Published on Dec 14, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-33b4-4dc5

Leveraging evolution to identify novel organismal models of human biology

Researching just a handful of organisms limits biological discovery. We developed an approach pairing organisms with biological questions to expand research biodiversity.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Erin McGeever, Austin H. Patton, Ryan York

Version 1 · Mar 31, 2025

Purpose

Biomedical research heavily relies on a few "supermodel organisms." Research using these organisms often fails to translate to human biology, limiting progress and clinical success. Recognizing these limitations, there's growing interest in expanding the diversity of research organisms. However, there's, as of yet, no optimal way to pair organisms with biological problems. Depending on the research question, each organism possesses distinct features that can be assets or liabilities. We developed a method to identify organisms best suited to specific problems and applied it to an “organismal portfolio” representing the breadth of eukaryotic diversity. We found that many aspects of human biology could be studied in unexpected species, broadening the potential for new biomedical insights.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.
- All associated **code** is available in this GitHub repository.
- **Data** from this pub, including input proteomes, NovelTree outputs, molecular conservation values, and associated metadata are available on Zenodo.
- For a more conceptual overview of our **organismal selection framework**, read our companion pub, "A data-driven approach to match organisms and research problems **[1]**."
- Check out an **example of this approach in action**, "Rescuing *Chlamydomonas* motility in mutants modeling spermatogenic failure" **[2]**.

Background and goals

Organismal models play a crucial role in biomedical research, shaping what can be discovered, developed, and understood. Research on model organisms has revealed many of the foundational principles of modern biology. Our knowledge of *human* biology has largely stemmed from studies of *non-human* species. Every drug progressing to clinical trials necessitates *in vivo* experimentation, which relies on selecting the appropriate organism for the specific research question.

For most biologists, only a limited number of organisms are typically considered. A select group of "supermodel organisms" such as mice, flies, nematodes, frogs, and zebrafish dominate current research, and their use is increasing **[3]**. Trends in grant proposals **[4]**, publications **[5][6]**, and clinical trials **[7]** indicate a narrowing focus on these specific organisms.

This narrowing of focus might be acceptable if supermodel organisms provided universal biological insights. Unfortunately, they don't. Research findings from these organisms often fail to generalize to other contexts **[8]**. Only 8% of basic research — primarily involving supermodels — translates successfully into clinical settings **[9]**. Additionally, 95% of drug candidates fail during clinical development **[9]**. The drug response profiles observed in common model organisms often don't predict those of

humans **[10]**. In the worst-case scenarios, years of research and millions of dollars may be spent investigating traits unique to a supermodel organism that doesn't apply to humans **[8][11]**.

The limitations of using supermodels in research have long been acknowledged **[3][8][12][13]**, leading to a growing interest in broadening the diversity of organisms used in biomedical studies **[11]**. Inspired by frameworks like Krogh's principle — which suggests that "for a large number of problems, there will be specific animals that can be studied most conveniently" **[14][15]** — researchers are increasingly exploring organisms beyond the traditional supermodels. This shift is facilitated by the availability of generalizable genetic and molecular tools, prompting more biologists to engage with diverse research organisms **[16][17][18]**.

Choosing which of the millions of existing species to study isn't a simple task. While all organisms have their merits for research **[19]**, selecting a species that aligns with a specific question requires careful consideration of various biological, technical, and practical factors **[20]**. Each organism has unique evolutionary traits — some highly conserved, others distinct — that can either aid or hinder research, depending on the question being addressed **[12]**. Research failures often occur when these features are overlooked in the design of biomedical studies. By better understanding the evolutionary histories of these research organisms, we can navigate the potential advantages and challenges they present.

In this study, we developed an evidence-based approach to match research organisms with specific biological problems. We employed novel methods to analyze the evolutionary landscape of an organism's protein-coding genome and identify which genes are most conserved with humans. By applying our method to a diverse portfolio of 63 eukaryotic organisms, we discovered that the similarity in proteins often didn't align with what neutral evolutionary expectations would predict.

Contrary to the "Scala Naturae" model (often called the "great chain of being"), which suggests that complexity increases linearly with similarity to humans, our findings revealed a more complex reality. Many human traits can be found in the eukaryotic tree's unexpected and distantly related branches. This greatly expands the potential avenues for addressing some of biology's most challenging problems.

The approach

Organismal curation

We used publicly available data to curate a portfolio of 63 diverse eukaryotic species. We performed a literature review and surveyed public databases to identify eukaryotes with publicly available proteomes. Since our goal was to identify potential models for human biology, we then determined which species had available tools for genetic perturbations. Finally, we selected species based on taxonomic breadth — ensuring representation from major eukaryotic lineages — and depth, which involved spanning vertebrate and metazoan diversity to facilitate gene family inference. Taxonomic classifications were assigned to each species following the conventions in the EukProt database **[21]**.

> **SHOW ME THE DATA**: The sources and metadata for these species and their proteomes are available on Zenodo (DOI: 10.5281/zenodo.14425432).

Phylogenomic inference

Proteomes were pre-processed by filtering out redundant and short sequences and curating functional annotations (e.g., KEGG annotations) **[22]**. Filtering was executed by a Snakemake workflow, the details of which are described in a previous publication **[22]**. The sample sheet used as input to the Snakemake workflow and the filtered proteomes and intermediate outputs can be found here **[23]**.

We used the filtered proteomes as input to NovelTree (v1.0.2) to infer gene families, multiple sequence alignments, gene family trees, and species trees **[24]**. We ran NovelTree on NextFlow Tower with run-specific parameters specified in the configuration file on Zenodo. We assessed a range of inflation parameters (from 1.25 to 4.5; 0.25 increments) to identify the optimal choice for use with OrthoFinder (v2.5.4) **[25][26]** and cogeqc (v1.2.1) **[27]**. We filtered out gene families that contained fewer than five proteins, represented fewer than five species, and/or were shorter than 30 amino acids in length. We then used WITCH (v0.3.0) **[28]** to perform multiple sequence

alignments and inferred gene family trees using IQ-TREE 2 (v2.2.0.5) **[29]**. We then used Asteroid (v1.0) (git sha: 3aae117) **[30]** and SpeciesRax **[31]** (as implemented in GeneRax (v2.0.4) (git sha: 56f3ed0)) to infer species trees. Species trees were inferred using gene families containing at least 75% of species in the portfolio and had a mean per-species copy number ≤ 10.

Protein physicochemical property calculations

We calculated ten protein physicochemical properties for each protein in our dataset using the ProtParam **[32]** module implemented within Biopython **[33]**. The properties were: 1) molecular weight, 2) aromaticity, 3) instability index, 4) flexibility, 5) GRAVY (grand average of hydrophobicity), 6) isoelectric point, 7) charge at PH 7, 8) helix fraction, 9) sheet fraction, and 10) molar extinction coefficient of cysteines. These protein features were calculated using the genefam_aa_summaries.py script. In addition to the above properties, we also calculated two other GRAVY metrics, four other charges (at PH 3, 5, 9, & 11), turn fraction, the molar extinction coefficient of cystines, and amino acid composition, but given their redundancy with other properties, they weren't used in downstream analyses.

Accounting for evolutionary non-independence

Species' traits (e.g., physicochemical properties) are evolutionarily (and, thus, statistically) non-independent. Closely related species will often have similar traits. This similarity is most likely due to shared ancestry, which, if not accounted for, can mask the signal of biological processes of interest. To control this, we used a phylogenetic transform to identify residual variation not explained by shared evolutionary history (i.e., phylogeny/gene tree) for each physicochemical property.

Specifically, we applied a phylogenetic generalized least-squares (PGLS) **[34]** transformation. PGLS effectively adjusts the observed data to unit variance after correcting for the covariance in traits induced by evolutionary non-independence under Brownian motion. The PGLS transformation assumes elements of the phylogenetic covariance matrix correspond to the amount of time (i.e., branch lengths) from the root of the tree to the common ancestor of each pair of taxa. That is, the phylogenetic tree that's used to conduct the transformation is expected to be time-

calibrated, with branch lengths corresponding to units of time, rather than substitutions-per-site as is common for trees inferred using molecular data as is the case in NovelTree **[24]**. We thus sought to time-calibrate each gene family tree before the application of the transform to the protein physicochemical property data.

We employed a two-step approach that used congruification **[35]**. First, we time-calibrated our species tree, enabling us to time-calibrate each gene family tree. In summary, the congruification method involves mapping divergence times from an existing time tree onto an uncalibrated phylogeny with partially overlapping taxa, followed by rate smoothing to calibrate the divergence times in the target phylogeny. While this method may be less accurate than others, it's highly efficient, making it well-suited for our high-throughput use case, which required the time calibration of 14,067 gene family trees covering 629,320 proteins.

Specifically, we obtained a time-calibrated tree that included 59 of the 64 species in our dataset from timetree.org. We then congruified this tree with the species tree inferred by SpeciesRax using the `congruify.phylo` function in the R-package geiger (v2.0.11) **[36]**. Using the time-calibrated species tree, we subsequently congruified each gene family tree and applied the PGLS transformation to the protein physicochemical property data for each gene family.

The PGLS transformation was implemented in a custom R function, `phylo_gls_transform`. This function uses the `vcvPhylo` function from phytools (v2.1-1) **[37][38]** to obtain the phylogenetic variance-covariance from a species or gene tree. It then calls a custom Rcpp function (`phylo_correction`) to perform the phylogenetic GLS transformation.

Quantification of protein (dis)similarity

Using these transformed protein physicochemical property data, we quantified multivariate Mahalanobis distances between all pairs of proteins within each gene family containing a human homolog. This distance metric accounts for covariances between variables to determine the distance between observations, making it well-suited to complex datasets like ours. However, the calculation of Mahalanobis distances is computationally intensive — a problem that's exacerbated by the high dimensionality of our dataset (10 physicochemical properties) and by the large number of observations among which we needed to compare (9,260 gene families; > 51 million

comparisons in total). Consequently, we developed our own highly efficient, parallelized implementation of its calculation in Rcpp: `pairwise_mahalanobis`.

Phylogenetic visualization and gene family distribution comparison

The time-calibrated SpeciesRax species tree was used for all downstream analyses. The phylogenetic visualization in Figure 1 was generated using the `ggtree` function in the R package ggtree **[39]**. Cophenetic distances of the species tree were calculated using the function `cophenetic.phylo` in the R package ape **[40]**.

We employed a permutation-based method to simulate the number of gene families shared between humans and non-human species, as shown in Figure 2. First, we developed a linear model to predict the number of gene families shared based on the evolutionary distance from humans for each species (using the R function `lm`). We then extracted the predicted values from this model and normalized them by dividing them by the total predicted count. This process provided us with a proportion for each species, allowing us to pose the question: "Given *n* random draws from the set of gene families containing human homologs, how many would we expect to have a homolog belonging to species *x*?"

We created a hypothetical "pool" of proteins to sample from, consisting of 100,000 unique proteins, each representing a different species. The frequency of each protein was determined based on previously calculated expected proportions. Sample sizes were established based on observed gene family sizes, which ranged from four to 45,364 proteins.

For each sample size, we randomly sampled proteins 100 times. For example, when sampling from a gene family size of 10, we randomly selected 10 proteins from the pool and identified the species represented in each sample. This process was repeated 100 times. Finally, we analyzed all permutations to determine the gene family size from which we began sampling proteins across all 63 species.

Describing patterns of molecular (dis)similarity

In a previous section, we explained how we quantified the similarity between human proteins and their non-human homologs within each gene family based on proteins' physicochemical properties. Using a more evolutionarily informed approach, this analysis enables us to identify species that may serve as better model organisms than the traditional "supermodel" species. We recognize that non-human homologs exhibiting a high degree of similarity to their human counterparts are also likely to be functionally similar.

This functional similarity can result from different evolutionary processes: conservation and convergence, or other forms of non-parallel evolution **[41]**. Similarity due to conservation arises from long-term evolutionary stasis, while convergence refers to the independent evolution of similar traits from unrelated common ancestors. Since our primary goal is to identify non-human proteins that likely share functions with their human homologs, we don't attempt to distinguish between these hypotheses in this discussion.

For clarity, we'll refer to protein similarity as molecular conservation throughout the rest of the publication, using our multivariate distance measures to indicate levels of conservation; specifically, smaller distances correspond to more significant conservation.

In Figure 3, B, we compare the distributions of molecular similarity across all gene families. To achieve this, we first characterized the distribution of protein conservation within each gene family by computing a frequency histogram. These histograms were binned in an equivalent way, allowing for a direct comparison of gene families based on their frequency distributions. As a heuristic approach, we applied hierarchical clustering using the R function `hcl`, to illustrate the relationships among gene families based on these binned similarity data.

Next, we investigated how the evolutionary distance from human homologs predicts molecular conservation and how this relationship varies among different gene families (examples can be seen in Figure 5). We conducted a regression analysis of the cophenetic distance from human homologs and molecular conservation for each protein, using the R function `lm`. This analysis identified the homolog most similar to humans for each species. The fitted models and their slopes were then used to illustrate the four examples in the figure.

To better understand and visualize the interaction between evolutionary relatedness and overall patterns of molecular conservation to humans, we constructed a phylomorphospace **[42]** (Figure 6). We first generated a matrix of similarity values, where the columns corresponded to the number of human proteins in the dataset, and the rows represented different species. The matrix was populated as follows: for each species and a specific column (representing a human protein), we identified the homolog in the species most similar to the human protein. If that species lacked a homolog, we used the global maximum conservation value instead. We then applied principal component analysis to create a lower-dimensional embedding of this matrix. The correlation between the principal components and gene family number/phylogenetic distance was assessed using the R function `cor.test`. Finally, we used the first two principal components to create the phylomorphospace with the `phylomorphospace` function from the R package phytools (v2.1-1) **[38]**.

Elo ratings

We quantified per-species conservation enrichment using the Elo rating system **[43]**. Since Elo ratings are sensitive to match order, we used a permutation-based approach that used repeated random starts to ensure robustness, following previous work **[44] [45]**. Matchups were only constructed *within* gene families to control for differences in gene family number across species *and* variation in molecular conservation across gene families.

We first identified all possible matchups within each gene family. All non-human proteins were given a score representing the conservation value of their homolog most similar to any human protein in the gene family. We selected the more conserved protein if a species shared multiple homologs with a given human protein. Furthermore, we only considered gene families with at least 10 possible matchups. When compiled, this resulted in 269,050 possible matchups. Each matchup pitted proteins from two species against each other. The “winner” was the species with the protein most similar to human.

We then constructed 50 series of 10,000 randomly selected matchups. Essentially, each series could be considered a “season” over which 10,000 matchups are played, each containing a different set of matchups. Species that ended each season with a similar Elo rating could be considered robust to matchup order. Species began each season with an Elo rating of 1,500. Ratings were updated after each match using the

`elo.cal` function from the R package elo **[46]**. We then averaged across all seasons to get a mean Elo rating for each species. The relative probabilities of the mean Elo ratings were compared using the function `elo.prob`. Species mean Elo ratings were compared to the number of gene families shared with humans using a linear model implemented by the R function `lm`. Two-way comparisons of mean Elo ratings were done with a Kruskal-Wallis test using the function `kruskal.test` in R.

> All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.14479310)

Additional methods

We used Grammarly Premium to suggest wording ideas, reorganize text using a template, and help clarify and streamline text that we wrote. We also used ChatGPT to help write code and comment our code.

The results

Mapping over 1 billion years of molecular evolution

Genomes aren't singular units. Genomes are configurations of the tangled paths a set of genes has taken. These paths involve gain, loss, duplication, change, and/or re-purposement **[24][47]**. Given this complexity, the relationships inferred between any two genomes (and species) will depend on which genes are considered. For example, genes that share a common ancestor will often possess similar sequences (i.e., they're *homologous*) **[47]**. However, similar sequences sometimes arise independently in distantly related species (i.e., they're *convergent*). If only convergent genes were considered, one might (wrongly) conclude that these two species are closely related. While this genome complexity presents challenges in some situations (such as phylogenetic inference), it may be a boon in others.

If genomes *were* singular units, the answer to “Which organism is best for modeling disease X, Y, or Z?” would always be the same (and likely always be “mice”). Yet, like all other genomes, the human genome is a mixture of evolutionary histories **[48][49]**. Some genes have been gained, lost, or duplicated **[50]**. Others are conserved to varying degrees; some are shared with the last universal common ancestor, and others with animals, vertebrates, mammals, or primates **[48]**. Some have evolved convergently.

What’s more, these patterns aren’t unique to humans. The genomes of popular organismal models are also complex amalgamations. For example, mice have evolved unique immune **[51]**, metabolic, and life history characteristics **[8]**. This all means that, from a genetic perspective, there's no single best organismal model for all aspects of human biology. Instead, an *organismal portfolio* is needed.

The evolutionary history of genes can guide the design of such a portfolio. Deeply conserved genes open up the possibility of studying more tractable yet distantly related species. More recently conserved genes will make closely related species better choices. However, in some cases, these close relatives may be on divergent evolutionary paths, leading them to lack traits relevant to an aspect of human biology. Convergent genes can only be studied in organisms where they've evolved, offering challenges (those species must be identified) and opportunities (they're likely to share important aspects of the relevant biology). Genes specific to humans will require very different modeling approaches since they lack naturally occurring analogs. Capturing these diverse patterns involves the reconstruction of each gene’s evolutionary history.

We set out to build a eukaryotic organismal portfolio for human biology. We selected 63 species as candidate models (see Approach for inclusion criteria). These species had a last common ancestor over one billion years ago and represent many eukaryotic lineages (Figure 1). They span the uni- to multicellular transition, live in most of Earth’s major biomes, and implement various life history strategies. Some are parasitic; some are photosynthetic. Some are endosymbiotic; others filter feed in the oceans’ pelagic zones. There are well-established supermodels (mice, zebrafish, *C. elegans*, *D. melanogaster*, *S. cerevisiae*) and comparatively understudied protists (e.g., Euglenozoa, Percolozoa, and the hyper-diverse TSAR clade).

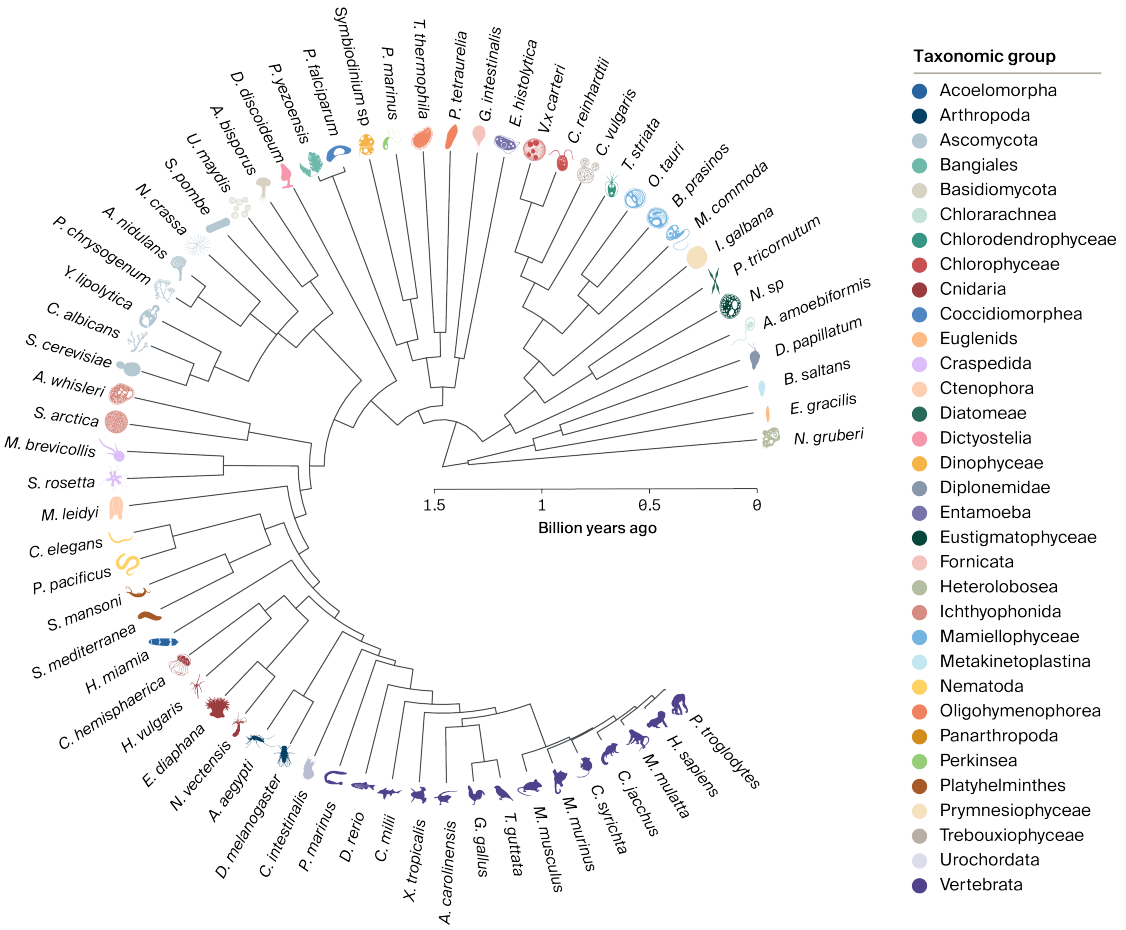

Figure 1

A eukaryotic organismal portfolio.

Time-calibrated species phylogeny created with SpeciesRax. Taxonomic groups correspond to taxogroup1 described by EukProt.

We used the NovelTree workflow **[24]** to infer gene families and evolutionary relationships (i.e., phylogenies) among proteins within each gene family and among species, incorporating information across gene families. After filtering, we identified 9,260 human-containing gene families, encompassing 17,644 human proteins (see Approach for filtering details). The taxonomic distribution of these gene families approximated evolutionary relationships; the more related a species was to humans, the more gene families were shared between them (Figure 2, A). For example, vertebrates possessed twice the number of gene families than non-vertebrates on average (vertebrates = 7,996, non-vertebrates = 3,075; $p = 6.73 \times 10^{-8}$, Kruskal-Wallis test). Chimpanzees were associated with the most gene families (*Pan troglodytes*;

9,158 gene families), while the Ichtheosporean *Abeoforma whisleri* was associated with the least (1,217 gene families). Intriguingly, they also suggest that even the least represented species within the portfolio had a roughly 1:9 (1,217/9,260 gene families) chance of being a potential model candidate. The portfolio, therefore, empowers us to identify organismal models across the phylogenetic breadth of eukaryotes.

We were next interested in assessing our sensitivity for discriminating between candidate models. Variation in the presence/absence of gene families would strongly decay with phylogenetic distance, meaning that related species might differ little in the genes they share with humans. This would be a scenario in which organismal selection might be straightforward (albeit a bit boring): species more closely related to humans will always be favored as model organisms. On the other hand, we might observe substantial variation in species' molecular conservation with humans. In this "high-sensitivity" scenario, the species favored as model organisms will be more variable, necessitating a more involved and nuanced species selection process. Because each gene family would show a different conservation pattern, other aspects of natural history and evolutionary biology could be leveraged to pinpoint an organismal model.

As predicted by such a scenario, we found that gene family presence varied substantially within and across phylogenetic scales (Figure 2, A). For example, the anemone *Exaiptasia diaphana* shared more gene families with humans (5,663) than the early-branching vertebrate *Petromyzon marinus* (sea lamprey; 4,618) despite the latter being more closely related to humans. Furthermore, the even more distant ctenophore *Mnemiopsis leidyi* was about evenly matched with the lamprey (4,583 gene families). This variation was also present at greater phylogenetic distances. The unicellular algae *Chlamydomonas reinhardtii* shared more gene families with humans than similarly distant species (such as the parasite *Giardia intestinalis*) (Figure 2, A). These patterns indicate substantial variation in gene family presence/absence across evolutionary scales within the portfolio, even among the most distant species.

These individual examples were also reflected at global taxonomic scales. The counts of unique species within gene family swiftly increased with total gene count (Figure 2, B) and significantly faster than expected in a simulated low-sensitivity scenario (i.e., where the number of gene families shared with humans linearly decays with evolutionary distance) (Figure 2, B; permutation-based sampling, see Approach). The smallest gene family representing all 63 species contained 70 genes. The equivalent measure in the simulated data was almost four times greater (264 genes). The relationship between the count of unique species within a gene family and that gene

family's age (i.e., time to the most recent common ancestor of all gene copies) revealed diverse species combinations across all sizes (Figure 2, C). The age of gene families increased linearly to ~20 species, after which the relationship plateaued (Figure 2, C).

Interestingly, gene families with as few as five species spanned the full evolutionary range of the portfolio, meaning these small gene families contained everything from the most closely related species to those most distantly related in our dataset (Figure 2, C). For example, gene family OG0013524 (human protein A6NEQ0) contained proteins from primates (humans, macaques, chimpanzees, marmosets) and the unicellular Euglenozoan *Bodo saltans*. These observations make clear that our portfolio is thus both broad — encompassing much of eukaryotic diversity — *and* sensitive, allowing for targeted and flexible selection of research organisms.

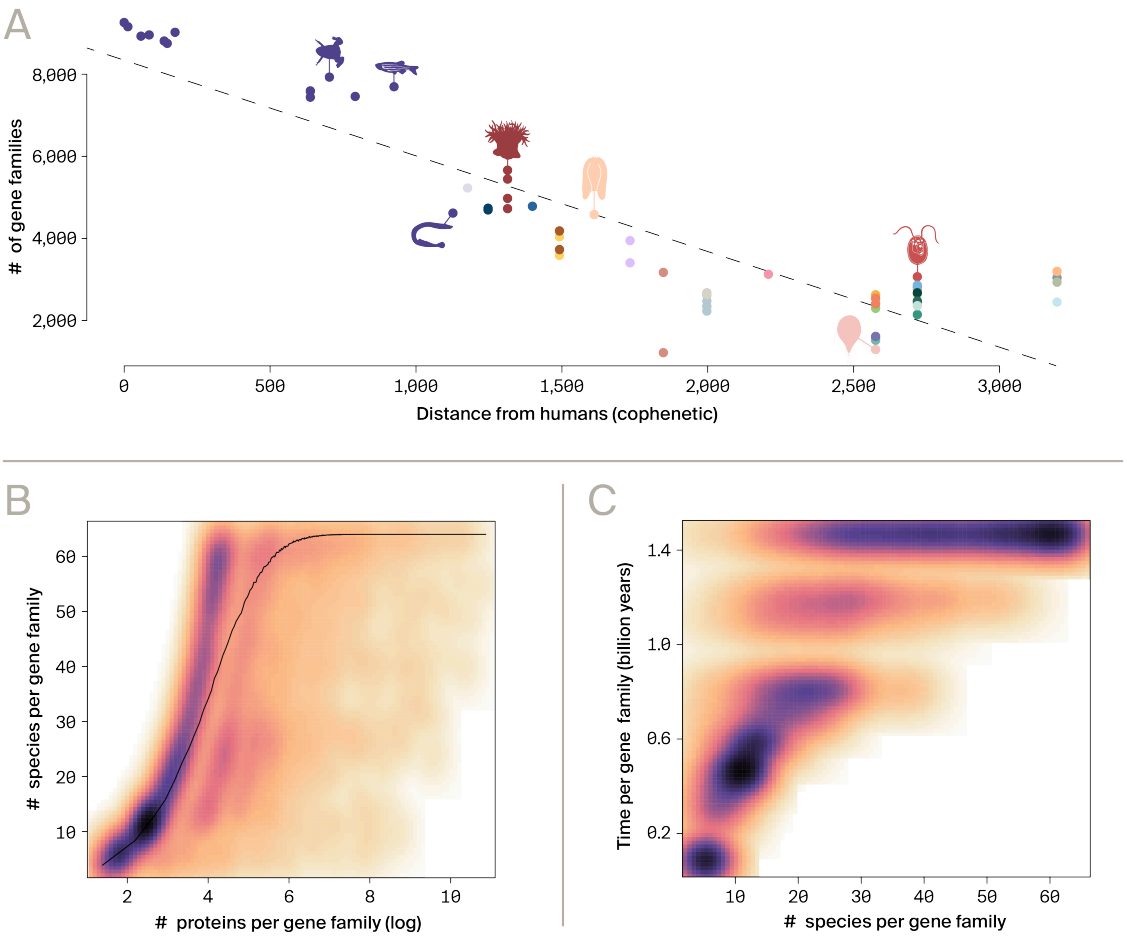

Figure 2

Evolutionary distribution of human gene families.

(A) Number of gene families shared with humans as a function of cophenetic distance from humans. Labeled organisms are (from left to right): *Xenopus tropicalis*, *Danio rerio, Petromyzon marinus*, *Exaiptasia diaphana*, *Mnemiopsis leidyi*, *Giardia intestinalis*, and *Chlamydomonas reinhardtii.*

(B) Density scatter plot comparing protein (x-axis) and species number (y-axis) across gene families. As estimated by simulations, the expected relationship between these values is denoted by the black line.

(C) Density scatter plot of species number (x-axis) and all gene families' evolutionary scale (y-axis).

A novel measure of molecular similarity

Next, we turned our attention to measuring the similarity of molecular properties of the proteins encoded by each gene with their corresponding human homologs. Conservation is commonly inferred by sequence similarity; the more shared a sequence is, the more similar two genes or proteins are presumed to be **[17]**. We wanted to address the limitations of this approach. For one, sequence similarity doesn't always mean functional similarity. It's possible to have two proteins with low overall sequence similarity but share critical portions determining structure and function. In other words, not all portions of a sequence are the same. Sequences are also tied up with species' relatedness. More closely related species will, on average, necessarily have more similar and shared sequences than more distantly related species. This can make it hard to detect cases wherein very distantly related species share sequences that perform the same function through conservation, convergence, or other evolutionary means. Given our portfolio's massive range of evolutionary diversity, we concluded that relying on sequence similarity alone wouldn't cut it.

To address the insufficiency of sequence similarity for our purposes, we developed a novel molecular conservation measure incorporating phylogenetic and protein physicochemical properties (see Approach for details; Figure 3). First, various physicochemical measures and secondary structural properties are calculated from the amino acid sequences of all proteins in a gene family (Figure 3, step 1). As previously described, however, proteins are evolutionarily (and thus statistically) non-independent of one another. To account for this non-independence, we adjusted each measure for evolutionary relatedness using a phylogenetic generalized least squares transformation (PGLS transform; Figure 3, step 2) rendering each protein statistically independent. Using these adjusted protein features, we quantified all pairwise (dis)similarities among proteins within each gene family using Mahalonobis distances (Figure 3, steps 3–4). Last, the distance from the closest human protein was identified for each protein, resulting in our final conservation measure (Figure 3, step 5).

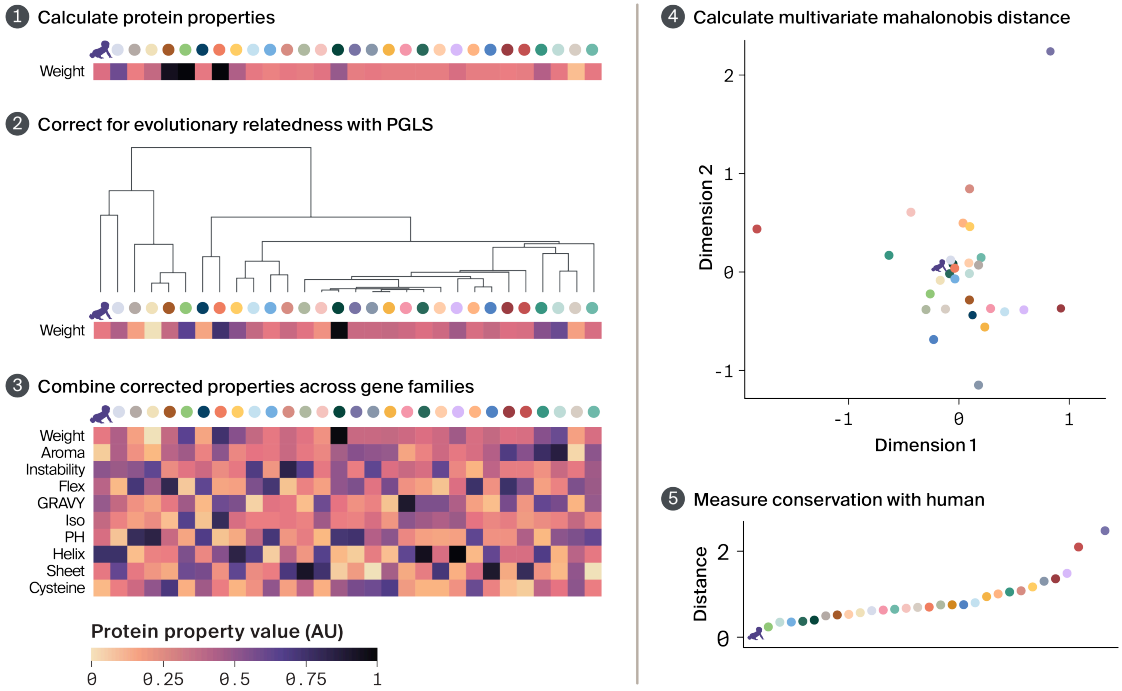

Figure 3

Calculating molecular conservation.

1) Heatmap of one protein physicochemical property. Here, molecular weight ("Weight") is an example. The colored points represent individual species. Colors correspond to the EukProt taxogroup1 (the purple infant cartoon indicates human). Each species' molecular weight is represented by color intensity.

2) We use a phylogenetic generalized least squares (PGLS) transformation to correct for evolutionary relatedness, rendering proteins statistically independent. The heatmap in this panel reflects molecular weight after this correction.

3) Cartoon of the combined matrix of 10 evolutionarily corrected physicochemical properties (naming key: "Weight" = molecular weight, "Aroma" = aromaticity, "Instability" = instability index, "Flex" = flexibility, "GRAVY" = GRAVY index, "Iso" = isoelectric point, "PH" = charge at PH 7, "Helix" = helix fraction, "Sheet" = sheet fraction, "Cysteine" = molar extinction coefficient of cysteines).

4) Cartoon 2-dimensional space representing the Mahalonobis distances measured between species' proteins.

5) Ranked distribution of distances from the human versions for all proteins considered.

Conservation with human homologs wasn't uniformly distributed across species (Figure 4). Gene families differed extensively in their distributions' shape, dynamic range, and magnitude (Figure 4), with many containing genes spanning the full range of conservation (Figure 4). Some were similar to humans, with little evolutionary variation (Figure 4), while others were uniformly distant (Figure 4). These observations reinforce that genomes aren't evolutionarily singular units.

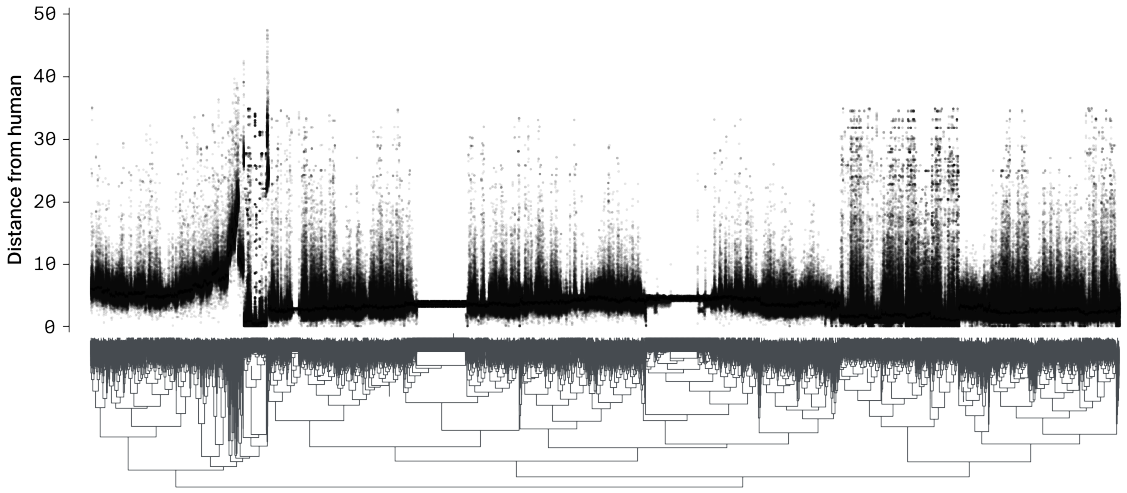

Figure 4

Landscape of molecular conservation between eukaryotes and humans.

Hierarchical clustering of gene families according to conservation patterns with humans across species in our portfolio. Each point corresponds to an individual protein. Conservation is measured using the multivariate distance metric described in Figure 3.

The distribution of conservation to individual human proteins further supports this observation, as shown in Figure 5. For example, PTN4 (UniProt: P29074) is a neurally associated phosphatase that matches evolutionary expectations under a molecular clock hypothesis; molecular conservation to this protein decreases linearly with evolutionary distance (Figure 5, A). The transcription factor FOXA1 (UniProt: P55317) also shows this pattern but, unlike PTN4, is generally not highly conserved (Figure 5, B).

In contrast, conservation to proteins such as ARF3 (UniProt: P61204) — an ADP-ribosylation factor — is uniformly high across the portfolio (mean conservation = 0.88, slope = $2.78e^{-05}$, r^2 = 0.09) (Figure 5, C). Finally, and intriguingly, molecular and evolutionary distance can display a negative relationship (i.e., more distantly related proteins are increasingly similar), as is the case for mitochondrial protein 3HIDH (UniProt: P31937; Figure 5, D). The observed variation of conservation profiles can refine our evolutionary hypotheses and help identify and take advantage of even counterintuitive patterns. It also underlines the importance of questioning Scala Naturae thinking in organismal selection for biomedical research.

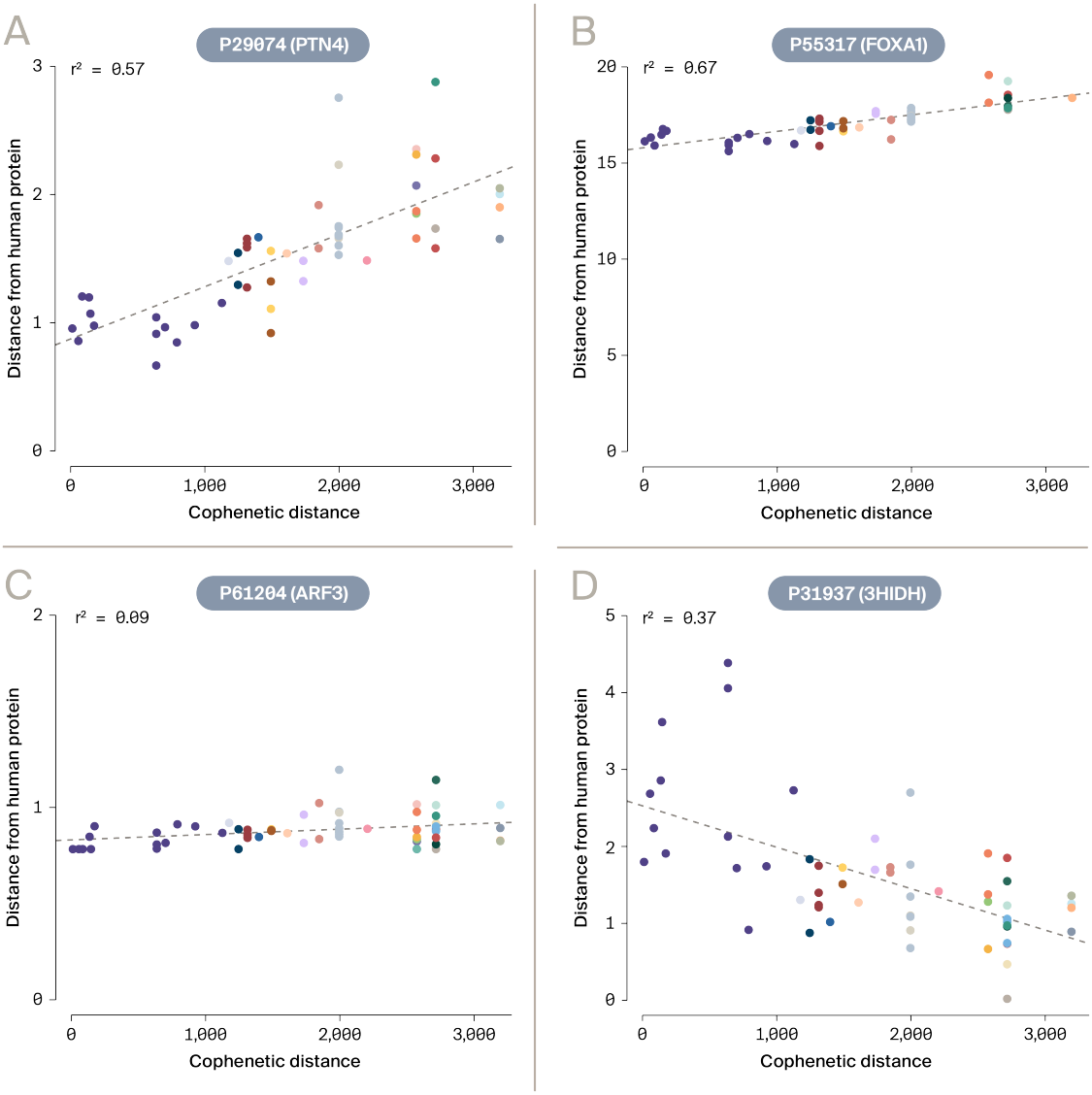

Figure 5

The diversity of conservation profiles.

Human proteins are characterized by the relationship between conservation ("Distance from human protein") and phylogenetic distance from humans ("Cophenetic distance"). Examples include proteins where similarity linearly decreases (A; PTN4, B; FOXA1; C; ARF3), is uniformly highly divergent (B) or deeply conserved (C), or even increases with phylogenetic distance (D; 3HIDH). r^2 = linear regression fit.

De novo identification of supermodel organisms

Our approach was founded on the idea that genome-wide conservation with humans can link potential organismal models with various aspects of human biology. By leveraging this idea, we posited that we could develop an organismal portfolio for each biological question by characterizing these connections. Just how specific might these portfolios be? As we saw above, individual gene family's evolutionary histories vary broadly. Whether or not these patterns translate to organismal-level differences is presently unclear. Are certain organisms disproportionately suited to modeling diverse aspects of human biology? If yes, then "general purpose" organismal models may be developed, potentially simplifying the model selection process. We sought to test this hypothesis.

To begin doing so, we first explored the extent to which evolutionary relationships predict genome-wide conservation patterns. Each species was characterized by a numerical vector containing binary (i.e., presence/absence) and continuous (i.e., molecular conservation) representations of conservation with all human proteins in the dataset. We assessed the relationships between these genome-wide conservation patterns using principal component analysis (PCA) (Figure 6, A). PC1 was significantly correlated with homolog presence/absence ($r = -0.98$; $p = 5.70 \times 10^{-46}$; Pearson correlation) and phylogenetic distance ($r = 0.92$; $p = 1.75 \times 10^{-26}$; Pearson correlation) and explained 45.89% of the observed variance. Projecting the species phylogeny onto PC space further highlighted these relationships (Figure 6, A). We found a clear phylogenetic path through the first two PC axes (Figure 6, A). Notably, of all the PCs ($N = 63$), only PC1 displayed significant correlations with ortholog presence/absence and phylogenetic distance (not shown). This means that most genome-wide conservation variation isn't captured by ortholog presence/absence and can't be directly predicted from phylogenetic relationships. Instead, the (more complex) patterns of protein conservation across each species' proteome must be considered.

Given these observations, we next sought to characterize the conservation profiles of each species' orthologs. We wanted to know if a given species' proteins were consistently more conserved with their human counterparts than expected. We needed a method robust to the uneven representation of species within our dataset; this led to identifying the Elo rating system as a candidate framework **[43]**. Developed initially to rate chess players, Elo ratings assess players' relative skills across a series of "matches" in a zero-sum framework. The Elo system is increasingly used to evaluate machine learning model performance **[45]**, and ratings have been used to identify

species-level biases on protein language model likelihoods **[44]**. Influenced by this work, we developed a permutation-based approach for assessing relative enrichment for conservation to human proteins for each species using Elo ratings (see Approach).

Elo ratings exhibited a range of variability across trials within and across species after summarizing across trials (Figure 6, B–C). In our implementation, scores greater than 1500 represented doing “better” than random. Similarly, scores less than 1500 are “worse” than random. Chimpanzees had the highest rating (mean Elo rating = 1618) whereas (as with gene family number) *Abeoforma whisleri* ranked last (mean Elo rating = 1414), meaning that chimpanzee proteins were more similar to human homologs 76.4% of the time (Figure 2). Vertebrate species, except for lamprey (*Petromyzon marinus*), had scores above 1500 and a median rating of 1571. Non-vertebrates had a median rating of 1478. Overall, ratings generally decreased with phylogenetic distance from humans (Figure 6, C). These expected evolutionary signals provided confidence in using Elo ratings for this task.

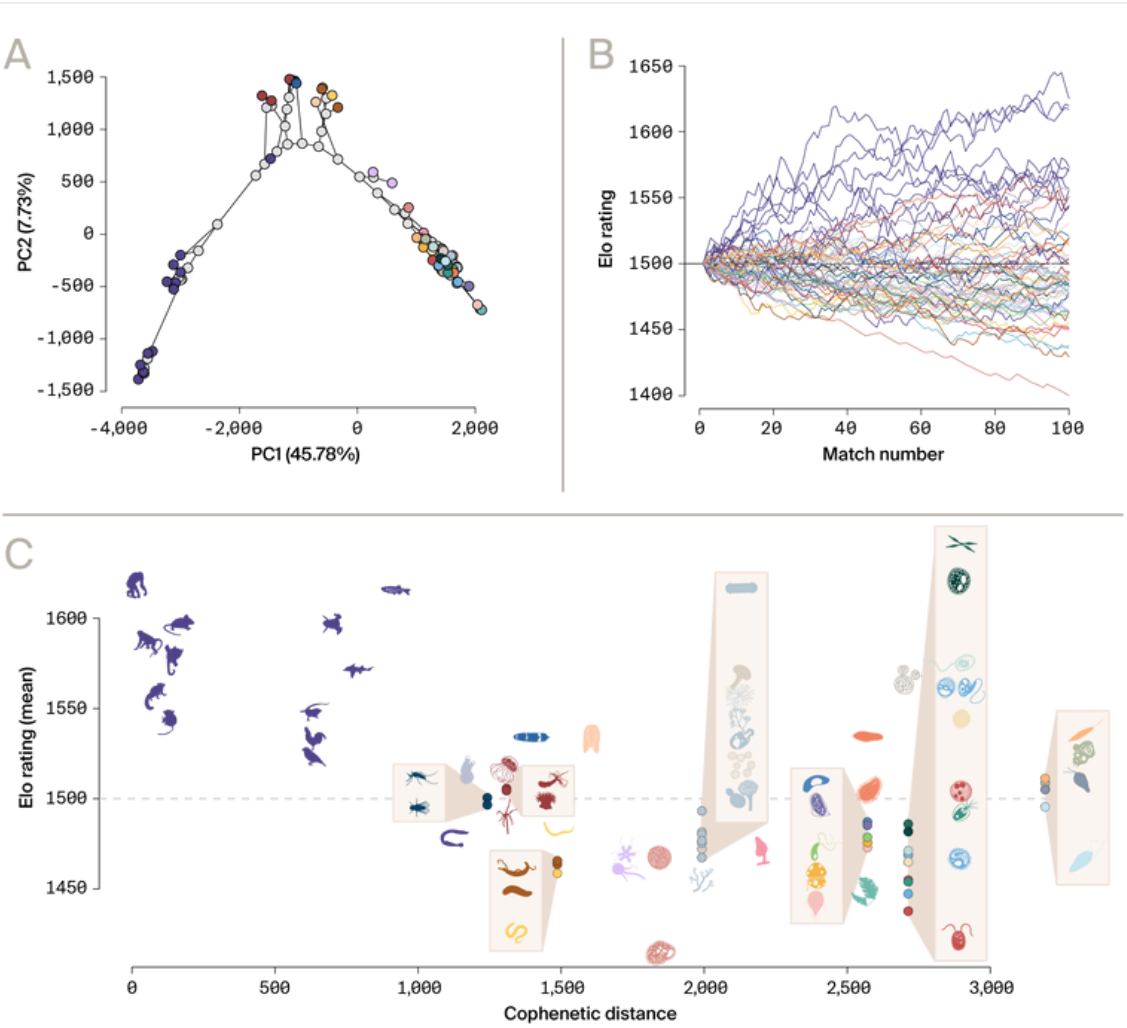

Figure 6

Using Elo ratings to rank research organisms.

(A) Phylomorphospace obtained using conservation to humans and gene family presence/absence for each species as measured across all 9,260 human gene families in our dataset. Percent values correspond to variation explained by each PC. Each point is a species, colored by taxonomic grouping.

(B) Example of Elo rating changes over a series of matchups (each line corresponds to a species). All species start with a rating of 1500, marked by the dotted line.

(C) Distribution of mean Elo ratings as a function of phylogenetic distance from human.

Elo ratings weren't linearly predicted by phylogenetic distances, exhibiting substantial variation at different taxonomic depths. Several outlying species could be readily identified (Figure 6, C). For instance, Zebrafish (*Danio rerio*) beat out primates and mammals to obtain the second-highest rating (Elo rating = 1615), just behind chimpanzees (Elo rating = 1618). Proteins from the unicellular algae *Chlorella vulgaris* (Elo rating = 1564) were 67.5% more likely to be conserved with humans than the closely related species *Chlamydomonas reinhardtii* (Elo rating = 1437). Although vertebrates possessed significantly larger Elo ratings than other taxa ($p = 7.72 \times 10^{-7}$; Kruskal-Wallis test), non-vertebrate multicellular species were indistinguishable from unicellular species ($p = 0.74$; Kruskal-Wallis test). Furthermore, the four most phylogenetically distant species from humans (*Bodo saltans*, *Diplonema papillatum*, *Euglena gracilis*, *Nageleria gruberi*) possessed Elo ratings comparable to invertebrates that arose hundreds of millions of years later ($p = 0.57$; Kruskal-Wallis test).

How unexpected are these patterns? To explore this, we performed a regression predicting Elo rating with variation in the count of human gene families in which each species was present. The model had a reasonably good fit (multiple $R^2 = 0.66$; $p = 4.49 \times 10^{-16}$), as might be expected given the presence of phylogenetic signal in both the Elo ratings and the counts of human gene families. However, we were interested in what *wasn't* described by the model, reasoning that species with exceptional molecular conservation would be associated with positive residual variance (i.e., Elo ratings higher than predicted by this null model).

Exceptional molecular conservation was observed across a wide range of eukaryotic diversity. Notable examples included *Chlorella vulgaris* (3.56; Studentized residual), *Paramecium tetraurelia* (2.42), zebrafish (2.35), chimpanzees (1.53), the frog *Xenopus tropicalis* (1.41), the ciliate *Tetrahymena thermophila* (1.05), the amoeba *Naegleria gruberi* (1.03), the malaria-causing parasite *Plasmodium falciparum* (1.03), the unicellular algae *Euglena gracilis* (0.96), and the ctenophore *Mnemiopsis leidyi* (0.95) (Figure 5, C). Interestingly, some well-studied model organisms exhibited less molecular conservation than anticipated. Nematodes (*Caenorhabditis elegans*) displayed a negative residual of −0.56, fruit flies (*Drosophila melanogaster*) had −0.49, and brewer's yeast (*Saccharomyces cerevisiae*) showed −0.12 (Figure 5, C).

At higher taxonomic levels, consistent patterns emerged. Heterotrophic and parasitic protists were notably enriched, including *Ciliophora* (1.73), *Heterolobosea* (1.03), *and Apicomplexa* (10.3). Fungi aligned with expectations, showing a result of 0.005, while

taxa representing the transition from unicellular to multicellular organisms, *such as Choanoflagellata* (–1.16) *and Ictheosporea* (–1.09), were underrepresented (Figure 5, C).

These observations lead us to conclude that the landscape of genomic conservation is complex and can't be easily predicted by evolutionary relationships alone. Additionally, Elo rating distributions may provide insights into the breadth of human biology that can be modeled using specific research organisms.

Key takeaways

Every species represents a combination of various evolutionary paths, making it difficult to predict which organisms will serve as effective models for understanding human biology. However, by examining the evolutionary context of a species' genome, we can make informed assumptions about the biological insights we might gain from studying that species.

We developed an approach to map the similarities between human genes and those of 63 eukaryotic research organisms. We identified a range of potential model organisms for each gene by analyzing conservation profiles across the human genome. Many of these profiles highlighted species that aren't typically supermodel organisms. Additionally, through global conservation analyses, we pinpointed species that share remarkable molecular similarities with humans based on their phylogenetic positions. Our findings revealed organisms throughout the eukaryotic tree that could serve as valuable model systems, expanding the range of possible organismal models in biomedical research. This approach allows researchers to test their assumptions regarding potential models and provides an evidence base that can free biologists from reliance on conventional wisdom.

Next steps

Experimental validation of our predictions is of great interest. We have begun using the conservation profiles of human genes to identify novel organismal models for genetic diseases. An example of our work can be found in a companion publication **[2]**, where we identified *Chlamydomonas reinhardtii* as a potential model for studying human spermatogenic failure. Through a small-scale drug screen, we demonstrated that the phenotypic effects of two human risk genes — *SPEF2* and *DNALI1* — are conserved,

supporting our evolutionary hypotheses. In the future, we'll focus on validating additional predictions and leveraging our approach to discover new research organisms for genetic and therapeutic explorations.

There are several potential computational extensions we could pursue. The findings in this publication primarily addressed the evolutionary patterns of single genes. A logical next step is to explore gene sets (e.g., molecular pathways, pairwise interactors, and polygenic disease targets) to enhance our ability to predict complex phenotypic conservation in research organisms. This could facilitate the development of innovative phylogenetic methods for comparing the evolution of genetic pathways. Additionally, it could help us generalize our approaches to other biological applications beyond human disease modeling.

Increasing the number of species analyzed would improve our coverage of eukaryotic diversity and enhance the precision of our predictions. An intriguing extension could involve creating a comprehensive organismal portfolio. By predicting more complex biological features across a broader range of species, we could outline a roadmap for biomedical research that effectively pairs specific problems with suitable organismal models and research designs. Even if achieving this goal proves challenging, working towards it should enhance our chances of identifying fundamental biological principles and determining where they can be most effectively applied.

References

1. Avasthi P, York R. (2024). A data-driven approach to match organisms and research problems. https://doi.org/10.57844/ARCADIA-48B0-607A
2. Essock-Burns T, Lane R, MacQuarrie CD, Mets DG. (2024). Rescuing Chlamydomonas motility in mutants modeling spermatogenic failure. https://doi.org/10.57844/ARCADIA-FE2A-711E
3. Alfred J, Baldwin IT. (2015) The Natural History of Model Organisms: New opportunities at the wild frontier. http://doi.org/10.7554/eLife.06956.001

4 Lauer M. 2016. A Look at NIH Support for Model Organisms, Part Two. https://nexus.od.nih.gov/all/2016/08/03/model-organisms-part-two/

5 Dietrich MR, Ankeny RA, Chen PM. (2014). Publication Trends in Model Organism Research. https://doi.org/10.1534/genetics.114.169714

6 Aitman T, Dhillon P, Geurts AM. (2016). A RATional choice for translational research? https://doi.org/10.1242/dmm.027706

7 Soufizadeh P, Mansouri V, Ahmadbeigi N. (2024). A review of animal models utilized in preclinical studies of approved gene therapy products: trends and insights. https://doi.org/10.1186/s42826-024-00195-6

8 Perlman RL. (2016). Mouse Models of Human Disease: An Evolutionary Perspective. https://doi.org/10.1093/emph/eow014

9 Arrowsmith J. (2012). A decade of change. https://doi.org/10.1038/nrd3630

10 Atkins JT, George GC, Hess K, Marcelo-Lewis KL, Yuan Y, Borthakur G, Khozin S, LoRusso P, Hong DS. (2020). Pre-clinical animal models are poor predictors of human toxicities in phase 1 oncology clinical trials. https://doi.org/10.1038/s41416-020-01033-x

11 Sánchez Alvarado A. (2018). To solve old problems, study new research organisms. https://doi.org/10.1016/j.ydbio.2017.09.018

12 Bolker J. (2012). There's more to life than rats and flies. https://doi.org/10.1038/491031a

13 Schnabel J. (2008). Neuroscience: Standard model. https://doi.org/10.1038/454682a

14 Krogh A. (1929). THE PROGRESS OF PHYSIOLOGY. https://doi.org/10.1152/ajplegacy.1929.90.2.243

15 Krebs HA. (1975). The August Krogh principle: "For many problems there is an animal on which it can be most conveniently studied." https://doi.org/10.1002/jez.1401940115

16 Russell JJ, Theriot JA, Sood P, Marshall WF, Landweber LF, Fritz-Laylin L, Polka JK, Oliferenko S, Gerbich T, Gladfelter A, Umen J, Bezanilla M, Lancaster MA, He S, Gibson MC, Goldstein B, Tanaka EM, Hu C-K, Brunet A. (2017). Non-model model organisms. https://doi.org/10.1186/s12915-017-0391-5

17 Yamamoto S, Kanca O, Wangler MF, Bellen HJ. (2023). Integrating non-mammalian model organisms in the diagnosis of rare genetic diseases in humans. https://doi.org/10.1038/s41576-023-00633-6

18 Goldstein B, King N. (2016). The Future of Cell Biology: Emerging Model Organisms. https://doi.org/10.1016/j.tcb.2016.08.005

19 Clark CJ, Hutchinson JR, Garland T Jr. (2023). The Inverse Krogh Principle: All Organisms Are Worthy of Study. https://doi.org/10.1086/721620

20 Dietrich MR, Ankeny RA, Crowe N, Green S, Leonelli S. (2020). How to choose your research organism. https://doi.org/10.1016/j.shpsc.2019.101227

21 Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C. (2022). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. https://doi.org/10.24072/pcjournal.173

22 Celebi FM, Chou S, McDaniel EA, Reiter T, Weiss ECP. (2024). Predicted genes from the Amblyomma americanum draft genome assembly. https://doi.org/10.57844/ARCADIA-9602-3351

23 York R, Patton A. (2024). Leveraging evolution to identify novel organismal models of human biology. https://doi.org/10.5281/ZENODO.14425432

24 Celebi FM, Chou S, McGeever E, Patton AH, York R. (2024). NovelTree: Highly parallelized phylogenomic inference. https://doi.org/10.57844/ARCADIA-Z08X-V798

25 Emms DM, Kelly S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. https://doi.org/10.1186/s13059-019-1832-y

26 Emms DM, Kelly S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. https://doi.org/10.1186/s13059-015-0721-2

27 Almeida-Silva F, Van de Peer Y. (2023). Assessing the quality of comparative genomics data and results with the *cogeqc* R/Bioconductor package. https://doi.org/10.1111/2041-210x.14243

28 Shen C, Park M, Warnow T. (2022). WITCH: Improved Multiple Sequence Alignment Through Weighted Consensus Hidden Markov Model Alignment. https://doi.org/10.1089/cmb.2021.0585

29 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. https://doi.org/10.1093/molbev/msaa015

30 Morel B, Williams TA, Stamatakis A. (2022). Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. https://doi.org/10.1093/bioinformatics/btac832

31 Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A. (2022). SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. https://doi.org/10.1093/molbev/msab365

32 Walker JM, editor. (2005). The Proteomics Protocols Handbook. https://doi.org/10.1385/1592598900

33 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. https://doi.org/10.1093/bioinformatics/btp163

34 Butler MA, Schoener TW, Losos JB. (2000). THE RELATIONSHIP BETWEEN SEXUAL SIZE DIMORPHISM AND HABITAT USE IN GREATER ANTILLEAN*ANOLIS*LIZARDS. https://doi.org/10.1111/j.0014-3820.2000.tb00026.x

35 Eastman JM, Harmon LJ, Tank DC. (2013). Congruification: support for time scaling large phylogenetic trees. https://doi.org/10.1111/2041-210x.12051

36 Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. (2014). geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. https://doi.org/10.1093/bioinformatics/btu181

37 Revell LJ. (2011). phytools: an R package for phylogenetic comparative biology (and other things). https://doi.org/10.1111/j.2041-210x.2011.00169.x

38 Revell LJ. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). https://doi.org/10.7717/peerj.16505

39 Yu G, Smith DK, Zhu H, Guan Y, Lam TT. (2016). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. https://doi.org/10.1111/2041-210x.12628

40 Paradis E, Schliep K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. https://doi.org/10.1093/bioinformatics/bty633

41 Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE. (2018). (Non)Parallel Evolution. https://doi.org/10.1146/annurev-ecolsys-110617-062240

42 Sidlauskas B. (2008). CONTINUOUS AND ARRESTED MORPHOLOGICAL DIVERSIFICATION IN SISTER CLADES OF CHARACIFORM FISHES: A PHYLOMORPHOSPACE APPROACH. https://doi.org/10.1111/j.1558-5646.2008.00519.x

43 Elo A. (1978). The Rating of Chessplayers, Past and Present. New York.

44 Ding F, Steinhardt J. (2024). Protein language models are biased by unequal sequence sampling across the tree of life. https://doi.org/10.1101/2024.03.07.584001

45 Boubdir M, Kim E, Ermis B, Hooker S, Fadaee M. (2023). Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. https://doi.org/10.48550/ARXIV.2311.17295

46 Heinzen E. elo. (2017). https://github.com/eheinzen/elo

47 Eisen JA. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. https://doi.org/10.1101/gr.8.3.163

48 Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. (2007). The human phylome. https://doi.org/10.1186/gb-2007-8-6-r109

49 Fernández R, Gabaldón T. (2020). Gene gain and loss across the metazoan tree of life. https://doi.org/10.1038/s41559-019-1069-x

50 Pollen AA, Kilik U, Lowe CB, Camp JG. (2023). Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. https://doi.org/10.1038/s41576-022-00568-4

51 Mestas J, Hughes CCW. (2004). Of Mice and Not Men: Differences between Mouse and Human Immunology. https://doi.org/10.4049/jimmunol.172.5.2731