)	Chordata	Actinopterygii	Cypriniformes
<i>Paulownia fortunei</i> (dragontree)	Tracheophyta	Magnoliopsida	Lamiales
<i>Penicillago nodositata</i>	Ascomycota	Eurotiomycetes	Eurotiales
<i>Percina squamata</i> (olive darter)	Chordata	Actinopterygii	Perciformes
<i>Petinomys mindanensis</i> (Mindanao Flying Squirrel)	Chordata	Mammalia	Rodentia
<i>Phaeolepiota aurea</i> (Golden Bootleg)	Basidiomycota	Agaricomycetes	Agaricales
<i>Phegopteris excelsior</i> (tall beech fern)	Tracheophyta	Polypodiopsida	Polypodiales
<i>Phellia robusta</i> (robust sea anemone)	Cnidaria	Anthozoa	Actiniaria
<i>Phoradendron rubrum</i> (narrow-leaf mistletoe)	Tracheophyta	Magnoliopsida	Santalales
<i>Phyllomys dasythrix</i> (Drab Atlantic Tree-rat)	Chordata	Mammalia	Rodentia
<i>Pictichromis porphyrea</i> (Magenta Dottyback)	Chordata	Actinopterygii	Perciformes
<i>Pinus teocote</i> (Aztec Pine)	Tracheophyta	Pinopsida	Pinales
<i>Plakina atka</i> (Atka membrane sponge)	Porifera	Homoscleromorpha	Homosclerophorida
<i>Poecilimon fuscii</i> (Fuss' Bright Bush-cricket)	Arthropoda	Insecta	Orthoptera
<i>Pollimyrus adspersus</i> (Adspersus Mormyrid)	Chordata	Actinopterygii	Osteoglossiformes
<i>Polycoccum psorae</i>	Ascomycota	Dothideomycetes	Pleosporales
<i>Pontohedyle kepii</i>	Mollusca	Gastropoda	Acochlidimorpha
<i>Priolepis agrena</i> (Network Reefgoby)	Chordata	Actinopterygii	Gobiiformes
<i>Propithecus coronatus</i> (Crowned Sifaka)	Chordata	Mammalia	Primates
<i>Protalembra brasiliensis</i> (Brazilian leafhopper)	Arthropoda	Insecta	Hemiptera
<i>Pseudagrion estesi</i> (Estes's Sprite)	Arthropoda	Insecta	Odonata
<i>Pseudemys peninsularis</i> (Peninsula Cooter)	Chordata	Reptilia	Testudines
<i>Pseudoastrophaeriella bambusae</i>	Ascomycota	Dothideomycetes	Pleosporales
<i>Pseudochromis springeri</i> (bluestriped dottyback)	Chordata	Actinopterygii	Perciformes
<i>Pseudodactylaria xanthorrhoeae</i>	Ascomycota	Sordariomycetes	Pseudodactylariales
<i>Pseudoschultzeiella cerciata</i> (Mweru Wantipa Bush Grasshopper)	Arthropoda	Insecta	Orthoptera

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
<i>Pteromonas aculeata</i>	Chlorophyta	Chlorophyceae	Chlamydomonadales
<i>Ptilopsis granti</i> (Southern White-faced Owl)	Chordata	Aves	Strigiformes
<i>Raatzbrockmannia boulardi</i>	Arthropoda	Insecta	Hemiptera
<i>Ramajendas frigidus</i>	Tardigrada	Eutardigrada	Parachela
<i>Ranunculus sardous</i> (hairy buttercup)	Tracheophyta	Magnoliopsida	Ranunculales
<i>Rastrognathia macrostoma</i>	Gnathostomulida	Unknown	Bursovaginoidea
<i>Rumex pallidus</i> (large-tubercled dock)	Tracheophyta	Magnoliopsida	Caryophyllales
<i>Sargocentron rubrum</i> (redcoat)	Chordata	Actinopterygii	Holocentriformes
<i>Sciaena umbra</i> (brown meagre)	Chordata	Actinopterygii	Perciformes
<i>Scorpaenodes engleri</i> (Englert's Scorpionfish)	Chordata	Actinopterygii	Scorpaeniformes
<i>Secubelmis limpida</i> (aquatic oligochaete worm)	Annelida	Clitellata	Unknown
<i>Serranobatrachus megalops</i> (San Lorenzo Robber Frog)	Chordata	Amphibia	Anura
<i>Silene haussknechtii</i> (Hausknecht's Catchfly)	Tracheophyta	Magnoliopsida	Caryophyllales
<i>Solenodon paradoxus</i> (Haitian solenodon)	Chordata	Mammalia	Eulipotyphla
<i>Spalangia plaumanni</i>	Arthropoda	Insecta	Hymenoptera
<i>Stenocercus rhodomelas</i> (Red-black Whorltail Iguana)	Chordata	Reptilia	Squamata
<i>Stugeta carpenteri</i> (Carpenter's Sapphire)	Arthropoda	Insecta	Lepidoptera
<i>Stylaster venustus</i> (fuchsia lace hydrocoral)	Cnidaria	Hydrozoa	Anthoathecata
<i>Sweltsa palearata</i> (Shenandoah Sallfly)	Arthropoda	Insecta	Plecoptera
<i>Symphurus rhytisma</i> (patchtail tonguefish)	Chordata	Actinopterygii	Pleuronectiformes
<i>Symposiachrus sacerdotum</i> (Flores Monarch)	Chordata	Aves	Passeriformes
<i>Tateshia naomiae</i>	Mollusca	Gastropoda	Neogastropoda
<i>Taxodium distichum</i> (bald cypress)	Tracheophyta	Pinopsida	Cupressales
<i>Tetraconus mocsaryi</i>	Arthropoda	Insecta	Hymenoptera
<i>Teucrium botrys</i> (cut-leaf germander)	Tracheophyta	Magnoliopsida	Lamiales
<i>Thelenota ananas</i> (Pineapple Sea Cucumber)	Echinodermata	Holothuroidea	Synallactida
<i>Tianella daamsae</i>	Arthropoda	Diplopoda	Chordeumatida
<i>Tortula muralis</i> (Wall Screw-moss)	Bryophyta	Bryopsida	Pottiales
<i>Trachemys scripta</i> (red-eared slider)	Chordata	Reptilia	Testudines
<i>Treron pembaensis</i> (Pemba Green Pigeon)	Chordata	Aves	Columbiformes
<i>Trichoribates columbreti</i>	Arthropoda	Arachnida	Sarcoptiformes
<i>Trichouropoda laevis</i>	Arthropoda	Arachnida	Mesostigmata
<i>Troglodytes troglodytes</i> (Winter Wren)	Chordata	Aves	Passeriformes

Table S1. Organisms Used for Deeper (n = 200) Validation

Organism	Phylum	Class	Order
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 cactusnail)	Mollusca	Gastropoda	Stylommatophora
Yabea microcarpa (California-hedge-parsley)	Tracheophyta	Magnoliopsida	Apiales

Table S2. Pattern Matrix from the 14-Factor Promax-Rotated Item-Factor Analysis.

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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													

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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													
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													

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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													
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												

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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							####						
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												
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													

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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		
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													

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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												
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									

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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					####						
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										
Does this organism have a rigid internal skeleton or structural support system?	-0.31			-0.73										

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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									

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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							

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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							
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?								####						
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					

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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			
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		

Item	Plant-Like	Fungus-Like	Aquatic	Insect-Like	Simple	Robust	Durable	Bird-Like	Dangerous	Shelled	Drought-Tolerant	Parasitic	Cold-Tolerant	Enviro-Sensitive
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

Note. Loadings are rounded to two decimals. Bold values indicate $|\lambda| \geq .50$. Blank cells indicate loadings not displayed in the uploaded sorted factor-analysis output. Abbreviations: PL = Plant-Like; FL = Fungus-Like; Aq = Aquatic; In = Insect-Like; Sim = Simple; Rob = Robust; Dur = Durable; Bird = Bird-Like; Dang = Dangerous; Sh = Shelled; Dr = Drought-Tolerant; Par = Parasitic; Cold = Cold-Tolerant; Env = Environmentally-Sensitive. The Drought-Tolerant header corrects a spelling error in the uploaded spreadsheet.

Table S3. Top five LifeDive items most strongly correlated with each of 44 external validation criteria, drawn from seven public trait databases plus the GBIF taxonomic backbone.

Validation criterion (source database)	n	LifeDive item	r
Is animal (Kingdom Animalia) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	Does this organism have a life stage that is predominantly haploid?	.84
		Can this organism survive being completely dried out and later rehydrate?	.80

Validation criterion (source database)	n	LifeDive item	r
		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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	Does this organism use gills or gill like structures for gas exchange?	.96

Validation criterion (source database)	n	LifeDive item	r
		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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	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) <i>GBIF/CoL backbone</i>	8,427	Does this organism form perennial structures that persist and regrow across multiple years or seasons?	1.00
		Does this organism produce cellulose?	.93

Validation criterion (source database)	n	LifeDive item	r
		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) <i>GBIF/CoL backbone</i>	8,427	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 <i>AmphiBIO</i>	957	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
Amphibian is viviparous <i>AmphiBIO</i>	957	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
Amphibian eats vertebrates <i>AmphiBIO</i>	1,051	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
Bird is nocturnal <i>EltonTraits</i>	2,283	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 <i>EltonTraits</i>	2,283	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 <i>EltonTraits</i>	1,464	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 <i>EltonTraits</i>	1,464	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

Validation criterion (source database)	n	LifeDive item	r
		Is this organism commonly found in polar or Arctic Antarctic regions?	.47
Mammal foraging stratum: aerial <i>EltonTraits</i>	1,464	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 <i>EltonTraits</i>	1,464	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 <i>EltonTraits</i>	1,464	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 <i>EltonTraits</i>	1,464	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 <i>EltonTraits</i>	1,415	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) <i>MammalDIET</i>	1,414	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) <i>MammalDIET</i>	1,415	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) <i>MammalDIET</i>	1,414	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) <i>MammalDIET</i>	1,411	Is this organism capable of trapping or capturing prey?	.64
		Is this organism venomous?	.59

Validation criterion (source database)	n	LifeDive item	r
		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) <i>PanTHERIA</i>	535	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 <i>SeaLifeBase</i>	1,792	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 <i>SeaLifeBase</i>	1,792	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

Note. Correlations are pairwise polychoric, polyserial, biserial, tetrachoric, or Pearson, as appropriate to the measurement type of each variable. The sample size *n* is the number of organisms with non-missing data on the criterion.

Supplementary References

30. Catalogue of Life, *Catalogue of Life checklist (version 2026-04-18 XR)*, dataset (2026); <https://doi.org/10.48580/dgxjw>.
31. xAI, *Batch API*, xAI Docs (2026); <https://docs.x.ai/developers/advanced-api-usage/batch-api> [accessed 21 May 2026].
32. R. Likert, A technique for the measurement of attitudes. *Arch. Psychol.* **22**, 1 (1932).
33. L. Breiman, Random forests. *Mach. Learn.* **45**, 5 (2001).
34. H2O.ai, *h2o.randomForest*, H2O-R package documentation (2026); <https://docs.h2o.ai> [accessed 21 May 2026].
35. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2026); <https://www.R-project.org/>.
36. W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, R package version 2.6.5 (Northwestern Univ., 2026).
37. L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, E. J. Strahan, Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**, 272 (1999).
38. A. E. Hendrickson, P. O. White, Promax: A quick method for rotation to oblique simple structure. *Br. J. Stat. Psychol.* **17**, 65 (1964).
39. J. L. Horn, A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179 (1965).
40. W. F. Velicer, Determining the number of components from the matrix of partial correlations. *Psychometrika* **41**, 321 (1976).
41. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
42. K. Slowikowski, *ggrepel: Automatically Position Non-Overlapping Text Labels with ggplot2*, R package version 0.9.8 (2026).
43. T. L. Pedersen, *patchwork: The Composer of Plots*, R package version 1.3.2 (2025).
44. U. Olsson, Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443 (1979).
45. S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, D. Borsboom, qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, 1 (2012).
46. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129 (1991).
47. N. Pérez-Harguindeguy et al., New handbook for standardised measurement of plant functional traits worldwide. *Aust. J. Bot.* **61**, 167 (2013).
48. R. Villanueva, V. Perricone, G. Fiorito, Cephalopods as predators: A short journey among behavioral flexibilities, adaptations, and feeding habits. *Front. Physiol.* **8**, 598 (2017).
49. S.-H. Lee, D.-S. Kim, S.-H. Park, H. Park, Phytochemistry and applications of *Cinnamomum camphora* essential oils. *Molecules* **27**, 2695 (2022).
50. R. A. S. N. Ranasinghe, S. D. T. Maduwanthi, R. A. U. J. Marapana, Nutritional and health benefits of jackfruit (*Artocarpus heterophyllus* Lam.): A review. *Int. J. Food Sci.* **2019**, 4327183 (2019).
51. L. P. Wilhite, J. R. Toliver, in *Silvics of North America: Volume 1, Conifers*, R. M. Burns, B. H. Honkala, Tech. Coords. (Agriculture Handbook 654, U.S. Department of Agriculture, Forest Service, 1990).
52. C. J. Jackson, A. Reyes-Prieto, The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanopteryx gloeocystis*: Multilocus phylogenetics suggests a monophyletic Archaeplastida. *Genome Biol. Evol.* **6**, 2774 (2014).
53. B. J. Harris et al., Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat. Ecol. Evol.* **6**, 1634 (2022).
54. J. Larraín, A. Fife, C. Atala, *Lorentziella* (Gigaspermaceae, Bryophyta) new for Chile, and lectotypification of the genus. *Bol. Soc. Argent. Bot.* **52**, 351 (2017).
55. A. M. A. Aguinaldo et al., Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489 (1997).
56. I. Cid del Prado-Vera, H. Ferris, S. A. Subbotin, Six new and one known species of *Geomonhystera* (Nematoda, Monhysteridae) from moss, an epiphytic plant and soil in México and Ecuador. *Zootaxa* **4471**, 76 (2018).