Published on Aug 01, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-570f-5cfb

The known protein universe is phylogenetically biased

Many protein prediction and design models rely on evolutionary comparisons. We show that popular databases are phylogenetically biased, influencing the statistical utility of the known protein universe in important ways.

Contributors (A-Z)

Prachee Avasthi, Erin McGeever, Ryan York

Version 2 · Mar 31, 2025

Purpose

Prediction and *de novo* generation of proteins is rapidly advancing. Much recent work relies on the comparison of diverse proteins — taken from massive public databases — to learn the evolutionary constraints on protein feature variation. By training on hundreds of millions of proteins, these models learn and, at least theoretically, generate beyond the structure of the "known protein universe." Central to this endeavor is the idea that the current "known protein universe" is sufficient for learning, and then implementing, the rules through which evolution has designed proteins.

Here, we explore the phylogenetic makeup of all 214 million proteins in the AlphaFold database (AFDB). We find strong phylogenetic biases in the AFDB. These biases are

associated with variation in prediction accuracy, influence the outcomes of downstream protein structural clustering tasks, and, when controlled for, greatly constrain the evolutionary diversity of this representation of the known protein universe.

These findings help delineate some of the promise and perils of evolution-informed protein models and should be relevant to researchers interested in the prediction and design of proteins.

- All **code** generated and used for the pub is available in this GitHub repository, including scripts for accessing data, performing analyses, and generating all figures.

Background and goals

We are entering an era of *de novo* biological design **[1]**. The application of machine learning/AI models to large biological datasets, it is believed, will unlock the potential to generate novel biological components not found in nature **[2][3]**. At the vanguard of this anticipated paradigm shift is the field of protein design. Models that can generate protein sequences and structures have rapidly advanced in recent years, attracting substantial scientific and financial interest **[4]**.

Proteins are appealing targets of generative design for several reasons. Like human language, proteins are information-complete, encoding their structure and function in amino acid sequences **[5]**. In addition, the sequences and structures of many proteins are available in public resources **[6]**. The sheer abundance and diversity of protein data has motivated the idea that we are on the brink of learning comprehensive generative rules for the “known protein universe” **[7]**.

The advent of protein structure prediction algorithms significantly catalyzed efforts to leverage the known protein universe. Approaches such as AlphaFold **[8]**, ESMFold **[9]**, and RoseTTAFold **[10]** contributed several key ingredients that laid the foundation for generative protein design models: *in silico* metrics for scoring protein prediction, abundant and re-usable structural training data **[11]**, and, importantly, an appreciation for leveraging the evolutionary diversity of proteins **[8]**. Indeed, though models vary

broadly (e.g., in architecture, type of training data, goals), almost all are based on a common assumption: the rules of biological design will fall out of evolutionary comparisons **[12]**.

This assumption is based on another: the sequences and structures available in 2024 are sufficient for learning general rules of protein design. While the amount of available protein data is indeed massive, it's important to remember that public protein databases grew randomly over time. There was no top-down roadmap to guide optimal sampling across the evolutionary diversity of proteins. Despite this, models have begun to assume that these databases define the true distributions of naturally occurring proteins **[13][14]**. Recent work has shown that this assumption can be problematic. Unequal sampling of proteins has been found to bias the behavior of protein language models; species that are better represented in training data have an outsized influence on generated proteins, limiting the contributions of rarer species and sequences **[14]**.

These findings highlight the fact that training data distribution is an important influence on the behavior of at least some protein models. Better characterizing the underlying distributions of training data would therefore be useful for understanding the potentials and limitations of protein prediction and design. Luckily proteins differ from many other types of training data — which can be hard to to characterize — in that we know the generative process underlying their sequence and structure: evolution. Even more luckily, the generative process of evolution leaves behind traceable signatures in the form of phylogeny.

We decided to see how much we might learn about the distribution of the known protein universe through the lens of phylogenetics. Have proteins been evenly sampled across the tree of life? Does the phylogenetic distribution of proteins influence model prediction? How much protein diversity have we actually sampled and is it universal? We reasoned that answers to these questions would contextualize current possibilities for protein models and provide guidelines for better leveraging evolutionary information in their creation.

The approach

Data

We downloaded **AlphaFold database (AFDB) structural identities**, **cluster designations**, and **associated metadata** from the Foldseek web server **[15]**.

We downloaded **Protein Data Bank (PDB) statistics** from the PDB website.

We collected **species taxonomies** from the NCBI Taxonomy database **[16]**. We accessed genome statistics from the NCBI Genome database.

The multi-domain phylogeny we used in all analyses was provided via personal communication from TimeTree **[12]** developers.

Measuring taxonomic completeness

We used the multi-domain scale phylogeny **[12]** as the basis for calculating taxonomic completeness measurements. Given that the identification and estimation of species diversity is more volatile than higher taxonomic levels, we measured taxonomic completeness by the diversity of families within each phylum of the phylogeny. To do so, we created a family-level phylogeny by randomly choosing a single species from each family and reduced the tree using the `keep.tip` function in the R package ape **[17]**.

The procedure for calculating taxonomic completeness was as follows. First, species within the phylogeny that contributed at least one protein structure to the AFDB were identified. These species were then associated with their family and phyla classifications. Using these classifications, we then identified the families present in the AFDB for each phylum. The most recent common ancestor of each phylum

(`getMRCA` function in ape) was identified and used to extract a subtree for all phyla (`extract.clade` function in ape). Family-level presence/absence in the AFDB was represented as a binary vector and used to measure Faith's phylogenetic diversity (`PD` function in the R package Picante **[18]**) for each phylum. Taxonomic completeness was then calculated by normalizing the phylogenetic diversity of families within the AFDB by the total phylogenetic diversity of each phylum. The distribution of taxonomic completeness across the phylogeny was visualized using the `contmap` function in the R package phytools **[19]**.

> This **full procedure** is available via the function `clade_PD` in the GitHub repository associated with this pub.

The distributions of domain-level taxonomic completeness were statistically compared using Dunn's test. The association between phylogenetic distance, number of families, and taxonomic completeness for each phylum was assessed by creating a linear model using the `lm` function in R.

Analyzing Foldseek representative proteins

The taxonomic completeness of Foldseek representative proteins was assessed using the method described above. Representative proteins were associated with their taxonomic classifications, which were then used to calculate family-level diversity per phylum (as was done with the AFDB in Figure 3). The results were visualized using the `contmap` function in the R package phytools, as above.

We assessed phylogenetic influences on the relationship between taxonomic completeness in the AFDB and Foldseek clusters using a phylogenetic generalized least squares (PGLS) regression. First, a variance-covariance matrix capturing phylogenetic relationships was calculated using the function `comparative.data` in the R package caper **[20]**. The PGLS was then constructed using the function `pgls` in caper (using maximum likelihood for branch length optimization).

The relationship between species abundance, pLDDT, and Foldseek clustering outcomes was first assessed by calculating representative protein number, total protein number in the AFDB, and mean pLDDT for each species. The relationship between total protein number in the AFDB and mean pLDDT was measured using

Spearman correlation. Spearman correlations were calculated over a range of cutoffs corresponding to minimum representative protein number (the distributions of which are presented in Figure 4, A). The distributions of pLDDT across domains were statistically compared using Dunn's test.

Assessing the effects of data balancing

The effects of data balancing were simulated by testing a range of protein sample sizes. Given the diversity of sampling across species in the AFDB, increasingly conservative sampling (i.e., requiring a greater number of proteins per species) has an inherent filtering effect on the phylogenetic diversity of the available data.

The specific effects of filtering were assessed by calculating Faith's phylogenetic diversity (using `clade_PD`) of species contributing at least *n* proteins to the AFDB over a range of minimum values (from 0 to 20,000 proteins). The distribution of these measurements is presented in Figure 5, A. Taxa diversity was also assessed at each cutoff by calculating the proportion of taxa left as a function of the total number for each level of the taxonomic hierarchy (as in Figure 5, B). The percent of cluster space was calculated by identifying all the number of unique clusters represented at each cutoff, divided by the total number of Foldseek clusters (as in Figure 5, C).

To assess per-species sampling completeness, we calculated the ratio between protein *n* in the AFDB and the total number of proteins per species in the NCBI Genome database. Given the broad dynamic range of this value — referred to as "protein ratio" in the results section — its logarithm was used for analyses. We compared the relationship between protein ratio and mean pLDDT over a range of minimum protein numbers (Figure 6) and visualized the results using contour plots via the R function `contour`.

> All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.13145188), including scripts for accessing data, performing analyses, and generating all figures.

The results

Protein databases are taxonomically biased

We first wanted to understand the basic taxonomic makeup of the known protein universe. A straightforward approach to this is to measure the number of protein structures in the database that are contributed by each species. Visual inspection demonstrated that, in both the Protein Data Bank (PDB) and AlphaFold database (AFDB), a small number of species represented orders of magnitude more proteins than all others (Figure 1, A–B). In the PDB, these structures were dominated by eukaryotic samples (likely owing to the bias toward solving human protein structures) (Figure 1, A), while the AFDB was weighted toward prokaryotes (likely owing to the bias toward sequence bacterial genomes) (Figure 1, B). Despite domain-level differences, both databases were associated with strongly left-shifted cumulative distributions, indicating that a significant proportion of their proteins come from a very small number of species (Figure 1, C). Gross taxonomic biases in species sampling therefore exist in the PDB and AFDB (this has also been noted about UniProt and other databases **[14]**).

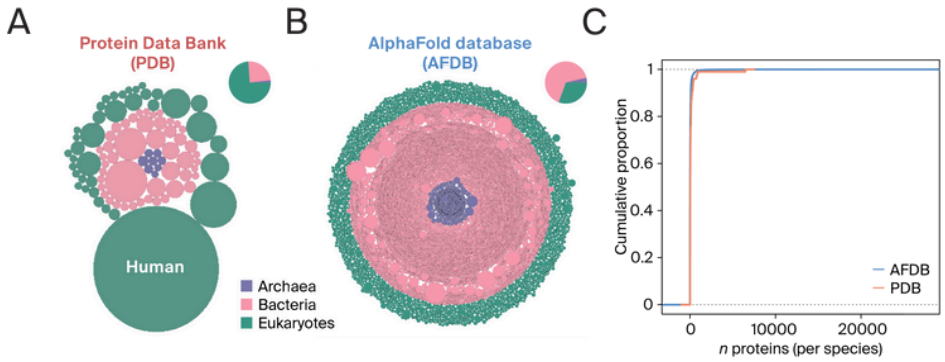

Figure 1

Species-level distributions of proteins in public databases.

(A) Circular packing plot of protein number per species in the Protein Data Bank (PDB). Circle diameter corresponds to protein number. Circles are colored by domain (green = eukaryotes; pink = bacteria; purple = archaea). The pie chart in the upper right corner the proportion of the database represented by each domain.

(B) Circular packing plot of protein number per species in the AlphaFold database (AFDB).

(C) Cumulative distributions of per-species protein number in the PDB (orange) and AFDB (blue).

What is the structure of these biases? Are they randomly distributed? Or are coherent groups of species well-represented and others not? To explore this, we measured how well-sampled phyla were in the AFDB using the complete TimeTree of Life phylogeny **[12]**. We assessed this “taxonomic completeness” by analyzing the ratio of observed and total possible phylogenetic diversity within each phylum (Figure 2, A; see Approach for details). We hypothesized that, if species were randomly sampled across the tree of life (ToL), the distribution of taxonomic completeness would be at least somewhat uniform across phyla. Conversely, strong taxonomic biases might lead to a strongly skewed distribution with only a few phyla well-represented.

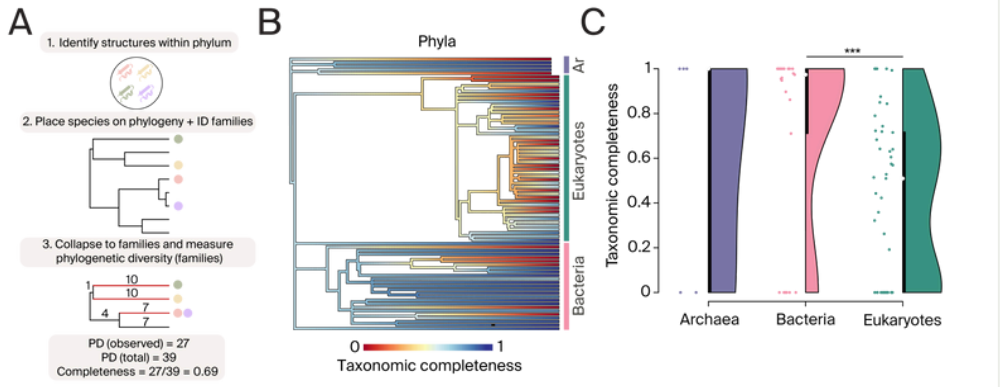

Figure 2

Taxonomic completeness of the AFDB.

(A) Graphical depiction of the approach used here to calculate taxonomic completeness.

(B) Taxonomic completeness of AFDB phyla. Domains are labeled on the right ("Ar" = archaea).

(C) Violin plot of taxonomic completeness distributions across domains of life. (*** = $p < 0.0001$; Dunn's test).

A quick note (and a bit of conceptual framing) before proceeding. All analyses presented hereafter explore patterns and distributions of *what is currently known* about the diversity of, and relationships among, species across the ToL. It's important to remember that what is known is a subset of *what actually exists* (i.e., the actual structure and composition of the ToL). The two should not be confused. A small example: > 1,300 bacterial phyla likely exist, the vast majority of which are uncultured and uncharacterized **[21]**. In this pub, we have phylogenetic data for 26 bacterial phyla. Therefore, any conclusions we make about taxonomic sampling concern phyla that have been sequenced and are at least somewhat characterized. To reiterate, the goal here is to understand the evolutionary structure of protein databases to better leverage them for training, prediction, and generation. Any claims about the structure of evolution itself should be interpreted within this context.

The distribution of taxonomic completeness was roughly bimodal across the ToL (Figure 2, B). Some phyla were completely sampled (18/77 phyla; 23%). Many others were not represented (25/77 phyla; 32%) and close to half were somewhat complete

(34/77 phyla; 44%). The most obvious trend was at the domain level: prokaryotic phyla (bacteria and archaea) were significantly better sampled than their eukaryotic counterparts (Figure 2, B–C; p = 0.0004, Dunn's test). Within eukaryotes, phyla were highly variable. Better-sampled phyla included fungi, Archaeplastida (land plants, green algae, red algae), and a handful of better-studied protist phyla (e.g., pathogenic oomycetes and diatoms). Many metazoan phyla were poorly sampled. Bacterial phyla that were not well represented included Fusobacteriota, Chlorobiota, Ignavibacteriota, Balneolata, Candidatus Melainabacteria, and Thermomicrobiota.

What accounts for this sampling disparity? Intuitively, the sheer size of phyla (i.e., the number of families per phylum) is a straightforward explanatory factor. Indeed, phylum size was significantly predictive of taxonomic completeness (linear regression; t-value = –2.2, p = 0.03). However, the model itself was not very explanatory (r^2 = 0.09). This suggests that other factors contribute to taxonomic sampling variation, the true landscape of which is likely a byproduct of both biological and historical influences. For example, the two largest phyla (Arthropoda, 1,574 families; Chordata, 1,060 families) — despite being some of the most studied in all of biology — each have modest levels of taxonomic completeness (0.51 and 0.64, respectively). These estimates are likely more accurate for these phyla than less well-studied ones. In general there may not be enough information to estimate what we have left to uncover for many phyla (as is very likely the case among many bacterial phyla). Therefore, sampling may be influenced as much by where biologists have decided to place their attention as by the complexity of taxonomy itself. Thus the current state of affairs: eukaryotes are substantially well-sampled within the known organismal universe, yet the known universe is likely itself just a fraction of the real diversity of life.

Biases in the AFDB are recapitulated by clustering methods

How might biases in database structure influence downstream applications? Given that structural clustering is among the more common uses of protein databases, we decided to assess one of the largest structural clustering datasets currently available: the Foldseek cluster database **[7]**. The Foldseek database comprises ~2.3 million clusters computed from 214 million AFDB proteins using a highly efficient structural clustering workflow **[7]**. Structural clustering is putatively able to identify remotely related proteins, allowing aspects of protein family evolution and function to be

potentially gleaned. If it is indeed true that a substantial portion of protein structural space has been sampled — as is often assumed — then large-scale protein cluster databases may be approaching comprehensive representation of protein structural diversity (and, hence, functions) across the tree of life **[7][22]**.

A key step in the Foldseek workflow is the identification of "representative proteins" after an initial sequence-based clustering step (via the MMseqs2 algorithm) **[23]**. Proteins with the highest prediction confidence (pLDDT; predicted local distance difference test) within the MMseqs2 clusters are chosen as representatives. These representative proteins are then used as input to Foldseek **[15]** which, using structural comparisons, identifies a smaller subset of clusters. Given the importance of these proteins for constructing the final clusters, we wondered the extent to which taxonomic bias might be present among the representatives. We hypothesized that, if taxonomic biases in the AFDB data influence prediction accuracy of the AlphaFold model, then these biases should also be present in the Foldseek representatives. Put another way, if there is a relationship between the number of proteins per taxa in the AFDB and pLDDT, taxa that are better represented in the AFDB should also be more likely to occur in the representative protein set.

We found that the taxonomic distribution of Foldseek representatives very closely mirrored that of the full AFDB dataset (Figure 3, A). Phyla that were well represented in the AFDB were, by and large, also well sampled among the representative proteins across the different domains of life (Figure 3, A). There was a strong relationship between the AFDB and Foldseek with respect to the number of proteins per phylum within each (R^2 = 0.92, linear regression; Figure 3, B). The distributions of taxonomic completeness were also strongly related (R^2 = 0.92, linear regression; Figure 3, C, black line). Notably, the strength of this relationship was consistent even when accounting for phylogeny via a phylogenetic generalized least squares (PGLS) regression (R^2 = 0.92, PGLS; Figure 3, C, red line), reinforcing the idea that taxonomic biases in the AFDB are non-randomly distributed. Furthermore, the non-random taxonomic makeup of the AFDB appears to strongly influence pLDDT-based representative protein selection as implemented in methods such as Foldseek.

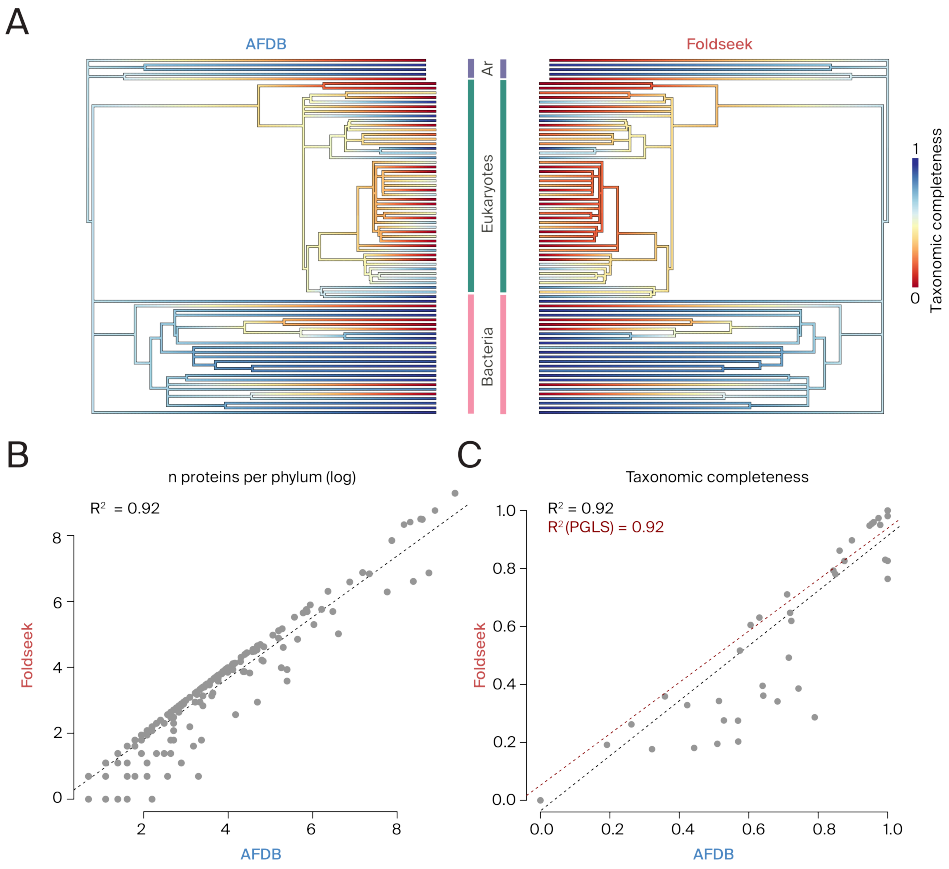

Figure 3

Comparing completeness of the AFDB and Foldseek.

(A) Comparison of the phylogenetic distribution of taxonomic completeness within the AFDB (left) and among Foldseek representative proteins (right).

(B) Distribution of the number of proteins within each phylum for the AFDB and Foldseek (linear regression R^2).

(C) Distribution of per-phylum taxonomic completeness within the AFDB and among Foldseek representative clusters (black line = linear regression; red line = PGLS).

As mentioned previously, it's possible that the concordance between AFDB and Foldseek representative proteins occurs because pLDDT is influenced by taxonomic biases. To explore this possibility, we compared species-level variation in pLDDT to the distribution of representative protein numbers in Foldseek. We reasoned that if higher

pLDDT values are achieved by species with more proteins in the AFDB, then there should be a linear relationship between these measures over the range of representative protein numbers. Indeed, we found that representative protein number was positively correlated with pLDDT (Figure 4, A; Spearman correlation). For example, at a cutoff of 1,500 proteins/species this relationship displayed a plateau of Spearman correlation ~0.7 (Figure 4, A). Interestingly, the correlation coefficients at cutoffs < 150 proteins were negative, suggesting that species contributing lower numbers of proteins had disproportionately high pLDDT values, leading to negative coefficients. Plotting joint distributions between pLDDT and protein number revealed that these correlations were driven by a small number of bacterial species with many proteins possessing mean pLDDT values > 70 (Figure 4, B). This reflects that pLDDT values were stratified by domain: bacterial and archaeal species were associated with significantly greater mean pLDDT than eukaryotic species (Figure 4, B–C; $p < 0.0001$ for both, Dunn's test). It's also notable that the shape of these relationships closely mirrored those seen by Ding & Steinhardt **[14]** when comparing the Progen2 **[24]** and ESM2 **[9]** predictions to the number of per-species input proteins.

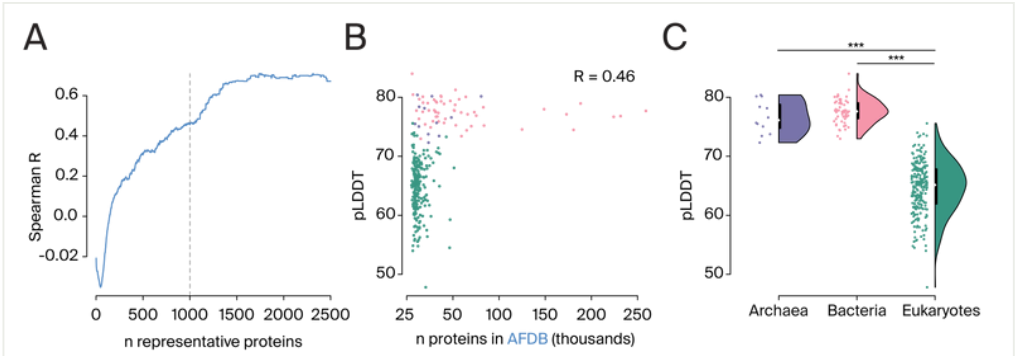

Figure 4

The relationship between training data structure and prediction accuracy.

(A) Distribution of Spearman correlation coefficients over a range of representative protein n cutoffs. The dotted line corresponds to the cutoff exemplified in panels (B) and (C).

(B) The relationship between mean pLDDT of representative proteins (y-axis) and number of proteins in the AFDB (x-axis). Points are colored by domain. (Spearman correlation).

(C) Comparison of mean pLDDT across domains (*** = $p < 0.0001$; Dunn's test).

Taken together, these results suggest that taxonomic biases covary with AlphaFold's pLDDT measurements and can impact downstream applications of AlphaFold that rely on pLDDT, such as Foldseek. This impact can be seen through the strong concordance between the taxonomic makeup of AlphaFold and the representative proteins used in Foldseek's clustering workflow (Figure 3). Notably, this also reflects effects on the behavior of other protein prediction models (Progen2, ESM2) arising from uneven species sampling **[14]**. In these cases, uneven sampling led to systematic biases in the output of protein language models and negatively influenced aspects of protein design **[14]**. A remedy for these issues is more intentional curation of protein datasets **[14]**. With this in mind, we explored how curation of the AFDB would impact the size of the known protein universe.

Data balancing greatly reduces the accessible protein universe

Taxonomic biases in the AFDB are reflective of it being an imbalanced dataset wherein certain classes — namely, taxa — disproportionately contribute. Dataset imbalances can be handled in a variety of ways. A common (and straightforward) approach is undersampling: even numbers of representatives are selected from each class in an attempt to ensure equal contributions from each. Undersampling's simplicity gives it a general utility but also makes it prone to some undesirable behaviors that are worth noting. For example, undersampling can lead to overfitting when working with small datasets and can generate unrealistic representations when classes vary substantially in size. This latter scenario may very well be the case here, as the upper limit of sample sizes will be lower for bacteria (smaller genomes, fewer proteins) than eukaryotes (bigger genomes, more proteins). Despite these caveats, we reasoned that undersampling is likely to be implemented elsewhere as a means for controlling phylogenetic bias and thus could provide a useful first approximation of the effects of data balancing on the makeup of diversity within the AFDB.

To assess the impact of undersampling, we generated a series of balanced datasets selecting partitions containing *n* proteins from each species (from 1 to 20,000 proteins). Species were excluded if they did not have at least *n* proteins in the AFDB. After exclusion, we calculated the phylogenetic diversity of species in each dataset (see Approach).

Balancing had a substantial effect on phylogenetic diversity (Figure 5, A). For example, the transition from a minimum protein *n* of 1 to a minimum *n* of 2 generated a loss of 23% of phylogenetic diversity (Figure 5, A). A minimum *n* of 1000 represented 38% of overall phylogenetic diversity in the AFDB (Figure 5, A). Phylogenetic diversity plateaued around $n = 5{,}000$ at ~5% of diversity captured (Figure 5, A). Diversity was most immediately lost at the species level: 48% of species were pruned when requiring > 2 proteins/species (Figure 5, B). The species distribution was mirrored by that of genera, with both plateauing at ~5% diversity when $n = 5{,}000$ (Figure 5, B). Overall, each taxonomic category lost substantial diversity as dataset partition sizes increased; less than half of phyla were represented when $n = 5{,}000$ (Figure 5, B). These results indicate that a substantial majority of phylogenetic diversity contained in the AFDB is driven by species associated with a small number protein structures, leading

to a rapid decrease in the size of the accessible protein universe after even modest filtering.

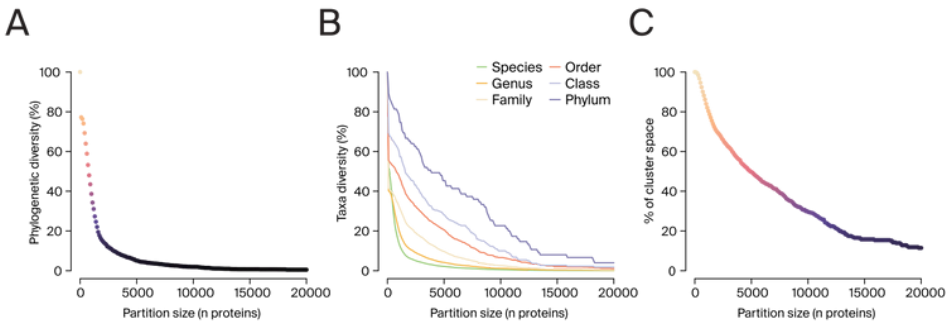

Figure 5

Effects of data balancing.

(A) Proportion of total phylogenetic diversity in the AFDB with increasingly conservative data balancing. Point color corresponds to phylogenetic diversity.

(B) Proportion of total diversity for each level of the taxonomic hierarchy. Colors indicating taxonomic levels are indicated in the upper right hand corner of the plot.

(C) Percentage of Foldseek clusters maintained with increasing conservative data balancing. Point color corresponds to the percentage of cluster space occupied at each cutoff.

We also assessed how data balancing affected the coverage of Foldseek cluster space. While balancing did lead to a consistent decrease in cluster space (Figure 5, C), the relationship was more modest than that observed with phylogenetic diversity (Figure 5, A–B). This robustness to balancing makes sense given that more abundant taxa drive the structure of Foldseek clusters while species with fewer proteins contribute proportionally less (Figure 3). However, though more modest, balancing still resulted in a relatively substantial decrease in the size of Foldseek cluster space, with > 20% of size lost at n = 1,000 and > 50% at n = 5,000 (Figure 5, C). These patterns further support the notion that Foldseek clusters recapitulate the taxonomic makeup of the AFDB.

The data balancing tests described above were agnostic to the real variation in proteome size among species within the AFDB. We hypothesized that, by accounting for proteome size, we might gain an orthogonal view into the effects of taxonomic biases on Foldseek clusters. Specifically, we were interested to see if species with under/over-represented proteomes were better modeled by AFDB and/or contributed more representative proteins in the Foldseek clustering workflow. To test this, we calculated the ratio of AFDB protein number and proteome size for each species (referred to as “protein ratio” in Figure 6). We then compared this protein ratio to the mean pLDDT of each species’ representative proteins and analyzed this relationship over a range of protein *n* cutoffs. This comparison allowed us to infer the effects of prediction accuracy (pLDDT), AFDB representation, and proteome size over sets of species that were increasingly influential on the structure of Foldseek clusters.

We noted a major difference in the behavior of eukaryotic and prokaryotic distributions (Figure 6). While the distribution of eukaryotic species stayed relatively stable over the range of cutoffs (Figure 6, A, i–vi), there was a substantial shift in the prokaryotic distribution (Figure 6, A, i–vi). As cutoffs became more stringent, there was an enrichment for species with very well-sampled proteomes and elevated mean pLDDT measures (Figure 6, A, v–vi). This is again reflective of the strong concordance between the taxonomic distribution of the AFDB and Foldseek representative proteins (Figure 3). It also demonstrates that these latent taxonomic biases are amplified with more conservative data balancing requirements (i.e., larger *n* proteins per species).

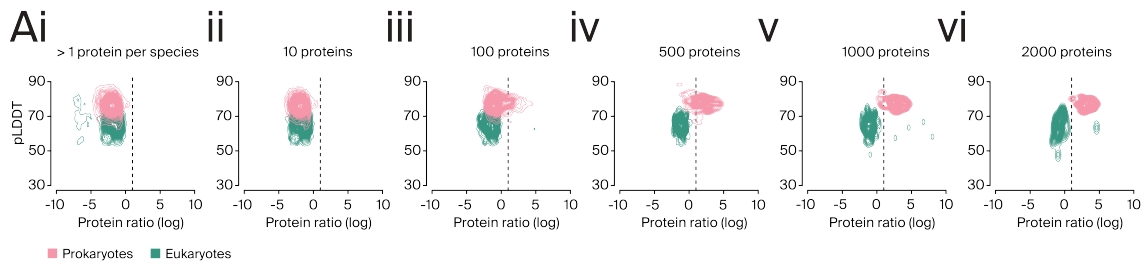

Figure 6

Data biases are amplified by balancing.

Contour plot comparing protein ratio (logarithm of the proportion of proteins in the AFDB and proteins in proteome) and mean pLDDT of individual species calculated over a range of cutoffs (from > 1 protein (i) to 2,000 proteins (vi).

Key takeaways

- Protein databases unevenly sample phylogenetic diversity (Figure 1)
- Sampling biases are taxonomically structured in the AFDB; established prokaryotic phyla are significantly better sampled than eukaryotic phyla (Figure 2)
- Sampling biases are predictive of protein cluster composition (Figure 3)
- Better sampled species possess higher pLDDT values in the AFDB (Figure 4)
- Data balancing leads to a substantial decrease in the phylogenetic diversity of the known protein universe Figure 5)
- Data balancing amplifies phylogenetic disparities in AlphaFold performance (Figure 6)

Implications

This pub lays out approaches to characterizing the structure and biases of the known protein universe. Given the broad scope of contemporary protein modeling, follow-up efforts will inherently be multi-faceted. Below we describe the implications of greatest interest to our work (and likely that of others).

Public protein databases are biased. The utility of protein models will therefore be contingent on whether, and how, training data are curated. Furthermore, generalization beyond natural protein distributions will likely be difficult without mitigating these biases **[14]**. Importantly, though, curation won't be a panacea. As seen here, data balancing decreased accessible phylogenetic diversity and exacerbated latent taxonomic biases in AlphaFold2 prediction accuracy. Appreciation of these constraints may substantially impact future model design, architecture, and implementation.

A simple example: prokaryotic proteins are better sampled in the AFDB than eukaryotic proteins. Better sampling appears to be related to more confident predictions (i.e., higher pLDDT). Better predictions lead to a disproportionate influence on structural clustering. If not accounted for, this bias will likely be recapitulated in other applications. Recognizing these constraints provides options. Treating prokaryotes and eukaryotes independently may make sense in some cases. Alternatively, the bias may be exploited to generate proteins possessing more

prokaryotic-like features. Whatever the goal, bias characterization should play a central role in comparative approaches.

However, even with better curation and model design, there is reason to believe that current approaches will continually fail to capture realistic evolutionary patterns. Most models infer evolutionary patterns (via lengthy and expensive training) by treating proteins as independent observations. This leads models to learn “star phylogenies”: evolutionary hypotheses lacking the hierarchical relationships that are hallmarks of natural diversification **[25]**. Crucially, these representations are very susceptible to a phenomenon known as — in the language of evolutionary biologists — phylogenetic non-independence **[26]**.

Evolution generally functions through gradual changes. Closely related species are likely to have been influenced by the same evolutionary events and, therefore, can be expected to possess similar traits. Given this, the traits of related species cannot readily be considered independent. Incorrect attribution of independence leads to the presence of pseudoreplication (overestimation independent sample number), severely limiting model power **[27]**. Models with pseudoreplication will fail to capture the true structure of the dataset, leading to overfitting and a general lack of interpretability **[26]**.

This may spell trouble for the future progress of protein prediction and design. The known protein universe is already massive, encompassing hundreds of millions of data points. It is (and has been) extremely tempting to believe that we can now learn — and generalize beyond — the generative rules of protein evolution given the sheer volume of the data. And why not? LLM-based chatbots such as ChatGPT achieve impressive feats from similarly sized datasets, learning generative features of human syntax, grammar, and semantics. Shouldn’t this be possible for biological sequences which, at first blush, seem to be not very different from words?

Unmitigated non-independence and phylogenetic biases make this currently unlikely for proteins. The known universe is effectively much smaller than appreciated. As shown here, these patterns vary across taxa and are unevenly distributed across the tree of life. Since the generalization of ML models is dependent on learning the true distributions of underlying data, until addressed, these factors will likely cap the generalizability of protein prediction and design.

There are some potential solutions. Future collection of protein data (i.e., sequences and structures) should be done with the goal of optimizing biological diversity. Undersampled, yet diverse, taxa should be prioritized across the ToL. Measures like

taxonomic completeness can help this type of “phylogenetic data engineering” by helping prioritize efforts and measure progress. This type of targeted approach will help us begin to infer the true distributions of naturally occurring proteins (or even, simply, know if we are getting close).

Finally, it’s worth noting that the statistical power and limitations of any dataset are determined by processes generating the data. For example, human language datasets also display the type of pseudoreplication and non-independence inherent to comparative biological data **[28]**. These are inborn features of language generation that, when unaccounted for, likely limit the generalizability of linguistic models. Luckily, the generative process underlying biological diversity is known: evolution. What’s more, phylogeneticists have been refining and implementing models of diverse evolutionary processes for decades. There are substantial opportunities to leverage evolutionary approaches to confront the biases described here. In general, explicit inclusion of phylogenetic information into protein models may reduce training cost, improve model accuracy, and expand generalizability.

References

1. Huang P-S, Boyken SE, Baker D. (2016). The coming of age of de novo protein design. https://doi.org/10.1038/nature19946
2. Verkuil R, Kabeli O, Du Y, Wicky BIM, Milles LF, Dauparas J, Baker D, Ovchinnikov S, Sercu T, Rives A. (2022). Language models generalize beyond natural proteins. https://doi.org/10.1101/2022.12.21.521521
3. Kortemme T. (2024). De novo protein design—From new structures to programmable functions. https://doi.org/10.1016/j.cell.2023.12.028
4. Jänes J, Beltrao P. (2024). Deep learning for protein structure prediction and design—progress and applications. https://doi.org/10.1038/s44320-024-00016-x
5. Ferruz N, Schmidt S, Höcker B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. https://doi.org/10.1038/s41467-022-32007-7

6 Bairoch A. (2004). The Universal Protein Resource (UniProt). https://doi.org/10.1093/nar/gki070

7 Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. (2023). Clustering predicted structures at the scale of the known protein universe. https://doi.org/10.1038/s41586-023-06510-w

8 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

9 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

10 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. https://doi.org/10.1126/science.abj8754

11 Winnifrith A, Outeiral C, Hie BL. (2024). Generative artificial intelligence for de novo protein design. https://doi.org/10.1016/j.sbi.2024.102794

12 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. https://doi.org/10.1093/molbev/msac174

13 Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL Jr, Xiong C, Sun ZZ, Socher R, Fraser JS, Naik N. (2023). Large language models generate functional protein sequences across diverse families. https://doi.org/10.1038/s41587-022-01618-2

14 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

15 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1038/s41587-023-01773-0

16 Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. https://doi.org/10.1093/database/baaa062

17 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

18 Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. (2010). Picante: R tools for integrating phylogenies and ecology. https://doi.org/10.1093/bioinformatics/btq166

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

20 Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. (2011). caper: Comparative Analyses of Phylogenetics and Evolution in R. https://doi.org/10.32614/cran.package.caper

21 Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. https://doi.org/10.1038/nrmicro3330

22 Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A, Tenson T, Hauryliuk V, Schwede T, Pereira J. (2023). Uncovering new families and folds in the natural protein universe. https://doi.org/10.1038/s41586-023-06622-3

23 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

24 10.1016/j.cels.2023.10.002

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

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

27 Hurlbert SH. (1984). Pseudoreplication and the Design of Ecological Field Experiments. https://doi.org/10.2307/1942661

28 Winter B, Grice M. (2021). Independence and generalizability in linguistics. https://doi.org/10.1515/ling-2019-0049