Published on Jun 17, 2025 by Arcadia Science **DOI**: 10.57844/arcadia-znum-bm22

Phylogenies and biological foundation models

Biological foundation models are, at their core, evolutionary comparisons on massive scales. As with all comparative studies, evolutionary nonindependence determines their power. We chart how this affects biological AI and propose practical routes to set the field on firmer ground.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Erin McGeever, Ryan York

Version 1 · Jun 17, 2025

Purpose

We're entering an era of biological foundation models (BFMs) — general-purpose biological prediction systems. BFMs generalize by inferring evolutionary patterns through massive comparisons of diverse data. However, using this strategy, they inherit a challenge long recognized by evolutionary biologists: biological data are inherently nonindependent due to the evolutionary process. Evolutionary nonindependence can make models overfit, biased, and capable of incorrect conclusions if unaccounted for.

Here, we show how nonindependence might affect BFMs and look for signs of its presence. We document patterns of data leakage, pseudoreplication, and model biases. Exploring possible solutions, we consider data rebalancing and using perplexity to characterize phylogenetic structure in model inputs, training regimes, and outputs. We conclude that, to realize the full potential of BFMs, machine learning and evolutionary biology must open a deep and ongoing dialogue.

- All associated **code**, including utility and analysis scripts, is available in this GitHub repository.
- All **data**, including example protein families, are available on Zenodo.

Background

Forty years ago, Joe Felsenstein published “Phylogenies and the Comparative Method” **[1]**. In it, Felsenstein points out that all comparative biological studies possess a standard limitation: the dependencies of evolutionary history limit statistical power. Many studies are underpowered, and some risk arriving at completely wrong conclusions. Evolution determines what's learned and the conclusions drawn. It eventually became an enormously influential paper.

Up to then, studies that compared traits from multiple species — say, for example, the relationship between tooth area and body size in mammals **[2]**, dioecy traits in angiosperm plants **[3]**, or chromosome number and insect eusocial behavior **[4]** — handled these traits as independent observations. What does this mean? In essence, biological characteristics are treated like coin tosses. With coins, outcomes aren’t influenced by previous throws; the probability of heads or tails will always be 50:50, no matter what came before. However, as Felsenstein shows, in biology, outcomes are extremely *non*independent. Evolution occurs via descent with modification. Organisms diversify from common ancestors at varying rates of change. The result is a vast hierarchy of structured relationships that can be modelled as phylogenetic trees. The various features we observe in an organism *now* are determined by what came *before*. Unlike coins, a species' traits *are* influenced by its history. Not all trait features or combinations are possible across evolutionary space; they depend on the evolutionary

history of where you're looking. This is why algae don't have cerebral cortices, humans can't photosynthesize, and finches don't produce milk.

Controlling for nonindependence can radically decrease statistical power. Say you've collected measurements from 200 species. Assuming independence, your dataset has an effective sample size of 200. However, let's say that just two species gave rise to this group, each leading to 100 identical daughter species. In this case, the sample size is two (corresponding to the ancestral states). Depending on whether you assume independence or nonindependence, the apparent statistical power of the dataset will differ by *two orders of magnitude* **[1]**. Assuming independence here could lead to various statistical problems: overfitting, inference biases, and, in worst-case scenarios, completely incorrect conclusions. Felsenstein demonstrates that accounting for phylogenetic relationships is the only way out. He proposes a method — phylogenetic independent contrasts — that uses phylogenetic information to compare trait values only at ancestral nodes, thereby controlling for relationships between individual samples. Many phylogenetic comparative methods have since been built on the foundation of this method.

In a section titled "What if We Do Not Take the Phylogeny into Consideration?", Felsenstein makes a remarkable admission toward the end of the paper. Certain reviewers thought the paper's message was too nihilistic. They wanted a simple method that could rescue comparative studies without a phylogeny. Given the difficulty of genetic sequencing in 1985, detailed phylogenies were a rare commodity. Felsenstein doesn't relent. In the paper's last sentence, he admits there are "considerable barriers to making practical use" of his method **[1]**. However, not doing so will always lead to statistical error and bias.

Things are very different today. Phylogenies abound due to modern sequencing and the development of comparative phylogenetic methods. Nonindependence is central to evolutionary biology. Trait comparisons are everyday occurrences and can be used to identify causal features of biology. Some argue that molecular and phenotypic histories can be so well-modelled that we may ultimately be able to predict evolution **[5]**. "Phylogenies and the Comparative Method" identified a ubiquitous research flaw without an immediate path forward. Decades later, advances coalesced to make its insights actionable. We're now in a similar situation in 2025, and a similar unresolved argument may be needed (albeit in a context alien to Felsenstein in 1985).

We're in an era of biological foundation models (referred to as BFMs going forward). BFMs are general-purpose models trained on vast, evolutionarily diverse datasets. If they live up to what's promised, what's on the table is head-spinning. Revolutionary medicines on demand. Elimination of cancer. Uncovering the deep principles of life **[6]**. By and large, the power of these generative models comes from training on massive, evolutionarily diverse sequence datasets. They're exploring increasingly massive parameter spaces rivaling large language models (LLMs) in training, cost, and complexity **[7][8][9]**.

Despite their complexity, BFMs are, at their core, beefed-up versions of the comparative studies Felsenstein cited in 1985. The difference is, instead of correlating two phenotypes across a dozen species, these models make billions (or trillions) of comparisons over the known biological universe. For example, the recently published Evo 2 model was trained on a "representative snapshot of genomes spanning all observed evolution" (~40,000 genomes; 9.3 trillion DNA base pairs). The model has 40 billion parameters and a 1-million-token context window **[8]**. By learning on such a massive scale, it's believed that the model will uncover an emergent set of rules governing molecular evolution (sometimes referred to as a "language" or "grammar").

However, what if Felsenstein's cautious points from 1985 still hold for these gargantuan comparative studies? It's often argued that generative models learn about evolution through co-evolutionary and/or phylogenetic relationships **[10][11]**. Is this enough to control for nonindependence? What if not? Where does that leave us, and how do we move forward? These are the questions of this pub.

An illustrative problem

A simple, illustrative example may help us connect Felsenstein's problem with biological machine learning. Cytochrome c oxidase subunit 1 (COX1) is a mitochondrial enzyme that's become a popular DNA "barcode" for species identification **[12]**. Most species on Earth possess a version of COX1, and its evolutionary history is multifaceted. For example, COX1 sequences have diversified extravagantly among animals **[13]** (Figure 1, A). Animal COX1 has changed so consistently that many species-specific versions exist; COX1 sequences of some sibling species differ at over 50% of sites **[14]**. Given this information richness, the use of COX1 as a barcode is especially useful among animal species. Here, the number of effective sequences may be close to the number of species.

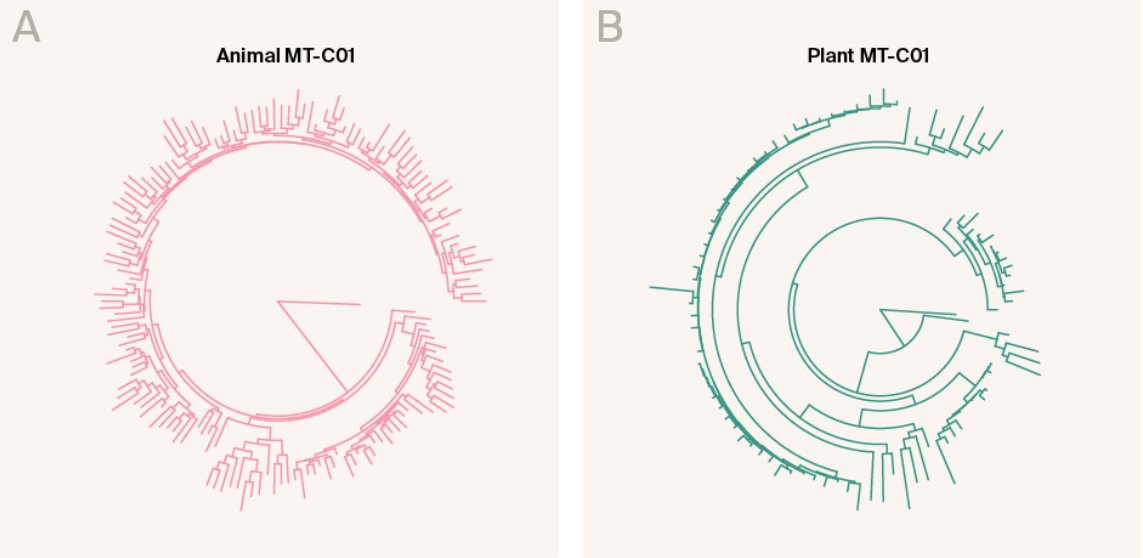

Figure 1

COX1 phylogenetic diversity.

Animal (A) and plant (B) COX1 phylogenetic trees. Tree from Zafeiropoulos et al. 2021 **[15]**.

Great for animals. The story is quite different elsewhere. Among most plants and fungi, COX1 evolution has been slow **[16][17]** (Figure 1, B). It has been so slow that, in some lineages, distantly related species possess nearly identical sequences. Here, COX1 as a species identity barcode doesn't work: a single sequence might connect to multiple (or many) species. There's no hope of identifying the correct one. Among these lineages, COX1 exhibits extreme *nonindependence*. As with the example in the introduction, these extant COX1 sequences will reflect ancestral versions. Though descended species may number in the thousands, given the slow rate of evolution, the number of effective sequences is far lower than this figure.

Imagine you're interested in creating a machine learning model that generates novel COX1 sequences. And say you collect primarily just plant sequences. There are many possible reasons for this. It could be because of history: the field has focused more on studying this gene in plants. It could be technical: it's easier to extract DNA or fully sequence the plant versions of the gene. It could be accessibility issues: maybe simply choose a plant-dominated database (maybe without intending to).

Whatever the reason, the number of effective sequences will be minimal, even with thousands of entries. What would the outcome be? The model might accurately predict plant sequences, gleaning what it can from the few effective sequences. However, performance on non-plants would be abysmal. Moreover, if specific lineages overcontributed training sequences (maybe because of the same technical, historical, or accessibility constraints mentioned above), it may not even be able to predict most plant sequences. The model will have learned to copy/paste a local evolutionary pattern, not the desired rules that govern COX1 sequence variation. The sampled evolutionary space thus limits what's learned and the conclusions drawn.

What if we expanded our model's scope to a "universe" of sequences? How many sequences look like COX1? How many don't? Do all gene families possess fewer degrees of freedom than the number of their members? How many genes have complex histories of which we have only sampled small portions (i.e., have we sampled just the equivalent of the animal *or* the plant/fungal portions of the tree)?

This leads us to the problem. Evolutionary nonindependence will influence all big biological models. Nonindependence will be unevenly distributed across evolutionary space, potentially invisible to model architectures, and possess an unknown distribution. Unpredictable errors and biases will be present, and their origins will be largely untraceable with current tools. Crucially, inferring the depth of the problem *a priori* is currently impossible.

Mapping the landscape of nonindependence

Phylogenetic relationships determine the effective sample size of comparative datasets. The effective sample size of COX1 was low among plants (few unique sequences) but proportionally high among animals (many unique sequences). Do most protein families resemble COX1, or is nonindependence an exception to the rule? How worried should we be?

Analyzing nonindependence across protein families might help. With this in mind, we looked at eukaryotic protein families from Ensembl's Compara database **[18]**. We computed the effective sample size for each family using Hill's diversity index **[19]**, a popular metric for inferring the biodiversity of datasets. Since Hill's diversity index

scales with dataset size and protein families vary significantly in size, we normalized the index by the number of proteins in each family. We refer to the resulting measure as evenness. An evenness of one means the effective sample size is close to the sample number; an evenness of zero indicates that a single sample dominates the effective sample size. Values close to one reflect greater independence and, in turn, statistical power.

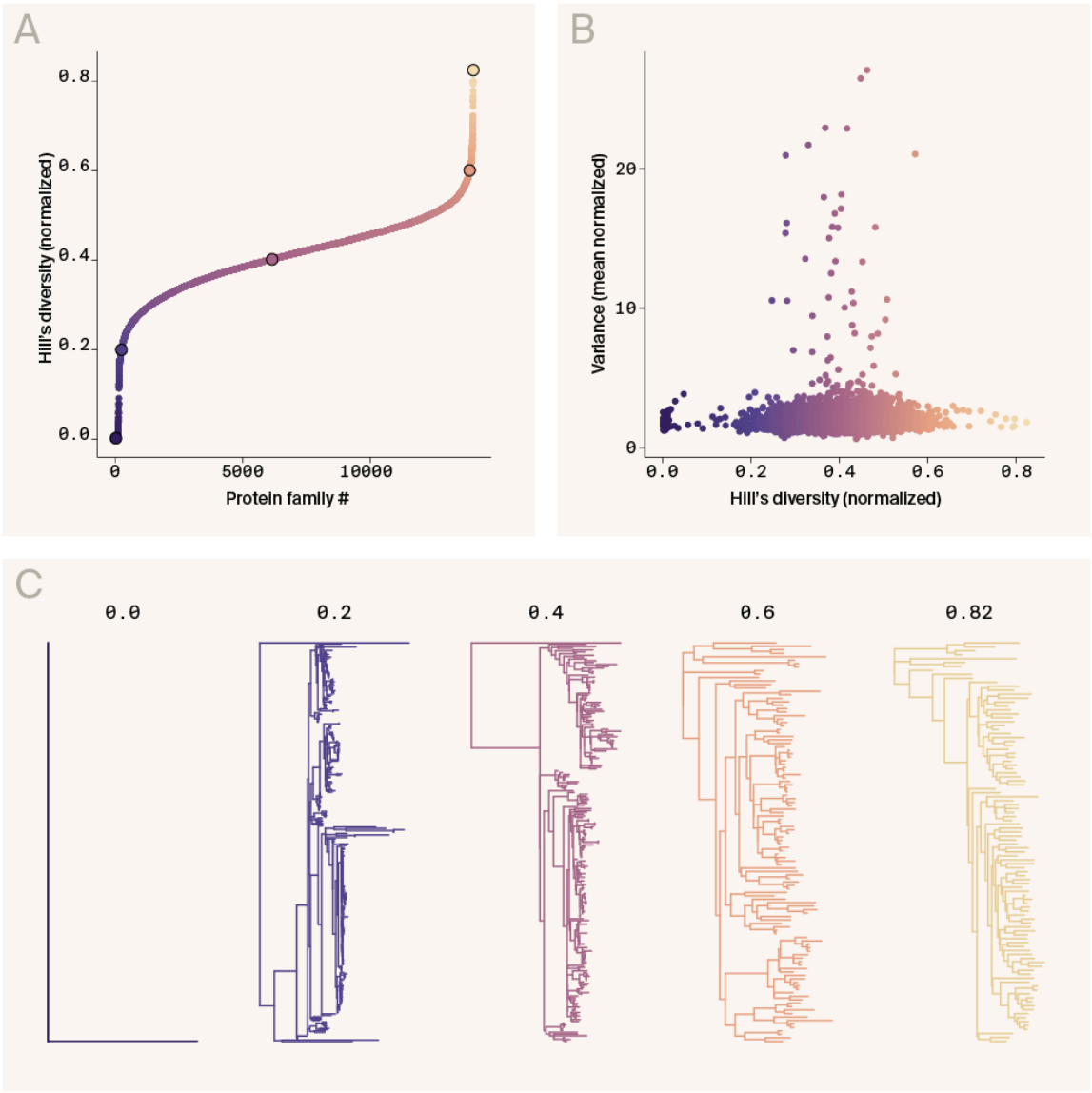

Figure 2

Effective sequence number variation.

(A) The distribution of Hill's diversity index (normalized by branch number) across vertebrate protein trees from the Ensembl Compara database (minimum # of branches per tree = 100). The position of example trees in (C) is enlarged and outlined in black.

(B) The joint distribution of Hill's diversity index and branch length variance (mean normalized) for the same protein trees in (A).

(C) Example protein trees. Hill's diversity index is labeled above each. Colors correspond to the trees' positions in (A).

Most protein families possessed an evenness of less than one (94.13%) (Figure 2, A). Across families, evenness was roughly normally distributed, with a mean of 0.41 and a left-skewed tail (Figure 2, B). Protein families closer to the mean displayed increased branch length variation, suggesting they contained multiple diversification patterns (as with COX1) (Figure 2, B). This makes abundant sense from an evolutionary perspective. Both consistent change (as you'd see with evenness values ≈ 1) and stasis (evenness values ≈ 0) are relatively rare; instead, things tend to evolve via a mixture of both over time, leading to the hierarchical and varied relationships indicative of most phylogenetic trees (Figure 2, C).

These observations lead to a couple of quick insights. First, we should expect most sequence sets to contain (at least some) nonindependence. Second, nonindependence will be unevenly distributed. As Felsenstein predicted, every comparative study (and, hence, BFMs) will be at risk of influence from nonindependence. Significantly, nonindependence will be influenced by evolutionary history (i.e., heterogeneity in diversification over time) and sampling (i.e., heterogeneity in data collection).

Tracing bread crumbs

Nonindependence is an expected problem for BFMs. Exactly how and to what extent this should affect BFMs is unclear, at least from first principles. Luckily, nonindependence can leave signatures. Affected ML models may possess data leakage and performance biases (among other features). How present are these signatures in BFMs?

Data leakage

Data leakage occurs when a training dataset has information intended to be restricted to its accompanying test set **[20][21]**. Information that was supposed to be off limits is learned, leading to overly optimistic error estimates and a tendency toward overfitting **[22]**. Nonindependence influences data leakage via pseudoreplication; similar (or identical) data points may enter training and test splits.

A recent paper by **[23]** looked at whether data leakage affects protein language model (pLM) pretraining. The authors assess the effects of two data split strategies on protein thermostability prediction by the popular pLM ESM2 **[24]**. One strategy to control leakage is to avoid the overlap of pretraining and test sets. The other employs a popular, clustering-based naive split approach. The naive split produced higher performance across the board **[23]**, suggesting that ESM2 can use leakage to boost performance on this task.

Bhatnagar et al. performed a similar test with ProGen3 **[7]**. Here, the authors admirably wanted to control for biases in their training data distribution. They used four schemes to balance training data, ranging from relaxed to stringent sequence similarity-based filtering **[7]**. In this framework, stringency is a proxy for the likelihood of data leakage. The relaxed filtering scheme allows the most sequence similarity between the training and test sets. The stringent scheme allows the least. Models were trained on each scheme and assessed using three different validation sets: sequences with 30%, 50%, and 90% similarity to the training data. Across the board, model performance was better for schemes where more sequences were shared between training/test splits **[7]**. Notably, the disparity between the relaxed and stringent schemes increased with validation set similarity: the difference in model losses between these schemes was 0.076 at 30% similarity, 0.105 at 50%, and 0.329 at 90%. The best-performing model was the least stringent scheme, predicting sequences with 90% similarity to its training set. Like ESM2, ProGen3 can leverage training/test similarity to boost performance.

Signs of data leakage aren't restricted to pLMs. Recent work has also pointed out leakage in applications using AlphaFold structural predictions **[25]**, protein–protein interaction inference **[26]**, and genomic language models (gLMs) **[27]**, among others **[28]**.

Performance biases

Even without data leakage, models can display preferential biases toward specific data patterns. More abundant data elements (e.g., highly conserved sequences or a well-represented species) can influence model learning and generate detectable performance biases. When present, there should be a traceable relationship between the distribution of the training data structure and the model output.

Multiple recent reports have connected pLM biases with the structure of their training data. For example, Gordon et al. 2024 **[29]** found that a preference for specific protein sequences during pretraining influenced the performance of ESM2 on various fitness prediction tasks. This preference couldn't be explained by model architecture alone. Instead, the authors note a substantial role for "user-level bias in curation of training data" **[29]**. What ESM2 could learn, and ultimately deem biologically plausible, was influenced by the randomness of data collection and latent evolutionary relationships.

Ding and Steinhardt (2024) recently demonstrated that species abundance disparities in protein databases lead to ESM2 and ProGen2 performance biases. They found that these models would preferentially generate proteins similar to abundant — and therefore high-likelihood — species in protein design tasks **[30]**. In a similar vein, recent work of ours identified taxonomic biases influencing AlphaFold2 structural prediction and clustering **[31]**. The phylogenetic distribution of the training data was predictive of AlphaFold2 output across various taxonomic levels.

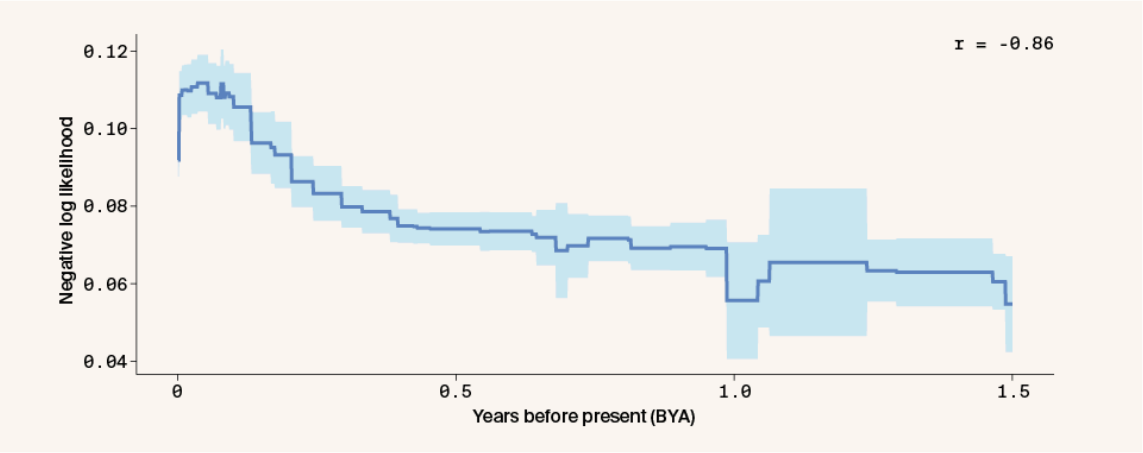

Figure 3

Gene age predicts Evo2 likelihood.

Average negative log likelihood as a function of gene age (billions of years before present; BYA) for human genes. The light blue box corresponds to the standard error. (*r* = Pearson's correlation coefficient).

Interestingly, even rougher correlates of data abundance may predict model performance. To explore this, we calculated the likelihood for every human gene using the recently published gLM Evo2 **[8]**, and compared the values to each gene's age. We

reasoned that older genes will likely be more abundant in training data. Speciation will have had more time to spread copies of their sequences. Fitting with this idea, gene likelihoods linearly increased with age (Figure 3). While more work is needed to flesh this out, it's intriguing to note the strength of this relationship and its parallels with the previously mentioned examples.

BFMs are a recent phenomenon. They're diverse in size, architecture, and goals. Many are sophisticated. With further refinement, it may be possible to overcome limitations mentioned here (at least in some contexts). The literature describing their limitations is nascent; the studies discussed here are a part of a small cohort just beginning to emerge. It therefore remains to be seen how universal their observations are. Still, the fact remains that, at their core, BFMs are comparative studies. For this reason, because predictable effects of phylogeny have already started to be detected, accounting for evolutionary nonindependence will be of universal benefit. But how?

Possible solutions

Rethinking data

Balancing datasets via clustering is a common approach (although not often used to control for nonindependence per se). For example, the known protein universe contains an enormous number of sequences (hundreds of millions to billions). Many of these sequences are copies of themselves; this is entirely predictable given the phylogenetic structure of biology (think back to our COX1 example). Duplicated sequences inflate dataset sizes and expand the compute needed. To deal with this, algorithms have been developed to identify, cluster, and remove sequences with a sequence similarity over some user-defined amount.

MMseqs2 **[32]** is a popular option. In training ProGen3, MMseqs2 was used to cluster and sample sequences across various values **[7]**. Comparing model performance across multiple clustering stringencies allowed the authors to select a desirable trade-off between compute and performance (albeit, as discussed previously, at the risk of increasing data leakage). Similarly, Fournier et al. 2024 **[33]** found that performant pLMs could be trained on MMseqs2 filtered data for an order of magnitude less cost than other models. ESM2 performance *increased* when trained on the filtered data,

benefiting sequence recovery *and* protein structural inference tasks. This suggests that data balancing can lead models to learn more biologically relevant patterns, potentially obviating the need for increased scaling of model complexity.

However, cluster-based approaches have their limitations. Global sequence-similarity filtering may be insensitive to phylogenetic structure. For example, animal and plant COX1 sequences will have different response curves when clustering; the number of retained sequences will likely depend on the sequence similarity threshold used. We clustered protein families from the Pfam database **[34]** at different similarity thresholds using MMseqs2 to demonstrate this. As expected, the number of protein clusters represented among the families decreased with thresholding, though not significantly; 88.3% of clusters were retained at a sequence similarity of 10%. Analyzing the same distribution by protein family tells a different story. The filtering effects were unevenly distributed across protein families (Figure 4, A). Some families were unaffected, while others rapidly shrank with any amount of filtering. Most were somewhere in between. Notably, sensitivity to filtering (the slope of cluster number regressed on sequence similarity) was positively correlated with effective sequence number ($r = 0.55$, $p = 4.8 \times 10^{-19}$; Pearson's correlation) (Figure 4, B). This suggests a couple of things.

First, retained sequence diversity after clustering will vary by protein family, creating a nonlinear relationship between similarity thresholds and training data distribution. In other words, sequence diversity (and, in turn, statistical power) won't be maintained as dataset size decreases. Second, phylogenetic relationships predict this relationship. Families with low effective sequence count will be most affected by clustering; those with high effective sequence count will be more robust. Most sit on a diverse, heterogeneous continuum (Figure 4, B). Evolution determines the behavior of the data and, ultimately, what can be learned.

Other considerations arise. Nonindependent relationships are ubiquitous. How big would the biological universe be if they were removed? Nonindependence is also unevenly distributed. Would filtering lead to the loss of whole chunks of biological phenomena (taxonomic groups, protein families, molecular functions)? More provocatively, is it possible we have already hit "peak biological data" (as has been suggested, this is the case for natural language models **[35]**)?

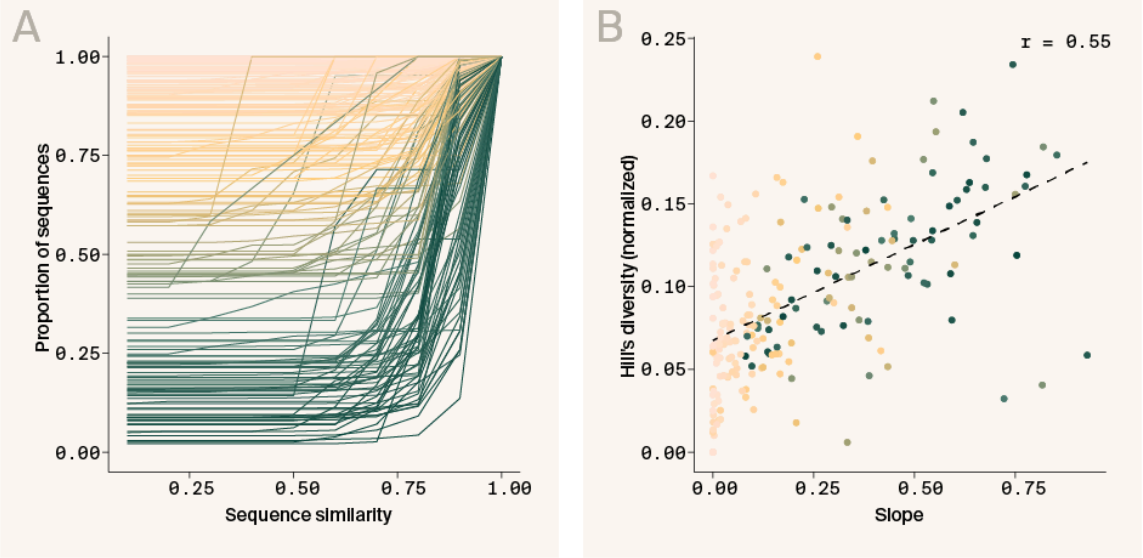

Figure 4

Nonindependence affects filtering outcomes.

(A) Proportion of retained sequences as a function of sequence similarity filtering with MMseqs2. Each line corresponds to a protein family. Color corresponds to the minimum retained sequence proportion for each family.

(B) The relationship between the slopes of each line in (A) (x-axis) and the Hill's diversity of each protein family (inferred from their multiple sequence alignments). Colors are the same as in (A) (*r* = Pearson's correlation coefficient).

Much of the known protein universe's diversity comes from species with just a few associated sequences or structures **[31]**. Even minimal data balancing leads to substantial loss of phylogenetic diversity. For example, ~48% of species are removed with at least two proteins **[31]**. Filtering also decreases the diversity of higher taxonomic groups and protein structural clusters **[31]**.

Similar patterns are present if filtering is done on sequence, instead of taxonomic diversity. Returning to the data in Figure 2, we wondered what the total number of effective sequences was across all the Ensembl Compara protein trees. In other words, how much data would be left if we filtered to just effective sequences? Given our estimates, we found this would reduce the dataset by about 58% (total sequences = 4,230,261; effective sequences = 1,782,627). Furthermore, as shown in Figure 2, these effects weren't evenly distributed (Figure 2, A). Suppose this pattern is

representative (remember, these are just vertebrate protein trees). In that case, it may be reasonable to expect that most of the data in a given sequence dataset will be pseudoreplicated. This substantially reduces usable data and, thus, statistical power.

In a peak data scenario, we'd expect models to have slurped up most available information in the sequence universe. Evidence supporting this may be accumulating. For example, Tule et al. 2025 **[11]** found that ESM2 predicts evolutionary relationships better than ESM3. Similarly, ProGen3 models with more parameters perform worse on zero-shot fitness prediction than smaller ones **[7]**. Better estimation of proteins' natural (phylogenetic) distribution may come at the cost of fitness prediction accuracy **[7]**. This fits with **[36]** and Gordon et al. 2025 **[29]**; phylogenetic relationships and protein fitness may reflect divergent processes. Only so much can be learned about one from the other (more on this later). And, fitting with patterns that may be expected from a “peak data” scenario, performance on both distributions may fall off as larger and more complex models are trained. Prediction worsens after a specific training threshold **[7]**. Time will tell if these patterns are borne out. Just where the hypothesized training thresholds exist will likely be model and task-specific. At the very least, initial signs indicate that the universe of usable data is smaller than we think.

Perplexity all the way down

So, clustering and data balancing may not be the solution. Could phylogenetic relationships be used to address nonindependence in BFMs? One way to do this would be direct inference of the latent phylogenetic structure of model inputs and outputs (i.e., training/test/validation data and predictions/generations). This would be generally useful. Datasets could be evolutionarily balanced. Training/test splits could be optimized. Loss functions accounting for nonindependence could be created. The amount of new biology a model has learned could be estimated. The whole of biological ML model development could be informed, motivating better-scoped problems and allowing generalizable rules to be explored more deeply. The problem is this: phylogenetic tree inference doesn’t scale to the size of datasets used by BFMs **[37]**. This feels reminiscent of Felsenstein’s argument in 1985 **[1]**. Accounting for phylogeny is the right thing; we often don’t have the phylogenies to do it.

Information theory may provide a solution. As calculated here, Hill’s diversity index is the exponential of a system's Shannon entropy **[19]**. In other contexts, this statistic is given a different name. “Perplexity” has been a popular measure of natural language

processing (NLP) model uncertainty since the late 1970s **[38]**. As you might have already guessed, perplexity is the exponential of Shannon entropy. Since language and protein models often share a common goal — token prediction — perplexity has also become popular in measuring the residue prediction confidence in pLMs. Predictions with lower perplexity mean the model can constrain its estimates to fewer equally possible options; lower perplexity means more certainty. In addition to assessing the certainty of individual tokens, average perplexity can also be used to benchmark a model's performance and scalability. For example, Lin et al. 2023 **[24]** showed that ESM2 perplexity dropped from 10.45 to 6.37 bits/residue when scaling from 6 million to 15 billion parameters. Perplexity is a flexible way to measure model performance locally, globally, and contextually.

At the same time, and as we have already seen, latent features of datasets can be inferred via perplexity. In Figure 2, we used perplexity to characterize and assess the effects of filtering a dataset containing millions of sequences. Perplexity efficiently characterizes phylogenetic patterns and can be cheap to compute; we could estimate a global distribution from quick measurements of individual protein family trees' perplexity. Here, and concerning datasets, more perplexity means more effective measurements, more statistical power, and, likely, more possibilities for generalization.

In this expanded way, perplexity may be broadly helpful in helping machines better learn phylogeny. Perplexity can help estimate and account for the evolutionary nonindependence of model inputs (i.e., data). On the other hand, minimizing perplexity helps build certainty around model outputs (i.e., predictions). It's desirable to build perplexity into model training. For example, the difference between the dataset and the prediction perplexity could be a potential loss function. In theory, the distributions of latent and learned perplexity could also be used for model interpretation, helping gauge the phylogenetic novelty of *de novo* designs or identifying which biological patterns have been captured. Finally, a generalized use of perplexity helps us understand how sequence preferences **[29]** determine the interplay between data and model performance, allowing for fairer comparative benchmarks for BFMs **[39]**.

Phylogeny won't be enough

Information-theoretic measures like perplexity may help shortcut some of the roadblocks faced by BFMs. However, it's worth briefly discussing why solving the

phylogeny problem alone likely won't be sufficient to realize the lofty goals of biological machine learning.

Many BFMs implicitly assume that the phylogenetic distribution of traits mirrors their fitness. Put plainly, things preferred by evolution (i.e., more "fit") will be selected for, propagate, and become more abundant. If something still exists after millions of years, it must have won some evolutionary jackpot, right? This assumption is why the likelihood calculations of many BFMs are based on their training data distributions. The "natural" distribution determines the assumed fitness of new observations. If the BFM has learned a deep "grammar" of biology, then observations far away from the natural distribution must be considered less likely and less fit.

However, evolution doesn't always generate traits that are optimally fit. It rarely, if ever, does. Trade-offs and constraints are permanent features. Genetic constraints mean only particular fitness trajectories are available **[40]**. Complex interactions and trade-offs limit phenotypic possibilities (again, this is why birds don't produce milk and algae lack nervous systems). Biophysical limits further narrow these possibilities. Only some pathways through this complex landscape exist at any given moment, determined by the combined weight of historical, genetic, and phenotypic reins. Extant biological features are, therefore, "fit enough" and fall somewhere on a (largely unknown) fitness continuum.

As we already saw, learning to navigate this complex landscape comes at a cost of learning phylogeny and vice versa **[7]**; one doesn't solve the problem of the other. The natural density of sequences (i.e., the phylogenetic distribution) consistently fails to estimate fitness **[36]**. Moreover, even if infinite sequence data were available, exact fitness estimation would be impossible **[36]**. For this reason, separating phylogeny, even if perfectly accounted for, and fitness (i.e., the sieve through which the "grammar" of biology passes) remains a complex problem for BFMs. Recent work suggests that the fitness/phylogeny disconnect may be addressed by simple modifications to model inference that account for shared selective pressures among sequences **[41]**. It'll be interesting to see how inference adjustments like this might mitigate limitations of current BFMs, even without modifications to data and model structures. However, model adjustments will remain necessary because of what's happening in the models.

In other words, we can scale, expand, and improve our models as much as we want. Ultimately, and as we hope the reader has come to appreciate, evolution will continue to determine what we can learn and the conclusions we can draw.

Methods

Data acquisition and processing

Code, including utility and analysis scripts, is available in our GitHub repo (DOI: 10.5281/zenodo.15678022).

All **data**, including example protein families, are available on Zenodo (DOI: 10.5281/zenodo.15644457)

COX1 tree

Zafeiropolous et al. published the COX1 gene tree in 2021 **[15]**. The tree was filtered to the phyla Streptophyta and Chordata, representing plants and animals (respectively).

Vertebrate protein trees

EMF (extended multi-format) files were downloaded from the Ensembl Compara database (release 114) **[18]**. Protein family trees (Newick files) were extracted using the script `fig2_extract_trees.py` , resulting in 54,399 tree files.

Human gene sequences

Human gene ages were collected from the GenOrigin database **[42]**. Human cDNA sequences (GRCh38) were downloaded from Ensembl (release 104) using the script `fig3_download_human_genes.py` . For genes with multiple cDNA sequences, the longest was used. This resulted in 22,796 individual FASTA files.

Pfam protein families

The complete Pfam database was downloaded from InterPro using their FTP server (release 37.3). A random subset of families for analysis was selected and is provided on Zenodo.

Measuring effective sequence count of protein family trees

We used Hill's diversity index to infer the effective sequence count of the Ensembl Compara protein trees. We required at least 100 proteins in a family to be considered for analysis. This resulted in a final set of 14,054 trees ranging in size from 100 to 1,499 proteins (median *n* proteins = 209).

We used a Hill's diversity *q* value of 1. This parameter value evenly weighs species count and relative abundance to estimate the effective count within a sample population. Each tree's terminal branch lengths were extracted and normalized by their sum to obtain probability distributions. Hill's diversity was then calculated as the exponential of Shannon entropy: $D_1 = \exp(-\Sigma p_i \log(p_i))$, where p_i represents the normalized branch length for tip i. Since trees varied broadly in size (from 100 to 1,499 proteins), we normalized Hill's diversity values by the number of tips in the tree, allowing us to compare the distribution of effective sequence counts across the tree dataset. The protein family tree structure variance was inferred by calculating the coefficient of variation (standard deviation/mean) of terminal branch lengths across protein families.

> Code for these analyses can be found in the script, `fig2_analysis.R`.

Evo 2 likelihood analysis

Evo 2 likelihoods for human gene cDNA sequences were calculated via API access to the evo2-40b model hosted on NVIDIA. We computed the mean likelihood for each gene and converted it to a negative log scale. We then used a rolling window approach to assess likelihood distributions as a function of gene age. The mean and standard

errors of likelihoods for genes falling within 250 million-year windows were computed from 1 million years ago to 1.5 billion years ago (1 million-year step size). The relationship between age and rolling-window likelihood was assessed using Pearson's correlation.

Code for these analyses can be found in the script, `fig3_analysis.R`.

Sequence clustering and diversity

MSAs and individual sequences for 219 example protein families were collected from the Pfam database. Each protein family was clustered using MMseqs2 **[32]** at multiple sequence identity thresholds (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%) with a coverage value of 80%. At each identity threshold, we calculated the proportion of retained sequences by dividing the number of detected clusters by the number of proteins in the family. A value of 1 would indicate maximum sequence diversity (*n* clusters = *n* proteins) while a value of 0 would reflect strong sequence conservation (*n* clusters < *n* proteins). For each family, we regressed sequence retention by sequence identity to obtain a slope of this fit.

Hill's diversity was calculated using column-wise amino acid frequency distributions for multiple sequence alignments. For each alignment position, we computed diversity as $D_1 = \exp(-\Sigma p_i \log(p_i))$, where p_i represents the frequency of amino acid i at that position. Mean diversity across all alignment positions was normalized using the formula $(D_1-1)/(S-1)$, where S = 20 represents the maximum possible amino acid diversity. These values were then compared to the slopes computed above using Pearson's correlation (as in Figure 4, B).

Code for these analyses can be found in the script, `fig4_analysis.R`.

Additional methods

We used ChatGPT to help write code, and we used Claude to help write code and clean up code. We used arcadia-themeR (v0.1.1) **[43]** to generate figures prior to

manual adjustment.

References

1. Felsenstein J. (1985). Phylogenies and the Comparative Method. https://doi.org/10.1086/284325

2. Pilbeam D, Gould SJ. (1974). Size and Scaling in Human Evolution. https://doi.org/10.1126/science.186.4167.892

3. Givnish TJ. (1982). Outcrossing Versus Ecological Constraints in the Evolution of Dioecy. https://doi.org/10.1086/283959

4. Sherman PW. (1979). Insect Chromosome Numbers and Eusociality. https://doi.org/10.1086/283445

5. Stern DL, Orgogozo V. (2008). THE LOCI OF EVOLUTION: HOW PREDICTABLE IS GENETIC EVOLUTION? https://doi.org/10.1111/j.1558-5646.2008.00450.x

6. Amodei D. (2024). Machines of Loving Grace. https://www.darioamodei.com/essay/machines-of-loving-grace

7. Bhatnagar A, Jain S, Beazer J, Curran SC, Hoffnagle AM, Ching K, Martyn M, Nayfach S, Ruffolo JA, Madani A. (2025). Scaling unlocks broader generation and deeper functional understanding of proteins. https://doi.org/10.1101/2025.04.15.649055

8. Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, Gonzalez GA, King SH, Li DB, Merchant AT, Naghipourfar M, Nguyen E, Ricci-Tam C, Romero DW, Sun G, Taghibakshi A, Vorontsov A, Yang B, Deng M, Gorton L, Nguyen N, Wang NK, Adams E, Baccus SA, Dillmann S, Ermon S, Guo D, Ilango R, Janik K, Lu AX, Mehta R, Mofrad MRK, Ng MY, Pannu J, Ré C, Schmok JC, John JSt, Sullivan J, Zhu K, Zynda G, Balsam D, Collison P, Costa AB, Hernandez-Boussard T, Ho E, Liu M-Y, McGrath T, Powell K, Burke DP, Goodarzi H, Hsu PD, Hie BL. (2025). Genome modeling and design across all domains of life with Evo 2. https://doi.org/10.1101/2025.02.18.638918

9. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, Badkundri R, Shafkat I, Gong J, Derry A, Molina RS, Thomas N, Khan

YA, Mishra C, Kim C, Bartie LJ, Nemeth M, Hsu PD, Sercu T, Candido S, Rives A. (2025). Simulating 500 million years of evolution with a language model. https://doi.org/10.1126/science.ads0018

10 Lupo U, Sgarbossa D, Bitbol A-F. (2022). Protein language models trained on multiple sequence alignments learn phylogenetic relationships. https://doi.org/10.1038/s41467-022-34032-y

11 Tule S, Foley G, Bodén M. (2024). Do protein language models learn phylogeny? https://doi.org/10.1093/bib/bbaf047

12 Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. (1996). The Whole Structure of the 13-Subunit Oxidized Cytochrome c Oxidase at 2.8 Å. https://doi.org/10.1126/science.272.5265.1136

13 Nabholz B, Glémin S, Galtier N. (2009). The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. https://doi.org/10.1186/1471-2148-9-54

14 Hebert PDN, Cywinska A, Ball SL, deWaard JR. (2003). Biological identifications through DNA barcodes. https://doi.org/10.1098/rspb.2002.2218

15 Zafeiropoulos H, Gargan L, Hintikka S, Pavloudi C, Carlsson J. (2021). The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. https://doi.org/10.3897/mbmg.5.69657

16 Wolfe KH, Li WH, Sharp PM. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. https://doi.org/10.1073/pnas.84.24.9054

17 De Chiara M, Friedrich A, Barré B, Breitenbach M, Schacherer J, Liti G. (2020). Discordant evolution of mitochondrial and nuclear yeast genomes at population level. https://doi.org/10.1186/s12915-020-00786-4

18 Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, Spooner W, Kulesha E, Yates A, Flicek P. (2016). Ensembl comparative genomics resources. https://doi.org/10.1093/database/bav096

19 Hill MO. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. https://doi.org/10.2307/1934352

20 Kaufman S, Rosset S, Perlich C, Stitelman O. (2012). Leakage in data mining. https://doi.org/10.1145/2382577.2382579

21 Kapoor S, Narayanan A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. https://doi.org/10.1016/j.patter.2023.100804

22 Sasse A, Ng B, Spiro AE, Tasaki S, Bennett DA, Gaiteri C, De Jager PL, Chikina M, Mostafavi S. (2023). Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. https://doi.org/10.1038/s41588-023-01524-6

23 Hermann L, Fiedler T, Nguyen HA, Nowicka M, Bartoszewicz JM. (2024). Beware of Data Leakage from Protein LLM Pretraining. https://doi.org/10.1101/2024.07.23.604678

24 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574

25 Dobson L, Tusnády GE, Tompa P. (2025). Regularly updated benchmark sets for statistically correct evaluations of AlphaFold applications. https://doi.org/10.1093/bib/bbaf104

26 Bushuiev A, Bushuiev R, Kouba P, Filkin A, Gabrielova M, Gabriel M, Sedlar J, Pluskal T, Damborsky J, Mazurenko S, Sivic J. (2023). Learning to design protein-protein interactions with enhanced generalization. https://doi.org/10.48550/ARXIV.2310.18515

27 Rafi AM, Kiyota B, Yachie N, de Boer C. (2025). Detecting and avoiding homology-based data leakage in genome-trained sequence models. https://doi.org/10.1101/2025.01.22.634321

28 Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, List M. (2024). Guiding questions to avoid data leakage in biological machine learning applications. https://doi.org/10.1038/s41592-024-02362-y

29 Gordon C, Lu AX, Abbeel P. (2024). Protein Language Model Fitness Is a Matter of Preference. https://doi.org/10.1101/2024.10.03.616542

30 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

31 Avasthi P, York R. (2024). The known protein universe is phylogenetically biased. https://doi.org/10.57844/ARCADIA-570F-5CFB

32 Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988

33 Fournier Q, Vernon RM, van der Sloot A, Schulz B, Chandar S, Langmead CJ. (2024). Protein Language Models: Is Scaling Necessary? https://doi.org/10.1101/2024.09.23.614603

34 Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ, Grishin NV, Schaeffer RD, Clementel D, Tosatto SCE, Sonnhammer E, Wood V, Bateman A. (2024). The Pfam protein families database: embracing AI/ML. https://doi.org/10.1093/nar/gkae997

35 Sutskever I, Vinyals O, Le QV. (2014). Sequence to Sequence Learning with Neural Networks. https://doi.org/10.48550/ARXIV.1409.3215

36 Weinstein EN, Amin AN, Frazer J, Marks DS. (2022). Non-identifiability and the Blessings of Misspecification in Models of Molecular Fitness. https://doi.org/10.1101/2022.01.29.478324

37 Bepler T, Berger B. (2021). Learning the protein language: Evolution, structure, and function. https://doi.org/10.1016/j.cels.2021.05.017

38 Jelinek F, Mercer RL, Bahl LR, Baker JK. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. https://doi.org/10.1121/1.2016299

39 Dominguez-Olmedo R, Dorner FE, Hardt M. (2024). Training on the Test Task Confounds Evaluation and Emergence. https://doi.org/10.48550/ARXIV.2407.07890

40 Weinreich DM, Watson RA, Chao L. (2005). PERSPECTIVE: SIGN EPISTASIS AND GENETIC COSTRAINT ON EVOLUTIONARY TRAJECTORIES. https://doi.org/10.1111/j.0014-3820.2005.tb01768.x

41 Pugh CWJ, Nuñez-Valencia PG, Dias M, Frazer J. (2025). From Likelihood to Fitness: Improving Variant Effect Prediction in Protein and Genome Language Models. https://doi.org/10.1101/2025.05.20.655154

42 Tong Y-B, Shi M-W, Qian SH, Chen Y-J, Luo Z-H, Tu Y-X, Xiong Y-L, Geng Y-J, Chen C, Chen Z-X. (2021). GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life. https://doi.org/10.1016/j.jgg.2021.03.018

43 arcadiathemeR. (2024). https://github.com/Arcadia-Science/arcadiathemeR