Metagenomic-scale analysis of the predicted protein structure universe

Jingi Yeo[1,2,#], Yewon Han[1,2,#], Nicola Bordin[3,#], Andy M. Lau[4], Shaun M. Kandathil[4], Hyunbin Kim[1,2], Eli Levy Karin[5], Milot Mirdita[1], David T. Jones[3,4,] ✉ , Christine Orengo[3,] ✉ , Martin Steinegger[1,2,6,7] ✉

Affiliations:
1 - School of Biological Sciences, Seoul National University, Seoul, Korea
2 - Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
3 - Institute of Structural and Molecular Biology, University College London, London, UK
4 - Department of Computer Science, University College London, London, UK
5 - ELKMO, Copenhagen, Denmark
6 - Artificial Intelligence Institute, Seoul National University, Seoul, Korea
7 - Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea
# these authors contributed equally
Correspondence to: d.t.jones@ucl.ac.uk, c.orengo@ucl.ac.uk, martin.steinegger@snu.ac.kr

Abstract

Protein structure prediction breakthroughs, notably AlphaFold2 and ESMfold, have led to an unprecedented influx of computationally derived structures. The AlphaFold Protein Structure Database now provides over 200 million models, while the ESM Metagenomic Atlas includes more than 600 million predictions from uncultured microbes. Here, we combine these two resources into the AFESM, an 821-million-entry dataset, and cluster them using a two-step pipeline based on sequence and structure similarity, yielding 5.12 million non-singleton structural clusters. We identify common ancestors and biomes for these clusters to explore their environmental diversity and specificity, and we investigate their domain composition for structural novelties. Initial ESMfold-based predictions revealed no novel domain folds, re-predicting 2.3 million proteins with AlphaFold2 yielded only one new fold, suggesting near-saturation of the domain space and limitations of predictors. Nevertheless, we discovered many previously unseen domain combinations, highlighting how ESMatlas expands coverage of the known protein fold space. In particular, we find 11,941 multi-domain architectures not observed before, underscoring the importance of metagenomic data for illuminating underexplored regions of the protein structural universe.
Availability An interactive webserver and data are available at afesm.foldseek.com.

Introduction

Advancements in protein structure prediction methods, notably AlphaFold2 [1] and ESMFold [2], have led to an unprecedented expansion of available protein structural data. The AlphaFold Protein Structure Database (AFDB) [3] now contains over 214 million highly accurate protein structure models predicted by AlphaFold2, while the ESM Metagenomic Atlas (ESMatlas) contributes more than 600 million ESMFold-predicted structures of proteins derived from metagenomic assemblies in the MGnify database [4]. This dramatic expansion of structural data has revealed unexpected evolutionary relationships [5] and highlighted novel protein domains [6] that extend our functional and evolutionary view of the protein universe. Yet, the potential of the ESMatlas remains largely underexplored.

Organising and interpreting such a volume of data requires efficient computational tools. Recent algorithmic advancements, such as Foldseek [7], have enabled the rapid clustering [5] of proteins based on structural similarity. Using Foldseek, the 214 million structures in AFDB could be organised into approximately 2.3 million distinct structural clusters. This clustering not only reduced redundancy but also uncovered intriguing structural relationships, including previously unknown connections between components of the human innate immune system and bacterial proteins.

Another large-scale effort has been made to segment AFDB into constituent domains, which are the fundamental units of protein structure and function, and organise them in the Encyclopedia of Domains (TED) [6]. TED identified nearly 365 million domains across more than a million taxa. This work also included over 100 million domains previously undetected by sequence-based methods and reveals that 77% of non-redundant domains are similar to existing domains classified in the CATH [8] database, demonstrating the effectiveness of structure-based approaches. Existing superfamilies, particularly those with few experimentally derived structures, are greatly expanded by AFDB domain models. Additionally, domains with no similarity to structures deposited in the Protein Data Bank (PDB) and other major domain structure classification resources unveiled over 7,000 putative novel folds across the tree of life.

Large-scale clustering and domain-dissection approaches have already highlighted the breadth of structural diversity in AFDB. Notably, a recent integrative analysis revealed a shared functional landscape across AFDB clusters, the 36 million (4.5%) high-quality subset of the ESMatlas, and the Microbiome Immunity Project [9].

Unlike AFDB, which primarily focuses on protein sequences from culturable and well-characterised organisms, the ESMatlas derives its data from uncultured metagenomic sequences, capturing the vast, untapped 'dark matter' of the protein universe [10]. This dataset represents a unique opportunity to explore novel protein structures and domain combinations that are absent in existing structural databases and also measure their abundance in nature. We hypothesised that the ESMatlas will greatly expand our knowledge of protein structural diversity, offering insights into unexplored regions of the protein fold space and evolutionary processes.

To investigate the largely uncharted potential of these metagenomic protein structures, we

compiled a joint resource of 821 million predicted models from AFDB and ESMatlas, then performed large-scale sequence and structural clustering using MMseqs2 [11,12] and Foldseek. We annotated these clusters by identifying putative domains using TED, assigning taxonomic labels, and integrating biome (environmental) metadata. This allowed us not only to expand the known protein structural repertoire with metagenomic proteins but also to quantify their distribution across diverse taxa and highlight structures potentially adapted to specific habitats. Moreover, it enabled us to conduct a systematic dissection of structural novelty, from uncovering genuinely new folds to pinpointing novel combinations of domains, thereby offering deeper insights into the evolutionary processes that shape protein architecture across diverse ecosystems.

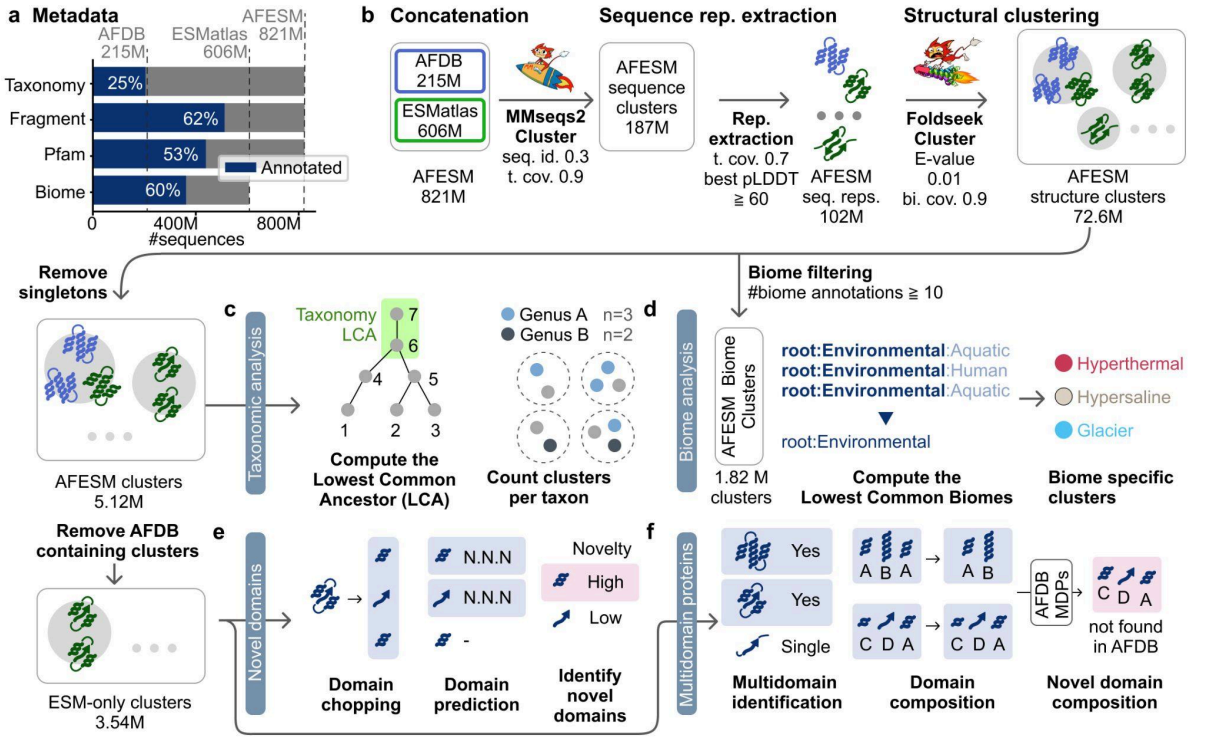


Figure 1. Overview of database, clustering steps and branches of analysis.
a, Distribution of the metadata over AFDB, the ESMatlas and their concatenation in this study: the AFESM. Fractions of the AFESM with various annotations are indicated in blue (Biome annotations are available for the ESMatlas only) **b,** Clustering pipeline. The AFESM was clustered using an above-the-twilight-zone sequence identity (seq. id.) cutoff and a target to query coverage (t. cov.) that associates fragments with longer representatives. Cluster representatives were selected as the protein with the highest predicted Local Distance Difference Test (pLDDT) that is also longer than 70% of the longest protein in the cluster, discarding clusters consisting of only poor predictions (pLDDT < 60). These representatives were clustered by their structures using Foldseek Cluster, applying E-value and bi-directional coverage (bi. cov.) thresholds. **c,** For each of the 5.12 million non-singleton clusters, we computed the Lowest Common Ancestor (LCA) based on the taxonomic annotations of all sequences within each cluster. **d,** We computed the lowest common biomes (LCB) of the 1.82 million AFESM clusters with least 10 biome annotations to identify biome-specific protein structures. **e,** We segmented the representatives of the clusters with

no AFDB entries among their members (ESM-only clusters), into putative domain regions, identified domains, and characterised their novelty. **f,** We identified proteins with multiple domains and characterised their domain composition. Comparing the composition against the domain arrangements in AFDB, we identified novel domain co-occurrences in ESM-only clusters.

Results

Clustering of 821 million predicted protein structures

We first constructed a computational pipeline consisting of two-step clustering and various taxonomic and functional analyses (see Methods, **Fig. 1**) and applied it to the combined set of 821 million entries from AFDB and ESMatlas, henceforth referred to as the "AFESM".

To reduce redundancy among proteins with similar sequences and to remove protein fragments (81% of the ESMatlas), we performed sequence-based clustering using MMseqs2 Cluster [12] to group fragments with their corresponding full-length proteins, if available. This clustering resulted in 187 million clusters, with the longest protein in each cluster proposed as the representative by MMseqs2. Focusing on structural information, we substituted an MMseqs2-proposed representative with a member protein if it satisfied two conditions: (1) its length was at least 70% of that of the representative, and (2) it had the highest predicted Local Distance Difference Test (pLDDT) score among all cluster members that met condition 1. Clusters represented by proteins with a pLDDT value lower than 60 were excluded, resulting in a set of 102 million high-quality representative proteins.

To discover protein similarities within the twilight zone [13], we performed structure-based clustering on these representative proteins using Foldseek Cluster [5,7](see Methods). This effort resulted in 72.6 million AFESM structural clusters. Of these, 5.12 million were non-singleton AFESM clusters, much less frequent than the fraction of non-singleton clusters in AFDB [5] (7% and 27% non-singleton clusters in the AFESM and AFDB, respectively).

We then assessed the quality of the 5.12 million non-singleton clusters. We found that the median values of pLDDT and TM-scores of member-to-representative alignments were 0.73 and 0.70, respectively (**Fig. 2a**), indicating generally high structural similarity within clusters. Additionally, due to the removal of low pLDDT structures and singletons during clustering, the percentage of entries with Pfam annotations increased from 53% to 70%, before and after clustering, while the percentage of fragments decreased from 60% to 55% (**Fig. 1a vs. 2c**).

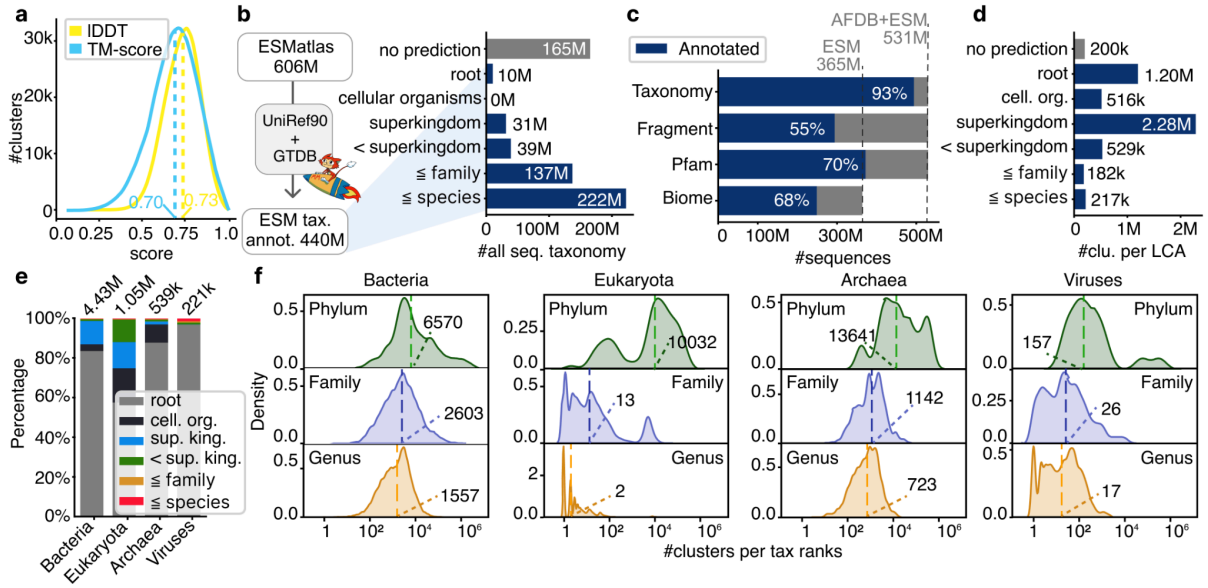


Figure 2. Structural similarity within AFESM clusters, their associated and projected metadata and taxonomic analysis. a, Distribution of structural similarity within AFESM clusters measured as the average over all member-to-representative TM and lDDT scores. **b,** Millions of unannotated ESMatlas entries were assigned taxonomic labels in this study primarily to the species level. **c,** Distribution of the metadata over the members of AFESM clusters after projection of labels. **d,** LCA computation labels of AFESM clusters at various taxonomic ranks, predominantly at the superkingdom level. **e,** Taxonomic distribution of AFESM clusters by superkingdom (Bacteria, Eukaryota, Archaea, and Viruses), showing the fraction of clusters at each taxonomic rank (LCA) containing at least one sequence from the respective group. **f,** The distribution of cluster counts per taxon at different taxonomic ranks. The vertical dashed lines indicate median counts.

Structural Diversity Across Taxa

Next, we explored structural diversity across different taxa. Since the ESMatlas does not include taxonomic information, we predicted taxonomic labels for the sequences by searching the ESMatlas sequences against UniRef90 [14] and the Genome Taxonomy Database (GTDB) [15] using MMseqs2 taxonomy [16], labelling 440 million (72%) of the ESMatlas. Projecting these and AFDB labels onto the clustered AFESM resulted in the taxonomic labelling of 93% of the non-singleton cluster members, compared to 25% before this effort (**Fig. 1a vs. 2b-c,** see Methods). Among these, 222 million annotations were resolved to species or more specific levels (**Fig. 2b**).

We integrated this information with species-level taxonomy annotations from AFDB and calculated the LCA for each cluster (**Fig. 2d**, see Methods). A total of 23.4% (1,198,802) of the non-singleton clusters were assigned the label “root”, indicating a structure found across Viruses and at least one of the three superkingdoms of cellular organisms. Similarly, approximately half of the clusters (2,283,970) were assigned to “cellular organisms”, encompassing sequences from two or more superkingdoms. Around 20% (927,603) were resolved to ranks below superkingdom, while 4% (199,879) remained unclassified.

We assessed the taxonomic ranks of clusters by superkingdom, including viruses (**Fig. 2e**). For Bacteria, Archaea, and Eukaryota, most clusters were resolved at the superkingdom level, indicating consistent classification within these lineages. By contrast, over 90% of clusters containing viral proteins could only be placed at the root of the taxonomic tree. This reflects likely frequent horizontal gene transfer in viruses and limitations in the current viral taxonomy framework rather than a mutual origin across cellular superkingdoms.

Next, we assessed the structural proteome diversity at different taxonomic ranks by counting the number of clusters that included any given taxon or its descendants. These counts were performed at phylum, family, and genus levels and visualised for each superkingdom (**Fig. 2f**). In Eukaryota, genera with one or two clusters mostly belonged to kingdoms such as Metazoa and Viridiplantae, which have larger cell sizes compared to prokaryotes or other eukaryotic supergroups such as the SAR clade (Stramenopila, Alveolata, and Rhizaria) [17]. These low counts thus likely reflect the bias of metagenomic approaches due to filtering larger cell sizes during extraction (**Supplementary Fig. 1**). In Viruses, the two most prevalent phyla are *Uroviricota* (165,923 clusters) and *Nucleocytoviricota* (30,955 clusters), covering 97% of the phylum-containing clusters (200,902 clusters). This suggests that the structures in the dataset are biased to dsDNA Viruses. At the phylum level, the structure counts increase with the number of species that belong to the phylum (**Supplementary Fig. 2**).

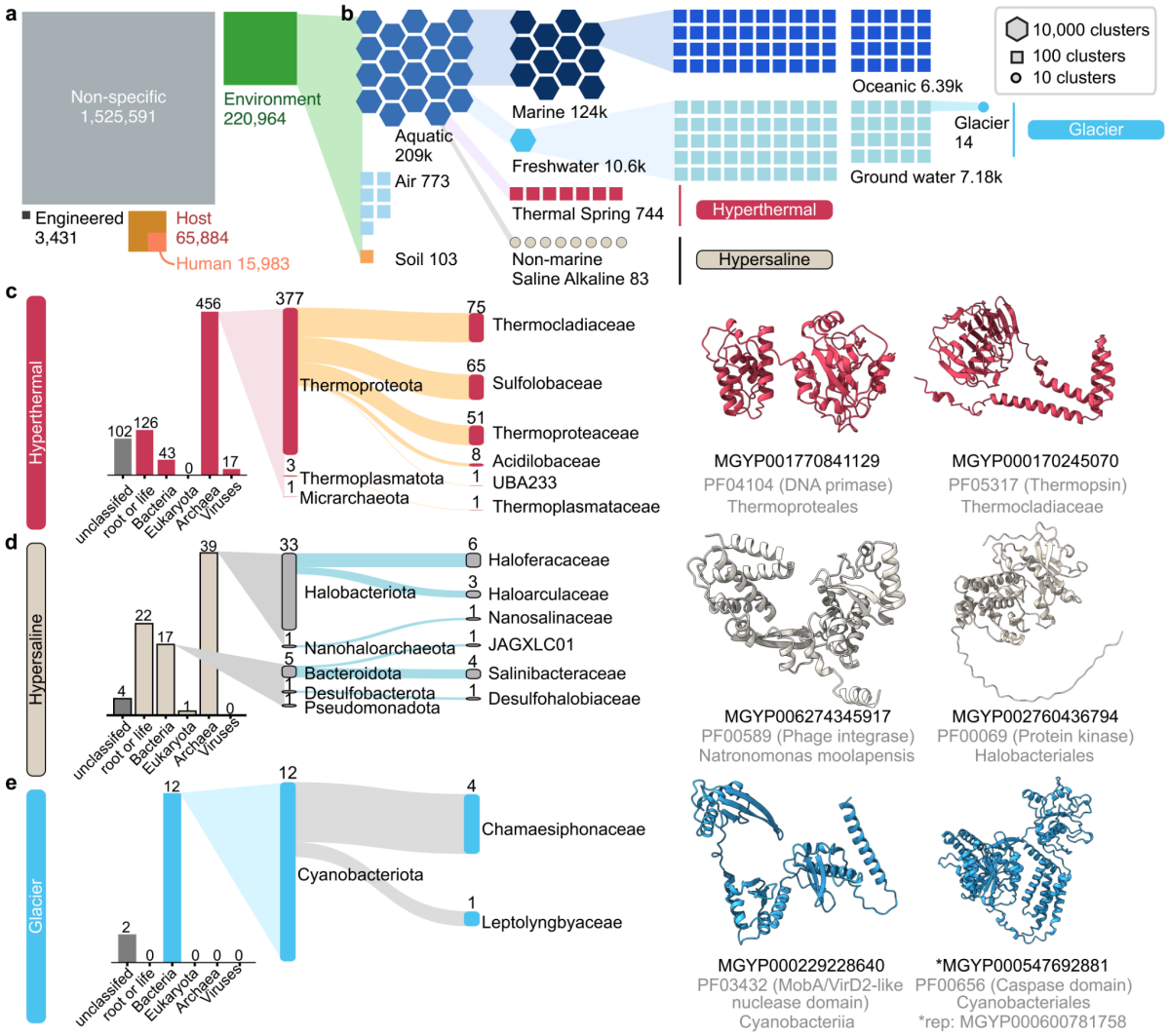


Figure 3. AFESM Biome analysis. a, LCB labels computed for the 1.82 million AFESM biome clusters covering 259 million annotations (see Results) indicate 12% are environmental (green). **b,** Further breakdown of the environmental labels reveals clusters unique to extreme environments: Hyperthermal, Hypersaline, and Glacier. **c-e,** Examination of the taxonomic labels for these clusters asserts the presence of clades known to be associated with extreme environments, such as archaeal Thermoproteota for Hyperthermal-associated clusters (c). On the right of each panel **c-e** are example proteins from each environment with their Pfam annotations.

Biome specific protein structures

Organisms adapt to their environments by altering their protein-coding sequences in response to different environmental pressures [18]. We could directly associate 362 million proteins (44% of the AFESM) with a biome label through the MGnify resource [4]. To reduce biases when determining the biome of clusters, we focused only on 1.82 million clusters with at least 10 biome annotations per cluster, hereafter “AFESM biome clusters”. These clusters included 259 million (71.7%) of the available biome annotations. For each cluster, we computed the lowest common biome (LCB) utilising the hierarchical structure of the annotations, analogously to LCA computation (see Methods). Most (83.5%) of the AFESM biome clusters were not specific to a single environment (root) and the rest were associated

exclusively with environmental biomes (12%), host-associated biomes (3.6%, including 0.9% that were human-specific) or engineered biomes (0.2%) (**Fig. 3a**).

We analysed protein structures associated with eight distinct environments, including three extreme biomes (**Fig. 3b**, see Methods): thermal springs (744 clusters), hypersaline environments (83 clusters), and glaciers (14 clusters). The LCAs of these clusters were consistent with ecological expectations: thermal spring-specific clusters were predominantly linked to archaeal Thermoproteota phylum, whereas hypersaline-specific clusters were mainly linked to archaeal Halobacteriota and bacterial Bacteroidota (Salinibacteraceae). Glacier-specific proteins were primarily associated with Cyanobacteria, particularly Chamaesiphonaceae and Leptolyngbyaceae, both commonly found in glacial habitats [19–21] (**Fig. 3b-e**). Beyond these, we examined AFESM biome clusters from five additional non-extreme environments. These also revealed dominant taxa that aligned with their environments: Air (fungal Basidiomycota), Human digestive system (Bacillota, Bacteroidota, Actinomycetota, and *Uroviricota*) [22,23], Human skin (Malasseziaceae) [24], Marine habitats (Marinisomatota) [25], and Soil (Acidobacteriota) [26] (**Supplementary Figs. 3–8**). Overall, these findings underscore that protein structures unique to particular environments are expressed by organisms adapted to those conditions.

We then investigated the function of proteins found in extreme environments using Pfam [27] annotations. In the hyperthermal clusters (266 total Pfam annotations), we identified a protein putatively involved in DNA primase activity [28,29], indicating potential adaptation to high-temperature conditions. We also discovered a protein harbouring a thermopsin domain, consistent with thermostable proteolytic activity typically found in hyperthermal habitats [30] (**Fig. 3c**, two rightmost panels).

In one hypersaline cluster (total Pfam annotations: 46), we found proteins containing domains linked to phage integrase [31,32]. Since this cluster is taxonomically classified as archaeal (*Natronomonas moolapensis*) and has structural similarity to phage proteins in the BFVD [33], these domains may reflect phage–host interactions specific to the environments (**Fig. 3d**, two rightmost panels). Additionally, while previous examples highlight highly specific functions, we also observed proteins associated with broader functions, such as kinases (**Fig. 3d**, two rightmost panels). Among glacier-derived clusters (total Pfam annotations: 3), proteins with MobA/VirD2-like nuclease domains were identified (**Fig. 3e**, two rightmost panels), known to be involved in horizontal gene transfer [34]. We also discovered a glacier-specific protein containing a caspase domain in Cyanobacteria, which is consistent with previous metagenomic studies reporting diverse metacaspases in this group [35].

We observed a broad range of protein functions, from DNA primase activity to gene transfer regulation, suggesting that adaptation to extreme environments involves modifications across diverse protein structures rather than relying solely on a few key proteins. Also, we structurally compared the examples to their closest counterparts in the non-specific biome clusters using FoldMason [36]. This analysis indicated structural features unique to the examples and absent from their strongest matches (lowest E-values), potentially related to their adaptation to the environments (**Supplementary Fig. 9**).

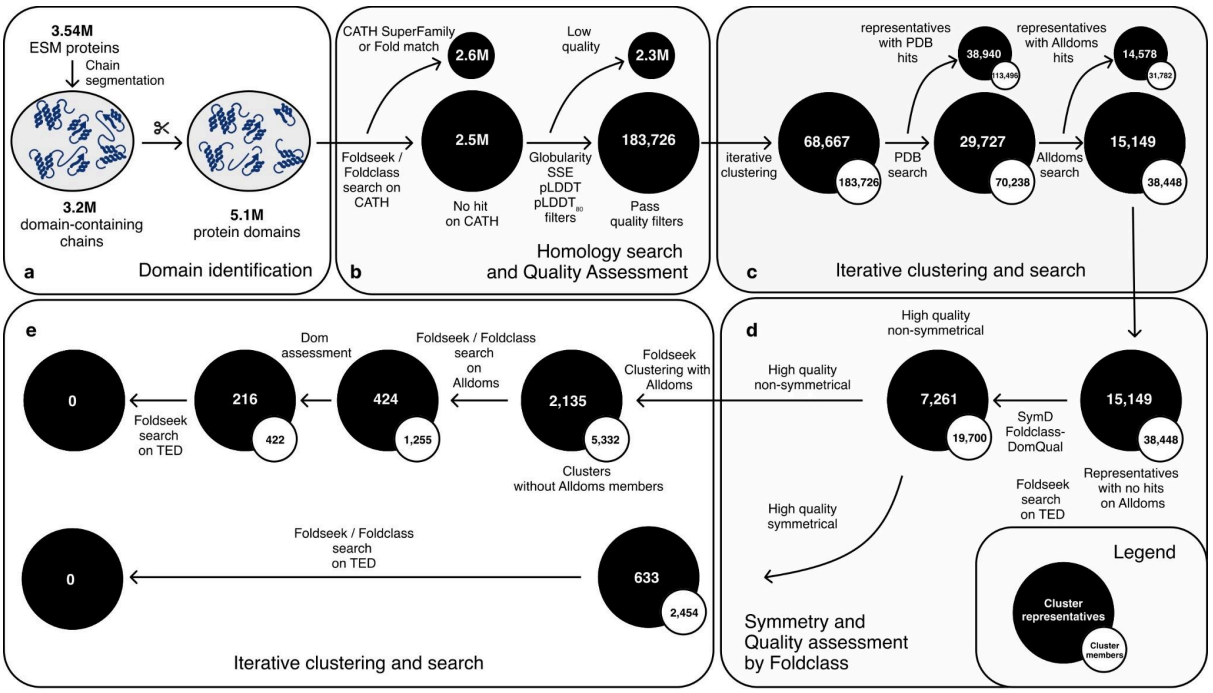


Figure 4. Novel domain identification workflow. a, From the full ESM-only chains to constituent domains using a consensus-based segmentation approach. **b,** Domains are searched for known homologs against representatives of CATH clustered at 5Å (CATH Structural Similarity Group 5, SSG5) and quality-assessed using various metrics, including pLDDT, globularity and amount of secondary structure elements). Homology search against CATH Structural Similarity Group (SSG) and quality assessment. **c,** Iterative clustering using Foldseek easy-cluster and search with Foldseek easy-search against a database containing representatives from CATH, SCOP and ECOD (Alldoms database). **d,** Quality assessment by a Foldclass-derived predictor. **e,** Further iterative clustering with Foldseek Cluster, search against Alldoms and TED followed by symmetry calculation using SymD.

Identifying novel folds unique to the ESM Metagenomic Atlas

We next set out to assess the domain composition of the 3.54 million ESM-only clusters (**Fig. 1e**) in AFESM. To that end, we applied the domain identification protocol outlined by Lau et al.[6] to the 3,213,406 protein chains associated with these clusters. We then identified their boundaries using Chainsaw [37], Merizo [38] and UniDoc [39], resulting in 5,114,907 high and medium consensus domains, where at least two predictors agree on their boundaries (**Fig. 4a**). Over 51% of the domains, 2,617,230 showed similarity to domains already classified in CATH (see Methods) and the remaining 2,497,677 were carried forward as putative novelties. We then removed poor quality models from consideration (pLDDT <70, $pLDDT_{80}$<90, fewer than six secondary structure elements, or non-globular; see Methods), retaining 183,726 sufficient-quality domains (**Fig. 4b**). These domains were clustered iteratively with Foldseek easy-cluster with increasingly stringent cutoffs, resulting in 68,667 cluster representatives. We further excluded 38,940 representatives for 113,496 domains for which we could find structural matches in the PDB (accessed September 2024). The remaining unmatched representatives were searched against the “Alldoms” database comprising the latest versions of CATH, SCOPe and ECOD, retaining 15,149 cluster representatives for 38,448 domains not showing similarity to this resource (**Fig. 4e**). Domains at this stage were assessed for internal symmetry, with 633 highly symmetric being

scanned against TED with Foldseek, removing them from further consideration for novelty.. Domains with no internal symmetry underwent a further quality assessment step with Foldclass DomQual [40], removing 7,255 clusters and their members from consideration (**Fig. 4d**). The remaining 7,261 cluster representatives were pooled with Alldoms and clustered (see Methods), with 5,424 domains clustering with other Alldoms members, suggesting similarity with domains classified in other resources. After this step, 2,135 domains were searched by means of structure similarity and embedding similarity using Foldseek and Merizo-search on Alldoms (see Methods). Ultimately, only 424 cluster representatives did not show similarity to known domain structures derived from the PDB. To verify that they were indeed single domains and were not mapping to domain databases due to having incorporated more than one domain, we assessed each of the 424 cluster representative structures and their boundaries using Dom[41]. 208 cluster representatives for 833 members were indeed multi-domain structures.
To assess novelty compared to the AlphaFold Database [3], we performed an exhaustive search on the TED database, and found no novelties from ESMatlas as all domains have fold-level matches in TED (**Fig. 4e**).

Intrinsic limitations of ESMfold in detecting novel folds

When compared to the 7,427 putative novel folds identified in the AFDB by the TED resource, we cannot identify additional domain novelties in the ESMatlas using the same discovery procedure, despite the high species diversity and sheer size of the ESMAtlas (**Supplementary Fig. 10a**). Three main questions arise from this behaviour:

Is the Known Fold Space Nearly Complete? First, is there a common 'domain toolbox' shared across all domains and metagenomes, comprising folds already present in the AlphaFold Database? As previous work utilising AlphaFold to model metagenomic sequences at large scale identified novelties not present in PDB [42] it is unlikely that we have now seen all folds in nature. Particularly, the majority of novel folds identified in TED are unique to bacterial sub-groups (4,982/7,427, or 67%), with a large part of the tree of life inaccessible to TED due to AFDB omitting a majority of the viral landscape [33,43].

Are Novel Folds Obscured by Low-Confidence Predictions? Second, are novel folds hidden in the discarded portion of our dataset where domain boundaries are not so confident (i.e. a low-confidence consensus prediction)? As shown in **Supplementary Fig. 10b**, we noticed a lower pLDDT for domains where the boundaries were identified with a consensus of two out of three methods, with a higher confidence seen in high-consensus models, i.e., all three methods agreed. In our large-scale TED analyses of AFDB, we observed that models with poor quality were very likely to have low-confidence consensus prediction.

Intrinsic Limitations of ESMFold for Novel Structures. We therefore ask whether ESMFold, and generally protein language model-based predictors, are suitable to explore these large metagenome datasets due to their inability to access homologs beyond their training data and the dependency to structures in the training set.
ESMatlas domains predicted by ESMFold are, even at high confidence, often small and non-globular (**Supplementary Fig. 10c, 11a**). Additionally, predictions that show low similarity to known templates from the PDB often return a lower confidence for both pLDDT and $pLDDT_{80}$, (**Supplementary Fig. 11c**) and while this behaviour is present in AlphaFold

and AFDB as well, AFDB contains a plethora of models that show no similarity to known structures, while retaining high confidence and plausible structural arrangements, benefiting from the additional information derived from their input MSAs [33,44].
To test ESMfold's performance we modelled all 7,427 novel folds identified by TED using ESMFold and noticed low confidence in the ESMfold models whereas the corresponding AlphaFold models all passed our quality metrics outlined in Methods (**Supplementary Fig. 11d**). While ESMFold is a good choice for an initial exploration of metagenomic space due to its speed and accuracy in modelling known structures, however when sampling structurally uncharacterised regions of sequence space it struggles to build good quality models.

Structure Rescue: AlphaFold2 Reveals Quality Structures Discarded by ESMFold
Subsequently, we used ColabFold [45] to predict 2.3 million domains, corresponding to 27% of residues in the ESM-only protein set, which were discarded due to their low-quality predictions by ESMFold, but might be rescued if modelled by AlphaFold2. We improved the pLDDT for 415,022 domains (18% of the dataset) (**Supplementary Fig. 11d-e**). We also revisited 2.3 million domains previously discarded due to low quality ESMfold predictions, and found that ColabFold could rescue 98,454 of them based on our quality thresholds —SSE content, pLDDT, pLDDT80, globularity, and gyration scores. However, only one non-symmetrical new fold (MGYP002252011424_02_256-274_315-478) was identified that is not already present in TED.

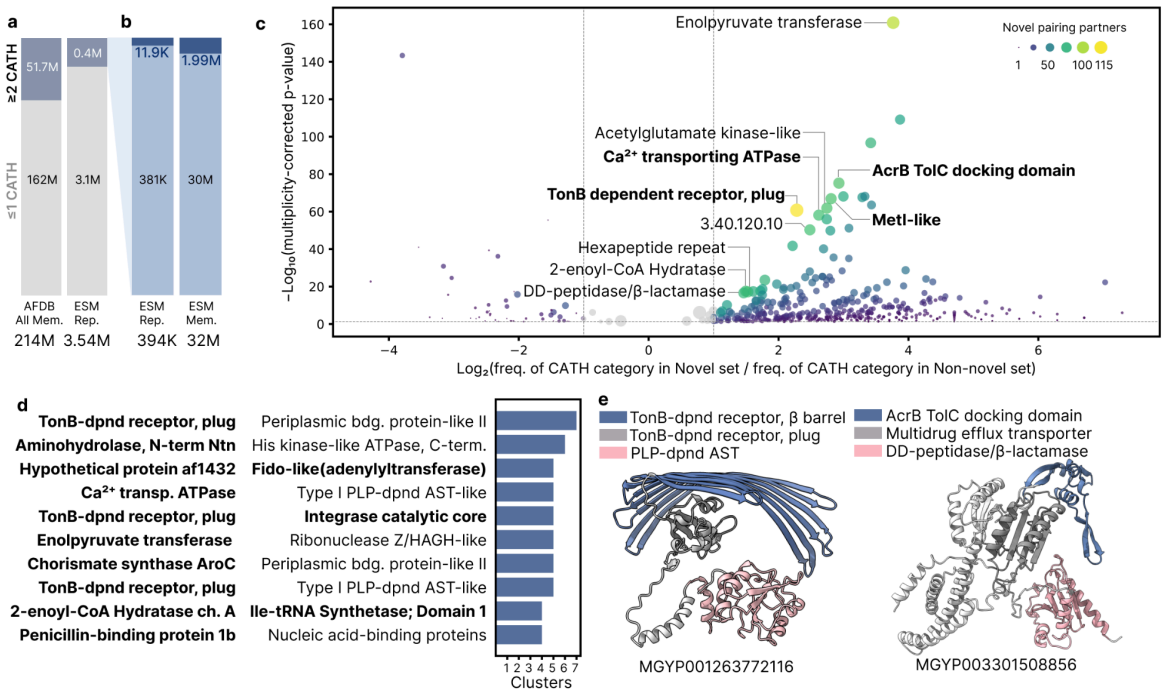


Figure 5. Domain combinations of ESM-only multi-domain proteins (MDPs). a, Fraction of MDPs: structures with ≥2 non-identical Topology-level (T-level) CATH annotations (blue) among AFDB members (left) and ESM-only representatives (right). **b,** Fraction of MDPs with *novel* (dark blue) co-occurring T-level CATH domains within ESM-only representatives (left) and members (right). **c,** Volcano plot of 530 H-level CATH domains that showed significant over/underabundance (see Methods) between ESM-only MDPs with novel domain combinations ("Novel set") and all other ESM-only MDPs ("Non-novel set"). The x-axis indicates overabundance (right) or underabundance (left) in the Novel set and the y-axis is a

function of the statistical *p*-value of each domain after multiplicity correction. Each domain is depicted as a circle whose colour and size reflect the number of its unique novel pairing partners, with textual labels for the top 10 and bold font for membrane-associated domains. **d,** Top 10 co-occurring H-level CATH domain pairs in the Novel set. Domains found to be overrepresented in the Novel set compared to the Non-novel set in panel **c** are highlighted in bold; the rest didn't show a significant difference. **e,** Examples of novel combinations of membrane-associated CATH labels.

Identifying combinations of known domain folds unique to the ESMatlas

Although 3.54 million representative structures are unique to the ESMatlas (ESM-only), the single novel domain we found cannot explain this finding. We therefore assessed how many of these unique proteins arise from previously unobserved combinations of established CATH domains rather than entirely new domains. For this we investigated the multi-domain protein (MDP) architectures of these representatives, focusing on 393,793 proteins with at least two different Topology level (T-level) CATH categories (**Fig. 5a,** See Methods). In order to assess how many of these harbour a CATH combination that is unique to the ESMatlas, we compared them to 51.7 million proteins from AFDB that matched this criterion (**Fig. 5a**). This revealed 11,941 clusters with MDP architectures containing T-level CATH combinations that are unique to the ESM-only set (**Fig. 5b,** See Methods). When examining the novel MDP architectures, we focused on a subset of 5,203 “Novel” MDP architectures containing ≥2 distinct H-level CATH categories — the most granular CATH classification — assigned to globular domains (see Methods). Using the same criteria, we also identified in the ESM-only set 134,576 “Non-novel” H-level architectures whose domain combinations can be found in AFDB.

The joint set of 139,779 MDP architectures consisted of 2,906 unique H-level categories: 1,233 were found both in the Novel set and in the Non-novel set, 33 were unique to the Novel set, and 1,635 were unique to the Non-novel set. That is, despite being over 20× smaller (5,203 vs. 134,576 MDPs), the Novel set captured 44% (1 - (1,635/2,906)) of the categories, indicating greater domain diversity. To further characterise the Novel set, we analysed the enrichment of CATH categories between the two sets. We tested the prevalence of the 2,906 H-level categories in the two sets using Fisher’s exact or Chi-squared tests, applying Benjamini–Hochberg correction (see Methods). This identified 530 H-level categories that were significantly over- or underrepresented in the Novel set (**Fig. 5c**, **Supplementary Table 1**). Of these, 431 (81.3%) were overrepresented in the Novel set, consistent with its broader domain repertoire relative to its size. Of the 10 significantly overrepresented H-level domains with the most pairing partners in the Novel set, four are membrane-associated: “Met I transporter”, “Multidrug efflux transporter AcrB”, “TonB dependent receptor”, and “Ca^{2+}-transport ATPase Cytoplasmic domain” (**Fig. 5c**, **Supplementary Table 2,** See Methods). Furthermore, H-level categories that are most prevalent in CATH are underrepresented in the Novel set (**Supplementary Fig. 12**), suggesting that the discovery of combinations among well-studied domains is nearing saturation. In contrast, novel combinations of domains predominantly involve least frequent H-level categories in CATH, highlighting the potential of this analysis to uncover functional combinations involving underexplored domains.

All ten most frequently co-occurring domain pairs in the Novel set include at least one domain category that is significantly overrepresented in the Novel set (**Fig. 5d**). We

observed pairings that appear unexpected—for instance, an outer membrane component paired with an integrase catalytic core domain, or with an aminotransferase-like domain. Unexpected domain pairings were also found in pairs involving the most overrepresented domains (**Supplementary Table 3**), such as combinations of cytosolic and nuclear domains, or cytochrome oxidoreductase and urease peptidase. One example for a protein with unexpected domain pairing is MGYP001263772116 (**Fig. 5e left**). This protein contains TonB beta-barrel domains and a PLP-dependent aminotransferase domain at the N-terminus—the former typically associated with the outer membrane and the latter with the cytosol, based on PDB entries linked to their respective CATH domains. MGYP001263772116 also includes a TonB plug domain. Such a multi-domain architecture deviates from the conventional TonB-dependent transporter structure [46], where the N-terminal TonB box is replaced by the aminotransferase-like domain, which may be an alternative mechanism for transport regulation. Another example is MGYP003301508856, which consists of Multidrug efflux transporter beta barrel domain and DD-peptidase/β-lactamase (**Fig. 5e right**). This novel pairing was observed twice, suggesting a possible mechanism for hydrolysis localised at membranes.

Our MDP analysis suggests that novel combinations of known CATH domains drive the functional diversification of many metagenomic proteins.

Limitations

The AFESM dataset contains a high fraction of fragments (56%), mostly from ESMatlas (81%). Clustering fragmented proteins by sequence and structure can miss entries which could be aligned as a member at its full length, leading to high numbers of singletons and misassigned members.

In taxonomy prediction, we observed discrepancies between reference databases. For Bacteria and Archaea, we used GTDB (a genome database), whereas for Eukaryota and Viruses, we used UniRef90 (a protein database). These database differences likely contributed to the multi-modality in protein count distributions in Eukaryota and Viruses, a pattern absent in Bacteria and Archaea. Metagenomic sampling differences also introduced biases, particularly for Eukaryota and Viruses. For the former, the median of just two protein structures per genus likely underestimates true diversity due to a metagenomic sampling bias toward small organisms, whereas for viruses, these same biases underscore the difficulty of accurately assigning viral sequences to taxonomic groups.

The biome lineages defined by our LCB approach are less clearly delineated compared to taxonomic hierarchies, largely due to limitations in the biome annotation system (e.g., "Hydrothermal vents" and "Thermal springs" are not within the same lineage). Additionally, many proteins in the ESMatlas lack consistent or complete biome metadata and some biomes have strong sampling biases such as marines or humans, further hindering the discovery of structures specific to particular environments.

In the MDP analysis, novel protein structures comprised only 0.3% of the ESM-only clusters and 2.3% of their 821 million members (**Fig. 5b**). This analysis is limited by incomplete CATH coverage, excluding MDPs with fewer than two CATH annotations. In our analysis, we

did not consider domain order, copy number, and spacing between domains, thus CATH superfamily names may not fully reflect domain function in multi-domain contexts.

Discussion

In this study, we analyzed 821 million predicted protein structures, generating AFESM, a combined and clustered resource of the AFDB and ESMAtlas. AFESM defines 5.12 million non-singleton structural clusters, consisting of 64.6% of the 821 million protein structures. Our analyses revealed taxonomic conservation of structures across broad evolutionary lineages, alongside structural adaptations associated with distinct environmental biomes. Although genuinely novel structural folds appear rare, possibly constrained by current protein language-model prediction methodologies, we discovered substantial novelty through previously unobserved domain combinations—identifying 11,941 unique multi-domain architectures. This underscores domain recombination, rather than fold discovery, as the primary mode of currently discoverable structural innovation captured within the ESMatlas.

This contribution via novel domain combinations is dwarfed by the parts that escape discovery within the 3.54M ESM-only cluster representatives (**Supplementary Fig. 13**). Definitive structural novelty was rare (~0.3%, comprising no new folds and ~11.9k combinations), 9.3% lacked identifiable domains, and 19% contained known domains (in known arrangements), however with potential additional unidentified regions. The majority (69%) consisted of fragments with known domains (16%) or were filtered out due to low prediction quality (53%, **Fig. 4b-e**). This large pool of fragments (likely due to poor metagenomic assembly [47] or poorly predicted structures may harbour additional structural novelty discoverable with full-length sequences or more robust prediction methods.

Given the many novel domain combinations we identified, future studies could leverage genomic context—such as gene neighborhoods—and multimer structure prediction methods like AlphaFold-Multimer [48], combined with large-scale comparative tools such as Foldseek-Multimer [49], to extend our analyses from single-chain proteins to multimeric complexes. We hypothesise that a significant fraction of the structural and functional diversity captured by metagenomic sequences may be encoded in protein multimers rather than monomers [50].

In conclusion, this study represents the first comprehensive structural analysis integrating nearly one billion predicted protein structures from both under-characterised metagenomic sequences (ESMatlas) and the known protein universe (AFDB). We established a computational framework to systematically explore structural diversity at unprecedented scale, we uncovered structural adaptations specific to environmental niches, notably within extreme habitats. While entirely novel domain folds seem elusive—likely reflecting either genuine biological constraints or methodological limitations—the numerous previously unseen domain combinations we discovered emphasise the critical role of domain recombinations in evolutionary innovation. Our approach demonstrates how large-scale integration of isolated and metagenomic datasets can illuminate the vast, yet underexplored regions of the protein structural universe.

Methods

Preparation of ESMatlas dataset

We obtained the ESM Metagenomic Atlas (ESMatlas) dataset by downloading tar archives for both the v0 and v2023_02 releases using aria2c from the ESM repository (https://github.com/facebookresearch/esm). To optimise data handling, these combined archives were compressed using Foldcomp [51], a protein structure-specific compression format. During compression, entries with discontinuous protein chains were excluded via the `--skip-discontinuous` option. This process, applied to the merged content of both ESMatlas releases, yielded 605,776,560 protein structure entries used for the subsequent clustering analyses.

Clustering the AFESM dataset

We performed the following steps: (1) Sequence clustering by MMseqs2 cluster with following parameters

```
--min-seq-id 0.3 --cov-mode 1 -c 0.9
```

(2) We initiated the selected representative sequence from each cluster by taking its longest sequence (mark its length as 'x'). Then, we examined all cluster members whose length was >0.7x, substituting the representative with the highest pLDDT sequence among them. Then, we removed 84,463,443 clusters (151,488,016 sequences) whose representative had a pLDDT score lower than 60 (**see Code Availability**), leaving 102,104,329 representatives.

(3) Structure clustering of the 102,104,329 entries using Foldseek cluster with following parameters

```
-c 0.9 -e 0.01
```

(4) During the clustering workflow, we repeatedly re-clustered representative protein structures. Because cluster representatives from one iteration become input to the next, this can result in cluster memberships that violate our defined thresholds (e.g. coverage or e-value). After the initial clustering, 38,386,513 entries formed non-singleton clusters, but 10,355,205 (27%) failed to meet our parameters. We therefore introduced a reassignment step by reassigning members that violated the thresholds to more appropriate clusters or grouping them into new clusters, reducing threshold violations of members from 10.3 million to 141,686—a 98.6% improvement among 34,624,609 non-singleton members.

```
--cluster-reassign 1
```

Taxonomy prediction and integration

We predicted the taxonomic labels of the ESMatlas dataset using MMseqs2 Taxonomy [16] with a new reference database, which was generated by merging the GTDB and UniRef90 databases as follows. Using the *databases* command, we downloaded GTDB (Apr 24, 2024/v220), a reference for Bacteria and Archaea and UniRef90 (Nov 8, 2023), which

includes sequences from the entire tree of life. By default, UniRef90 associates each sequence (cluster member) with the taxonomic identifier (taxid) of the LCA inferred for its cluster. This tends to associate sequences with the label “root”. To use more informative labels, we instead associated Uniref90 entries with the taxid of the cluster representative instead of its LCA taxid. Next, we filtered out all bacterial and archaeal sequences from UniRef90 using the command *filtertaxseqdb* (excluding taxids 2 for Bacteria and 2157 for Archaea). Then, we combined GTDB with the filtered UniRef90, resulting in a merged reference containing Archaeal, Bacterial, Eukaryotic and Viral references (ABEV) and we predicted the taxids of ESMatlas by aligning against ABEV.

Meanwhile, there were remaining bacterial and archaeal sequences in AFDB with the NCBI taxonomic labels which arise because of a discrepancy of the taxonomy system between the prediction results in ESMatlas (ABEV system) and AFDB (NCBI system). We converted their labels to corresponding GTDB labels. Referring to the table offered by GTDB that describes the mapping information between GTDB taxonomic ids and NCBI taxonomic ids, we computed the taxonomic LCA of the NCBI taxonomic identifiers connected to each GTDB taxonomy due to the many-to-many pairs between the NCBI and GTDB taxonomic IDs. Using the result - one GTDB id and its NCBI taxonomic LCA, we converted the bacterial and archaeal taxonomic IDs annotated to the AlphaFold database to the corresponding GTDB identifiers.

To streamline the results, we applied the same grouping strategy, categorising the assigned taxonomy into the six hierarchical groups as root, cellular organisms, superkingdom, < superkingdom, ≦ family, ≦ species.

Taxonomic Lowest Common Ancestor (LCA) computation

We computed the taxonomic LCA to infer the branching points of clusters using the lca module in MMseqs2. During the analysis, we excluded four taxa: (1) 12,908 unclassified sequences, (2) 28,384 sequences with "other" annotations, (3) 2 (NCBI Bacteria), and (4) 2,157 (NCBI Archaea).

Lowest Common Biome (LCB) computation

We aimed to determine the narrowest biome region associated with each protein cluster in the ESMatlas. The MGnify database defines 297 distinct biomes organised as a lineage tree that starts at the root and branches into four main categories: Environmental, Host-associated, Engineered, and Mixed. These categories further diverge into six hierarchical levels. For each cluster, we identified the LCB, representing the most specific (narrowest) biome element shared within the lineage across all cluster members. We excluded members labelled as “root” or “root:Mixed” from the analysis as their inclusion results in clusters annotated at higher-level biomes ignoring other annotations.

We computed the biome annotation on MGYP90 where ESMatlas entries were derived from. MGnify provides MGYP90 cluster information and biome annotations of the MGYP90 members. To fully utilise data, we included all members of MGYP90 clusters and annotated each ESMatlas entry with the resulting LCBs. For the AFESM Biome Clusters (see Results),

we took this approach by including the biome annotations of all members within the sequence and structure clusters to calculate the LCB.

Protein structures specific to extreme environments

We focused on three extreme environment categories: Hyperthermal, Hypersaline, and Glacier for the main analysis and five environment categories: Air, Human Digestive System, Human Skin, Marine, Soil for the supplementary analysis. The LCB annotations of each cluster were grouped into these categories based on their biome elements (**Table 1**). We examined and plotted the taxonomy distribution in each biome using Pavian [52].

Biome category	Lowest Common Biomes
Hyperthermal	root:Environmental:Aquatic:Thermal springs:Hot (42-90°C) root:Environmental:Aquatic:Thermal springs
HyperSaline	root:Environmental:Aquatic:Non-marine Saline and Alkaline:Salt crystallizer pond root:Environmental:Aquatic:Non-marine Saline and Alkaline:Saline:Microbial mats root:Environmental:Aquatic:Non-marine Saline and Alkaline:Saline root:Environmental:Aquatic:Non-marine Saline and Alkaline root:Environmental:Aquatic:Non-marine Saline and Alkaline:Alkaline
Glacier	root:Environmental:Aquatic:Freshwater:Ice:Glacier
Air	root:Environmental:Air
Human Digestive System	root:Host-associated:Human:Digestive system root:Host-associated:Human:Digestive system:Large intestine:Fecal root:Host-associated:Human:Digestive system:Large intestine root:Host-associated:Human:Digestive system:Large intestine:Sigmoid colon
Human Skin	root:Host-associated:Human:Skin root:Host-associated:Human:Skin:Naris
Marine	root:Environmental:Aquatic:Marine root:Environmental:Aquatic:Marine:Coastal root:Environmental:Aquatic:Marine:Hydrothermal vents root:Environmental:Aquatic:Marine:Intertidal zone root:Environmental:Aquatic:Marine:Intertidal zone:Coral reef root:Environmental:Aquatic:Marine:Intertidal zone:Estuary root:Environmental:Aquatic:Marine:Intertidal zone:Oil-contaminated root:Environmental:Aquatic:Marine:Intertidal zone:Salt marsh root:Environmental:Aquatic:Marine:Marginal Sea root:Environmental:Aquatic:Marine:Oceanic root:Environmental:Aquatic:Marine:Oceanic:Oil-contaminated

	root:Environmental:Aquatic:Marine:Sediment
Soil	root:Environmental:Terrestrial:Soil root:Environmental:Terrestrial:Soil:Permafrost

Table 1. The Biome categories and their corresponding LCBs.

Segmenting the ESMatlas chains into constituent protein domains

The ESMatlas chains underwent a consensus-based domain parsing workflow that incorporated three segmentation methods: Merizo [38], Chainsaw [37], and UniDoc [39]. Each method analysed the structure of the model and predicted domain boundaries. For UniDoc, which does not identify non-domain regions (NDRs), an initial step using Merizo was employed to filter out NDRs before domain segmentation was performed. This preprocessing step is essential, as the presence of NDRs can cause UniDoc to misinterpret entire chains as single domains. By first removing NDRs, the workflow ensured more accurate domain predictions using UniDoc.

ESMatlas domain assignments to CATH superfamilies

Domains obtained via the consensus-based segmentation process were assigned to 31,574 existing CATH Structural Similarity Groups at 5Å (SSG5), where each SSG5 is a cluster representative for CATH domains that superpose within 5Å. The assignment was performed using Foldseek easy-search (version a435618b95ba95cbfe74dcbc8b4bd2720547d285) with the following parameters:

```
--cov-mode 5 --alignment-type 2 -e 0.108662 -s 10 -c 0.366757 -a
--format-output
"query,target,fident,evalue,qlen,tlen,qtmscore,ttmscore,qcov,tcov"
```

Raw outputs were subsequently processed to retrieve assignments at the superfamily level (H-level) or fold-level (T-level) by using the following cutoffs identified in Lau et al. [6] for E-value (0.019000 and 0.108662), coverage (0.366757 and 0.786333), and TM-score (0.560000 and 0.416331), respectively.

Domain-centred quality metrics

Domains were assessed using CATH-AlphaFlow [53], a Nextflow pipeline that calculates metrics on PDB files. These include the average pLDDT of the domain, the number of secondary structure elements (SSEs) identified by STRIDE [54], and globularity metrics based on packing density and normed radius of gyration [53]. Domains were deemed of sufficient quality for subsequent analyses with pLDDT > 70, $pLDDT_{80}$>90 (at least 80% of residues have a pLDDT of 90 and above), at least six SSEs, packing density greater or equal to 10.333 and a normed radius of gyration less than 0.356.

Iterative clustering and search

Domains that passed the quality assessment were clustered with foldseek cluster using the following parameters

```
-s 4 -e 0.001 -c 0.7 --cov-mode 5 --tmscore-threshold 0.56 --min-seq-id
0.2 --seq-id-mode 1 --cluster-mode 1
```

And subsequently with foldseek easy-cluster:

```
-c 0.8 --tmscore-threshold 0.8 --cluster-mode 1
```

Representatives obtained via the two clustering steps identified above were searched against PDB with

```
easy-search models2 PDB result.m8 tmp -c 0.6 --cov-mode 2
--tmscore-threshold 0.5 --alignment-type 1 --format-output
query,target,qlen,tlen,qtmscore,cigar
```

Results were filtered retaining domains with qtmscore greater than 0.56 and query coverage over 0.6 calculated from the cigar alignment.
Domains not meeting these criteria were searched against the Alldoms database with

```
foldseek easy-search models3 alldoms result.m8 tmp -c 0.6
--tmscore-threshold 0.3 --alignment-type 1 --exhaustive-search 1
--format-output query,target,qlen,tlen,qtmscore,cigar
```

Valid domain matches to Alldoms required a qtmscore greater than 0.56, a cigar-computed coverage for both query and target above 0.6.

Quality assessment with Foldclass DomQual

Domain quality was further assessed with a version of the Foldclass [40] neural network with an additional EGNN layer to score domains with a binary output layer, as outlined in Lau et al. Domains with a score below 0.5 were not processed further.

Final exhaustive search and symmetry calculations

Domains with a Foldclass DomQual score over 0.5 were clustered with members of the Alldoms database using foldseek easy-cluster with the following parameters

```
--tmscore-threshold 0.5 -c 0.6 --alignment-type 1
```

Clusters containing ESM-derived domains and Alldoms members were discarded. The remaining domains were searched exhaustively against the Alldoms database (i.e. disabling the pre-filter step in Foldseek, guaranteeing that every domain is searched against all target database members) with

```
foldseek easy-search -c 0.6 --cov-mode 2 --tmscore-threshold 0.3
--alignment-type 1 --exhaustive-search 1 --format-output
query,target,qlen,tlen,qtmscore,ttmscore,rmsd,cigar
```

The output of the search was filtered using the following criteria
`(max(qtmscore,ttmscore) > 0.5 OR rmsd < 3.0) and cigar-computed qcov > 0.6`

Domains not meeting the criteria underwent one final search using Foldclass embeddings validated by TMalign. Hits meeting the following criteria were not considered novel:
`max(qtmscore,ttmscore) > 0.5 and cigar-computed qcov > 0.6`

In order to assess internal symmetry SymD [55] was used with default parameters, labelling domains as symmetrical if the domain Z-score was greater than 9.

Search against The Encyclopedia of Domains

In order to assess exclusivity to the ESMatlas database, we searched domains that underwent all previous steps against a Foldseek database for The Encyclopedia of Domains using the following parameters.

```
-c 0.6 --tmscore-threshold 0.3 --alignment-type 1 --exhaustive-search 1
--format-output query,target,qlen,tlen,qtmscore,cigar
```

Identifying CATH-MDPs in the ESM-only clusters and AFDB

From the set of cluster representatives of the 3.54 million ESM-only non-singleton clusters, we extracted 393,793 MDPs containing at least two CATH annotations—at either the T-level or H-level—with distinct T-level classifications.

In preparation for identifying novel domain co-occurrences in ESM-only proteins, we used the same criteria to identify MDPs in the AFDB. To ensure comprehensive coverage, we used all 214 million AFDB entries (“members”) rather than just the cluster representatives. From this set, we identified 92.9 million MDPs, of which 51.4 million contained at least two CATH-annotated domains, based on annotations from the TED database [6]. Additionally, we included 625K TED domains without CATH annotations but meeting high-quality standards (see 'Domain-centered quality metrics'). These domains were searched against the CATH database using the Foldseek exhaustive search with below specified parameters. For 25% of the 625K domains containing discontinuous regions, the starting and ending points were treated as domain boundaries.

```
--exhaustive-search 1 --format-output
query,target,qlen,tlen,qtmscore,cigar
```

For the domains identified with CATH hits through the Foldseek search results, the hits with the highest target TM scores were selected as CATH domain annotations, augmenting the TED dataset to create a comprehensive set of 51,741,054 CATH-annotated MDPs within AFDB.

Identifying MDPs with novel CATH domain pairings

To identify domain co-occurrences unique to ESM-only proteins, 393.8K MDPs from the ESM-only, non-singleton cluster representatives were compared against 51.7M MDPs from AFDB. Each MDP was mapped to a non-redundant, order-invariant set of CATH T-level categories (e.g., an MDP with three domains: '1.10.10.10', '1.10.10', and '2.20.20.20' maps to {1.10.10, 2.20.20}). All possible unordered domain pairs were extracted per MDP and compared to those from AFDB to identify novel pairs exclusive to the ESM-only set.

An MDP was considered "novel" if it contained at least one pair not observed in the AFDB; 11.9K MDPs met this criterion (the "Novel set"), while the remaining 381.5K formed the "Non-novel set." Novelty was assessed at the T-level rather than H-level for two reasons: (1) it maximises explanatory coverage using available data, and (2) it applies a stricter criterion, reducing potential false positives caused by inconsistencies of structural diversity within each H-level categories.

We further selected MDPs with ≥2 globular CATH H-level annotations as they are likely functional. From the 11.9K Novel MDPs, 5,203 met this criterion; from the 381.5K Non-novel set, 134,576 were selected. For example, MGYP2 ({Globular:2.20.20.20, Globular:3.30.30.30}) is included, while MGYP1 ({Globular:1.10.10.10, Non-globular:1.10.10.20, Non-globular:3.30.30}) is excluded due to having only one globular H-level domain.

Composition analysis of the Novel MDP architectures

We compared H-level CATH category representation between the Novel (5,203 MDPs) and Non-novel (134,576 MDPs) sets. Together, the sets consisted of 2,906 unique H-level domains and the frequency of each domain in each set was computed as the number of occurrences of that domain in the set divided by the total number of MDPs in the set. Using these frequencies, we computed the differential prevalence of each domain as $\log_2$(frequency_Novel / frequency_Non-novel).

We examined associations between the prevalence of each of the 2,906 domains and the type of set ("Novel"/"Non-novel") as follows. In its turn, we treated each domain as 'x', defining all other domains as 'not-x'. We then counted the occurrences of 'x' and 'not-x' in the Novel and Non-novel sets, constructing a 2×2 contingency table to test if the null hypothesis stating "no difference in the prevalence of x between Novel and Non-novel" can be rejected. Using the contingency table we performed a Fisher's exact test for 2,511 of the domains and a Chi-squared test with Yates' correction for 395 of the domains. Fisher's test was used when low counts caused errors in the Chi-squared test.. The tests were conducted using R version 4.4.3 and the complete list of 2,906 P-values was corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure [56] with FDR < 0.05, as the cutoff.

For the top 10 by the number of their unique novel pairing partners, we examined the FunFam [57] descriptions corresponding to their CATH identifiers on the CATH website (accessed April 2025), to ensure the consistency between the CATH name and the FunFam. H-level category names were assigned using `cath-names.txt` (CATH v4.3.0). To determine membrane association of CATH categories, we queried the CATH web service for

GO ‘Cellular Component’ terms linked to PDB entries. Categories were labelled as membrane-associated if the majority of annotations indicated membrane localisation (e.g., 87.4% of GO terms for 2.170.130.10 correspond to "outer membrane").

Software used for analysis

MMseqs2: MMseqs2 v14 for clustering, 407b31 for MSA generation of the domains to predict. Foldseek: ef768f for clustering. ColabFold: 1.5.5 (fdf3b235) for the domain prediction. Python 3.11: matplotlib 3.6.2 and pandas 2.2.2 for visualization, R 4.4.3, Pavian 1.2.1

Webserver

We adapted our previously-developed AFDB clusters webserver [5] to allow for exploration and visualisation of the AFESM. This adaptation includes additional metadata. We will implement biome search functionality in the future.

Data and code availability

All scripts to rerun the analysis are available at https://github.com/steineggerlab/afesm-analysis, the data can be explored and downloaded at https://afesm.foldseek.com.

References

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
3. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
4. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
5. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).
6. Lau, A. M. *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946 (2024).

7. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
8. Sillitoe, I. *et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
9. Szczerbiak, P. *et al.* Large protein databases reveal structural complementarity and functional locality. *bioRxiv* (2024).
10. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, e67667 (2022).
11. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
12. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
13. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
14. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
15. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
16. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab184.
17. Alexander, H. *et al.* Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *MBio* **14**, e0167623 (2023).
18. Brininger, C., Spradlin, S., Cobani, L. & Evilia, C. The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles. *Semin. Cell Dev. Biol.* **84**, 158–169 (2018).
19. Singh, P., Singh, S. M., Segawa, T. & Singh, P. K. Bacterial diversity and biopotentials of Hamtah glacier cryoconites, Himalaya. *Front. Microbiol.* **15**, 1362678 (2024).

20. Murakami, T. *et al.* Metagenomics reveals global-scale contrasts in nitrogen cycling and cyanobacterial light-harvesting mechanisms in glacier cryoconite. *Microbiome* **10**, 50 (2022).
21. Varliero, G. *et al.* Glacier clear ice bands indicate englacial channel microbial distribution. *J. Glaciol.* **67**, 811–823 (2021).
22. Hou, K. *et al.* Microbiota in health and diseases. *Signal Transduct. Target. Ther.* **7**, 135 (2022).
23. Pavia, G., Marascio, N., Matera, G. & Quirino, A. Does the human gut virome contribute to host health or disease? *Viruses* **15**, 2271 (2023).
24. Gaitanis, G., Magiatis, P., Hantschke, M., Bassukas, I. D. & Velegraki, A. The Malassezia genus in skin and systemic diseases. *Clin. Microbiol. Rev.* **25**, 106–141 (2012).
25. Pérez Castro, S. *et al.* Degradation of biological macromolecules supports uncultured microbial populations in Guaymas Basin hydrothermal sediments. *ISME J.* **15**, 3480–3497 (2021).
26. Flieder, M. *et al.* Novel taxa of Acidobacteriota implicated in seafloor sulfur cycling. *ISME J.* **15**, 3159–3180 (2021).
27. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.* **53**, D523–D534 (2025).
28. Lao-Sirieix, S.-H., Nookala, R. K., Roversi, P., Bell, S. D. & Pellegrini, L. Structure of the heterodimeric core primase. *Nat. Struct. Mol. Biol.* **12**, 1137–1144 (2005).
29. Zhao, D., Chen, X., Li, K. & Fu, Y. V. The application of thermophilic DNA primase TtDnaG2 to DNA amplification. *Sci. Rep.* **7**, 12809 (2017).
30. Fusek, M., Lin, X. L. & Tang, J. Enzymic properties of thermopsin. *J. Biol. Chem.* **265**, 1496–1501 (1990).
31. Aalto, A. P. *et al.* Snapshot of virus evolution in hypersaline environments from the characterization of a membrane-containing Salisaeta icosahedral phage 1. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7079–7084 (2012).

32. Emerson, J. B. *et al.* Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol.* **78**, 6309–6320 (2012).
33. Kim, R. S., Levy Karin, E., Mirdita, M., Chikhi, R. & Steinegger, M. BFVD-a large repository of predicted viral protein structures. *Nucleic Acids Res.* **53**, D340–D347 (2025).
34. Meyer, R. Identification of the mob genes of plasmid pSC101 and characterization of a hybrid pSC101-R1162 system for conjugal mobilization. *J. Bacteriol.* **182**, 4875–4881 (2000).
35. Asplund-Samuelsson, J. *et al.* Diversity and expression of bacterial metacaspases in an aquatic ecosystem. *Front. Microbiol.* **7**, 1043 (2016).
36. Gilchrist, C. L. M., Mirdita, M. & Steinegger, M. Multiple Protein Structure Alignment at Scale with FoldMason. *bioRxiv* 2024.08.01.606130 (2024) doi:10.1101/2024.08.01.606130.
37. Wells, J. *et al.* Chainsaw: protein domain segmentation with fully convolutional neural networks. *Bioinformatics* **40**, (2024).
38. Lau, A. M., Kandathil, S. M. & Jones, D. T. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat Commun* **14**, 8445 (2023).
39. Zhu, K., Su, H., Peng, Z. & Yang, J. A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics* **39**, (2023).
40. Kandathil, S. M., Lau, A. M. C., Buchan, D. W. A. & Jones, D. T. Foldclass and Merizo-search: embedding-based deep learning tools for protein domain segmentation, fold recognition and comparison. *bioRxiv* (2024) doi:10.1101/2024.03.25.586696.
41. Taylor, W. R. Protein structural domain identification. *Protein Eng.* **12**, 203–216 (1999).
42. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
43. Nomburg, J. *et al.* Birth of protein folds and functions in the virome. *Nature* **633**, 710–717 (2024).
44. Lee, S. *et al.* Petascale Homology Search for Structure Prediction. *bioRxiv*

2023.07.10.548308 (2023) doi:10.1101/2023.07.10.548308.

45. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
46. Noinaj, N., Guillier, M., Barnard, T. J. & Buchanan, S. K. TonB-dependent transporters: regulation, structure, and function. *Annu. Rev. Microbiol.* **64**, 43–60 (2010).
47. Orellana, L. H., Krüger, K., Sidhu, C. & Amann, R. Comparing genomes recovered from time-series metagenomes using long- and short-read sequencing technologies. *Microbiome* **11**, 105 (2023).
48. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
49. Kim, W. *et al.* Rapid and sensitive protein complex alignment with Foldseek-Multimer. *Nat. Methods* **22**, 469–472 (2025).
50. Levy, E. D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci.* **117**, 25–51 (2013).
51. Kim, H., Mirdita, M. & Steinegger, M. Foldcomp: a library and format for compressing and indexing large protein structure sets. *Bioinformatics* **39**, (2023).
52. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz715.
53. Waman, V. P. *et al.* CATH 2024: CATH-AlphaFlow doubles the number of structures in CATH and reveals nearly 200 new folds. *J. Mol. Biol.* **436**, 168551 (2024).
54. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500–2 (2004).
55. Tai, C.-H., Paul, R., Dukka, K. C., Shilling, J. D. & Lee, B. SymD webserver: a platform for detecting internally symmetric protein structures. *Nucleic Acids Res.* **42**, W296–300 (2014).
56. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**,

289–300 (1995).

57. Scheibenreif, L., Littmann, M., Orengo, C. & Rost, B. FunFam protein families improve residue level molecular function prediction. *BMC Bioinformatics* **20**, 400 (2019).

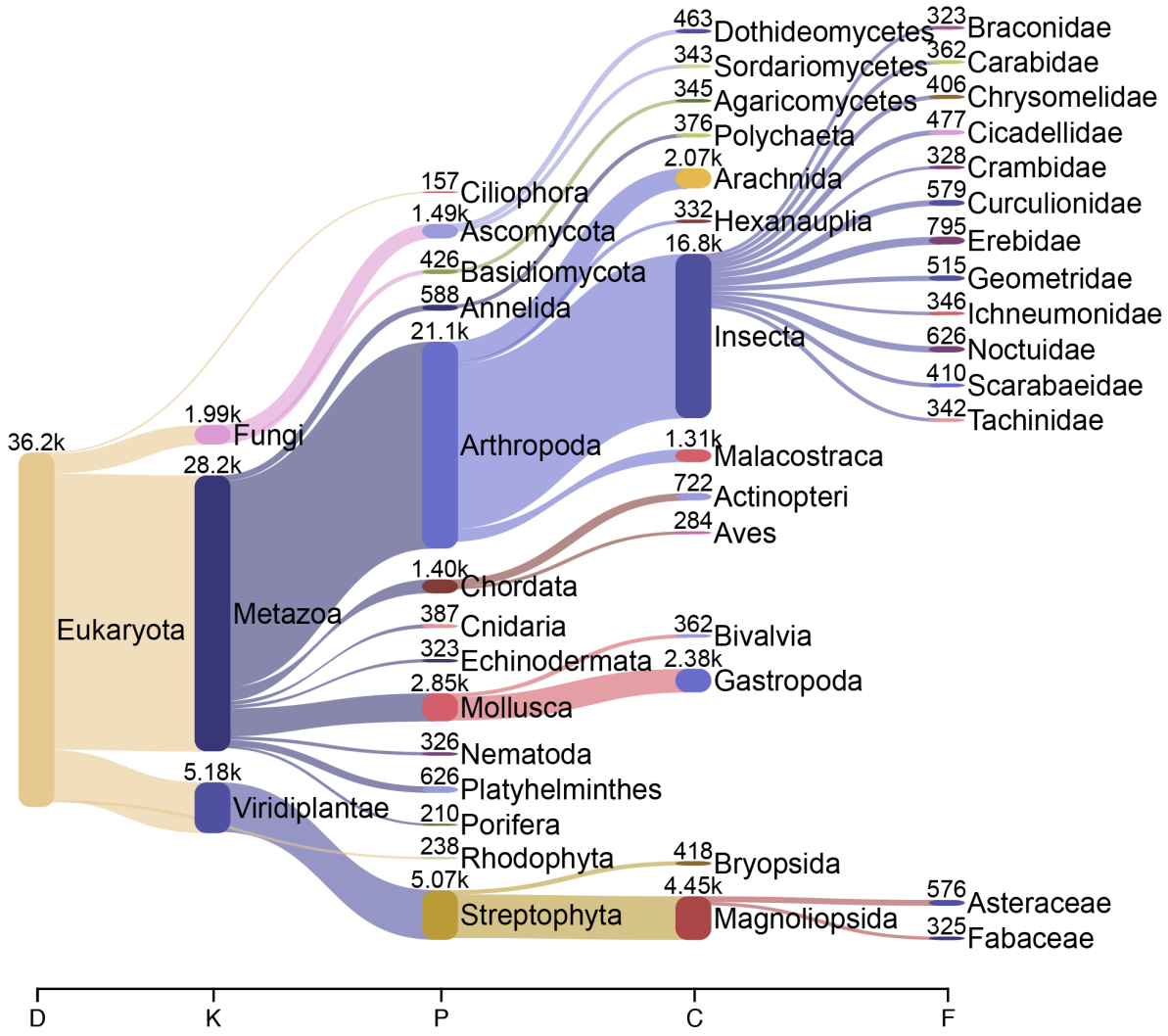
Acknowledgements

We thank Dr. Pedro Beltrao initial discussion regarding the analysis, and Dr. Robert D. Finn for insightful discussions regarding the MGnify database and for maintaining this essential resource, and Dr. Hyunbin Lee for valuable insights on the novel membrane protein examples. M.S. acknowledges support by the National Research Foundation of Korea grants (2020M3-A9G7-103933, 2021-R1C1-C102065 and 2021-M3A9-I4021220, RS-2024-00396026), Novo Nordisk Foundation (NNF24SA0092560), Samsung DS research fund, Creative-Pioneering Researchers Program through Seoul National University. M.M. acknowledges support from the National Research Foundation of Korea (grant RS-2023-00250470). This work was supported by the Biotechnology and Biological Sciences Research Council (grant BB/T019409/1 to A.M.L. and D.T.J., grant BB/W008556/1 to S.M.K. and D.T.J.) and the Wellcome Trust (grant 221327/Z/20/Z for N.B. and C.O).

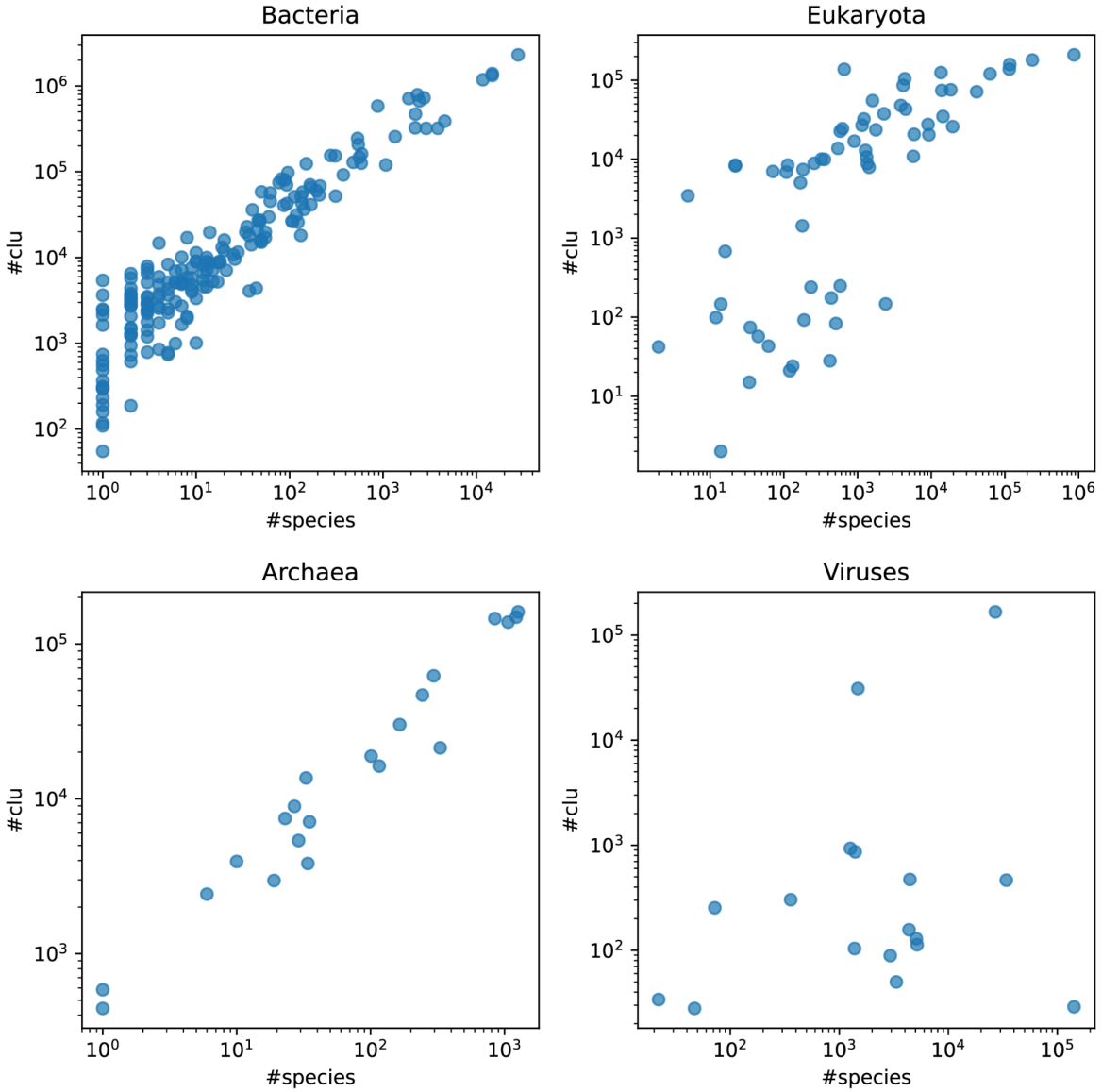
Competing interests

M.S. declares an outside interest in Stylus Medicine.

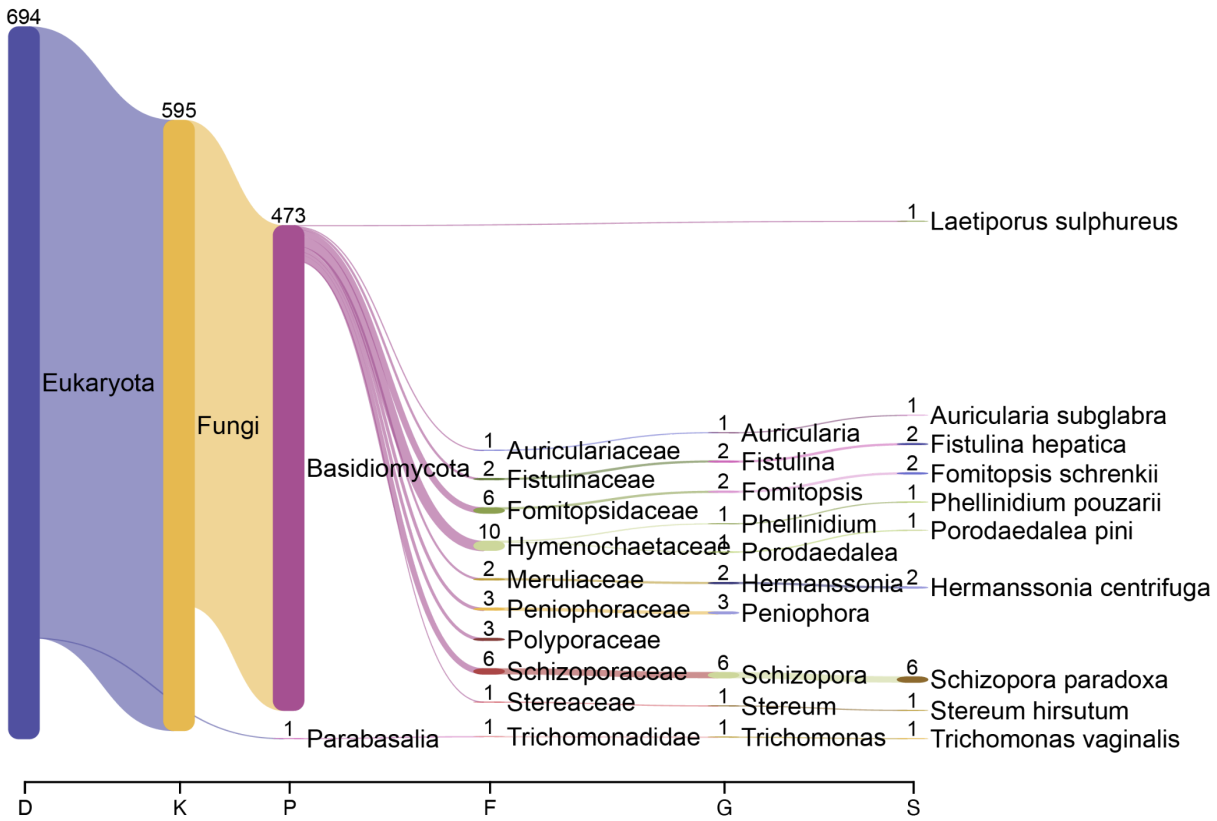
Supplementary materials and methods



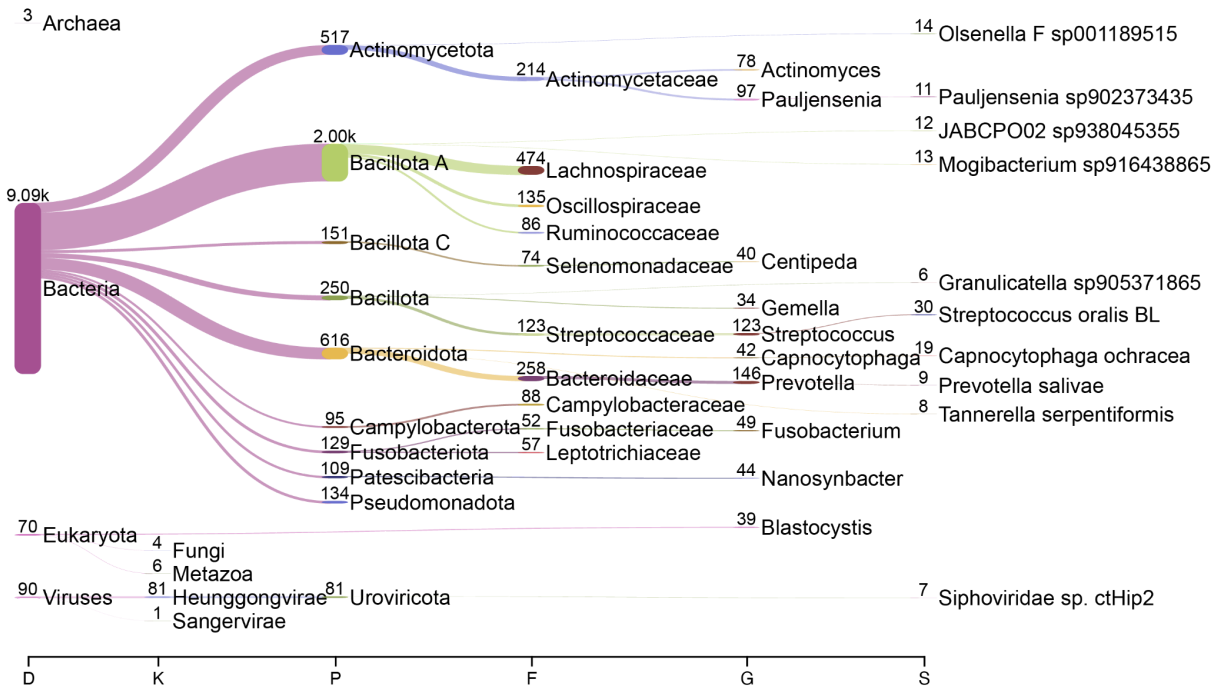
Supplementary Figure 1. Eukaryotic genera found in one or two structural clusters.



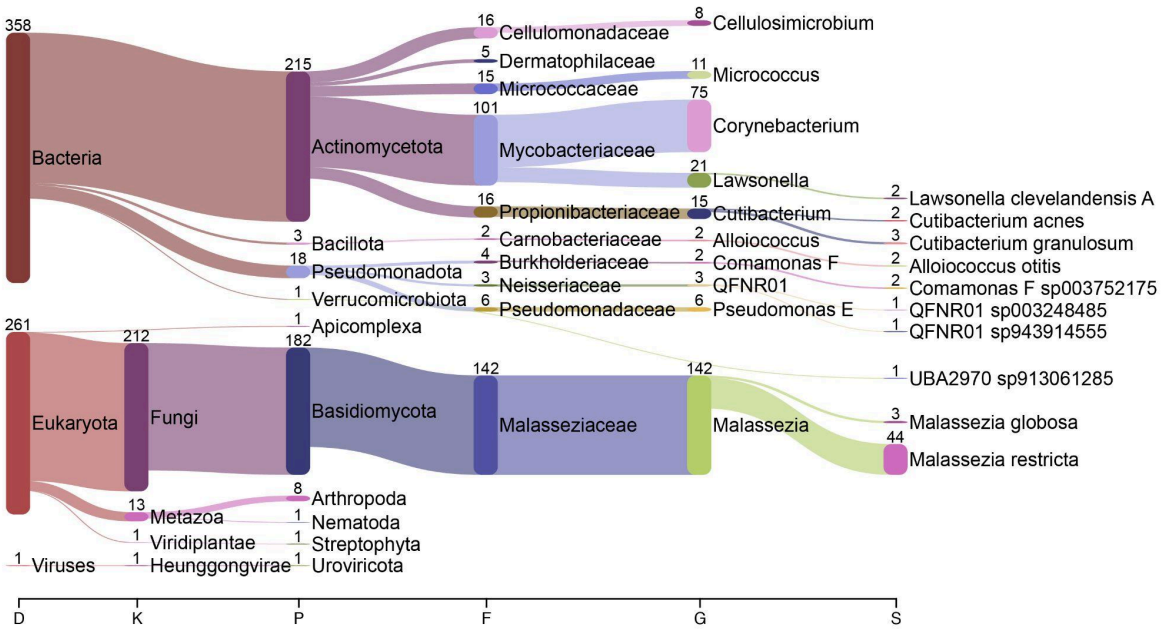
Supplementary Figure 2. The number of descendant species and strains and the number of protein clusters a phylum has in different superkingdoms and Viruses.



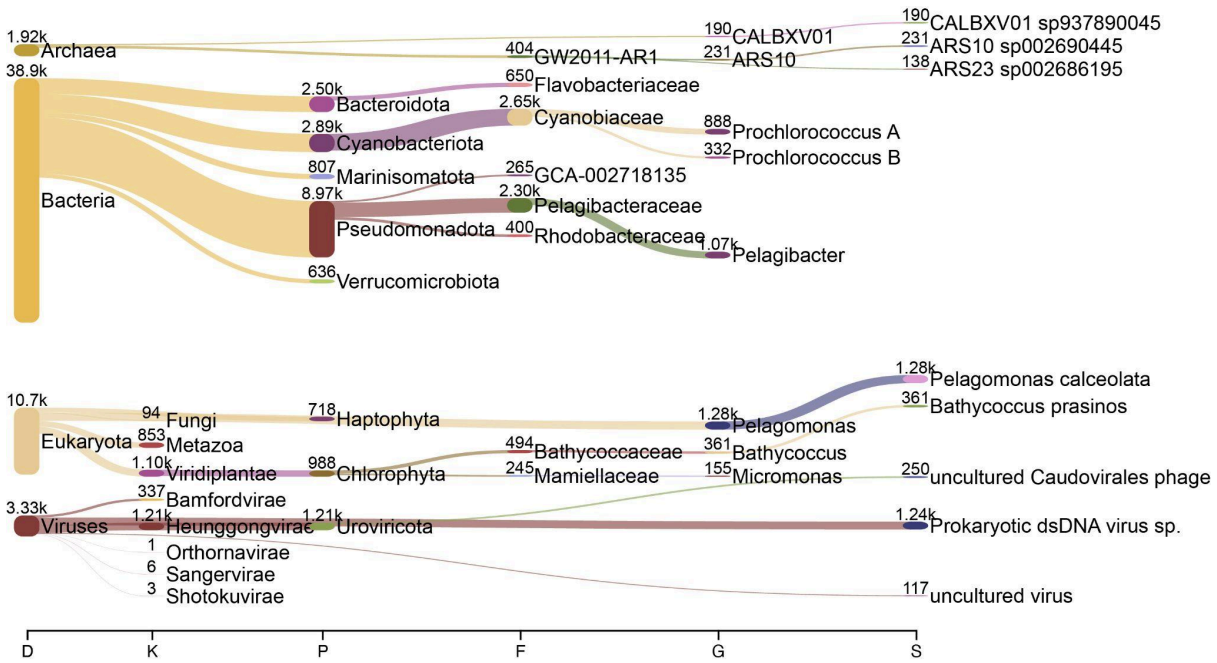
Supplementary Figure 3. The taxonomic distribution of the clusters specific to the biome: Air.



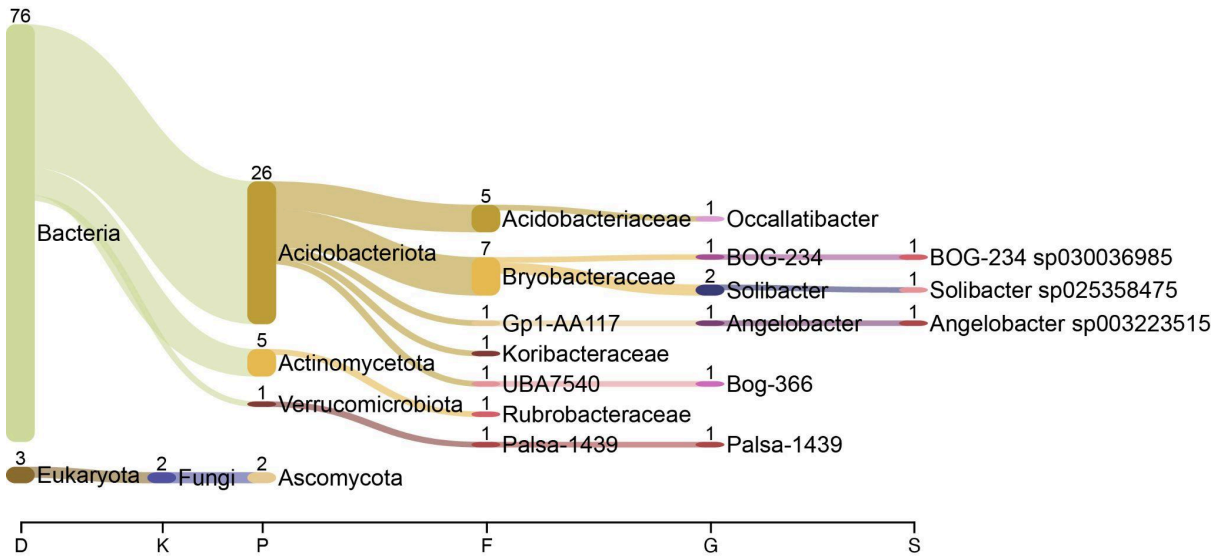
Supplementary Figure 4. The taxonomic distribution of the clusters specific to the biome: Human digestive system.



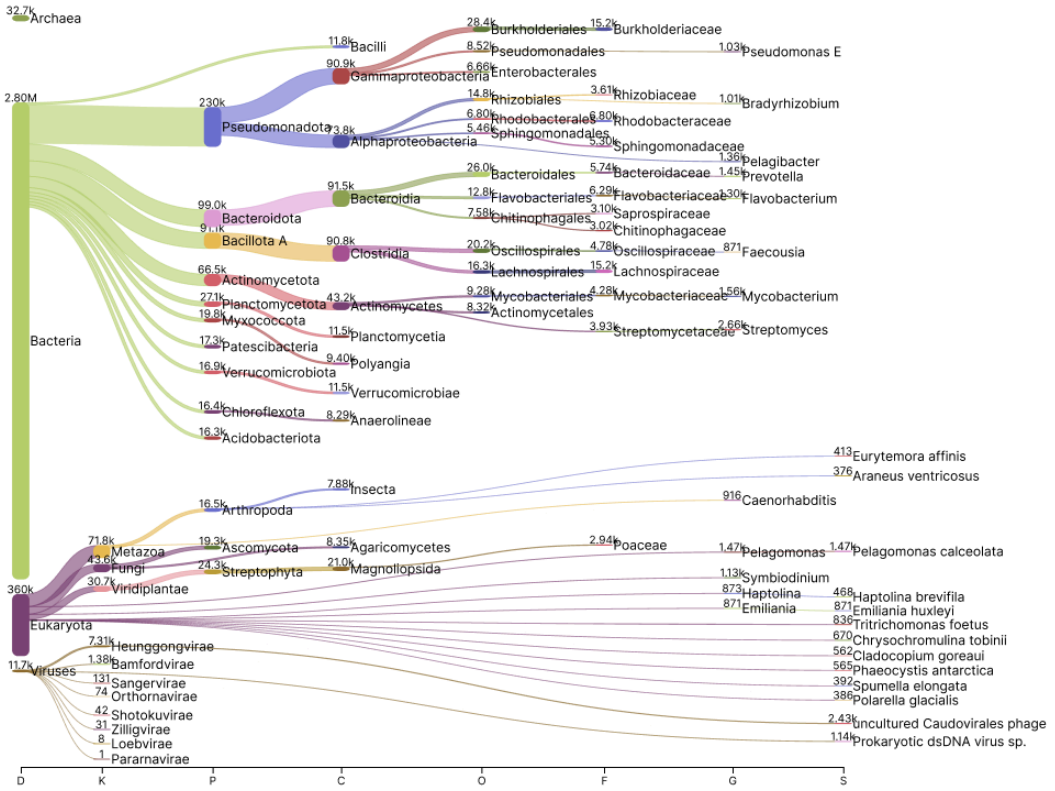
Supplementary Figure 5. The taxonomic distribution of the clusters specific to the biome: Human Skin.



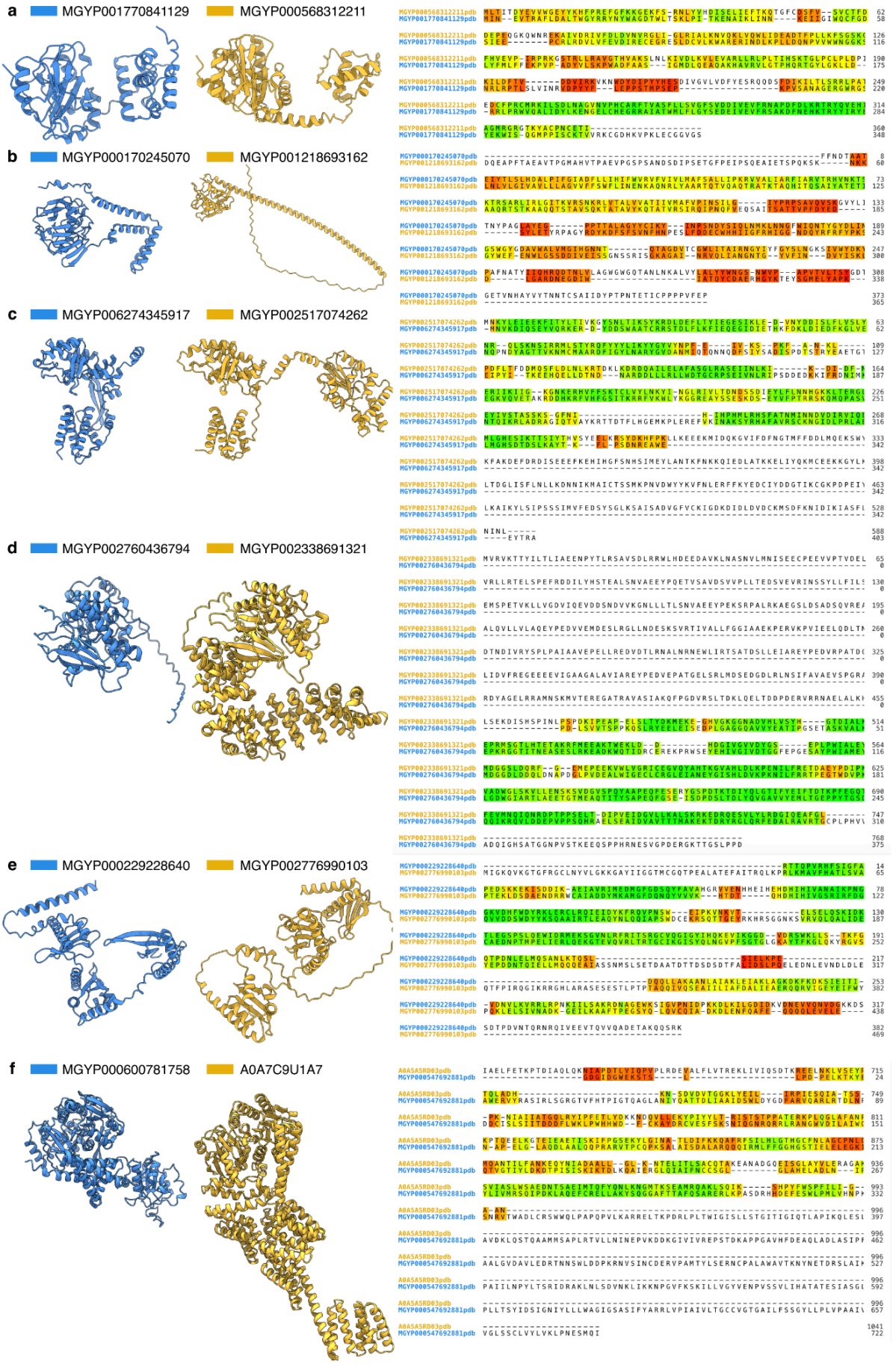
Supplementary Figure 6. The taxonomic distribution of the clusters specific to the biome: Marine.



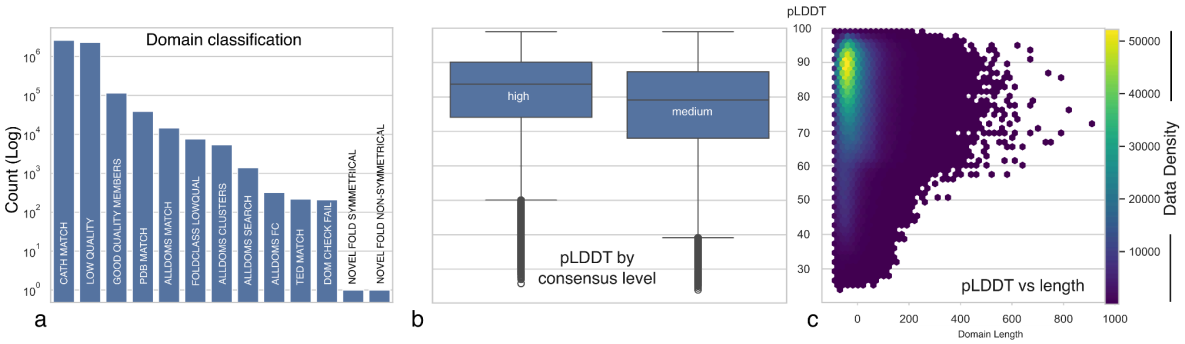
Supplementary Figure 7. The taxonomic distribution of the clusters specific to the biome: Soil.



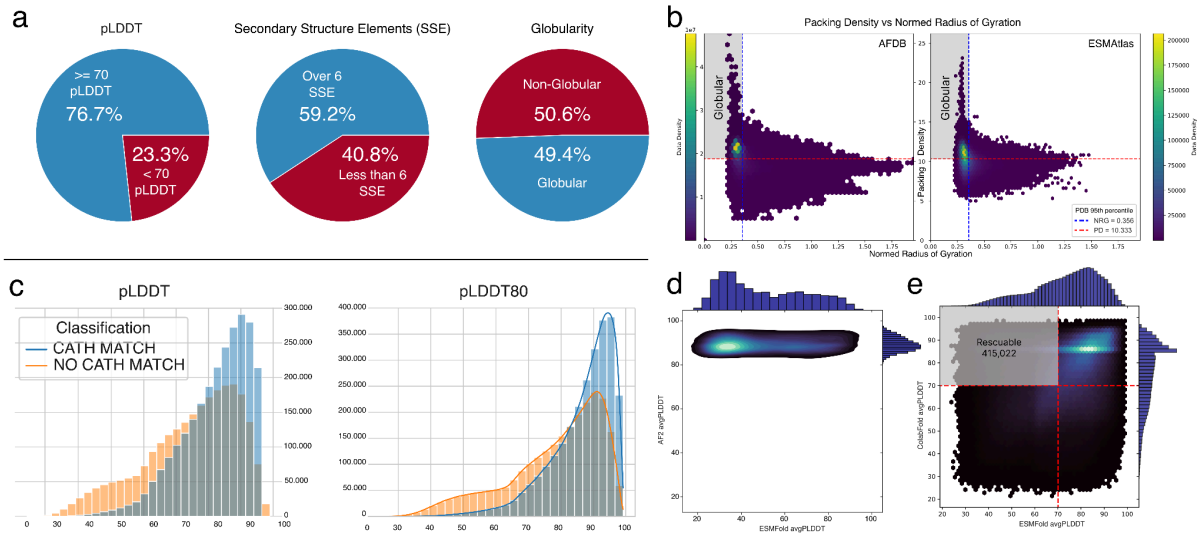
Supplementary Figure 8. The taxonomic distribution of the non-singleton AFESM clusters.


a
MGYP001770841129
MGYP000568312211
b
MGYP000170245070
MGYP001218693162
c
MGYP006274345917
MGYP002517074262
d
MGYP002760436794
MGYP002338691321
e
MGYP000229228640
MGYP002776990103
f
MGYP000600781758
A0A7C9U1A7

Supplementary Figure 9. The superposition of the extreme environment specific clusters to the proteins among the non specific clusters with the lowest E-value. The proteins - one from the environment specific clusters and the other from the non environment specific clusters - are at the left side and the LDDT scores between the proteins computed with Foldseek are at the right side. **a,** MGYP001770841129 (Hyperthermal) is aligned to MGYP000568312211 (Non-specific). **b,** MGYP000170245070 (Hyperthermal) superposed to MGYP001218693162 (Non-specific). **c,** MGYP006274345917 (Hypersaline) is superposed to MGYP002517074262 (Non-specific). **d,** MGYP002760436794 (Hypersaline) superposed to MGYP002338691321 (Non-specific) **e,** MGYP000229228640 (Glacier) superposed to MGYP002776990103 (Non-specific) **f,** MGYP000600781758 (Glacier) superposed to A0A7C9U1A7 (Non-specific).



Supplementary Figure 10. a, Distribution of ESMatlas domains and representatives by category. **b,** Average pLDDT by consensus level (medium - two predictors agree, high - all three predictors agree). **c**, pLDDT as a function of domain length.



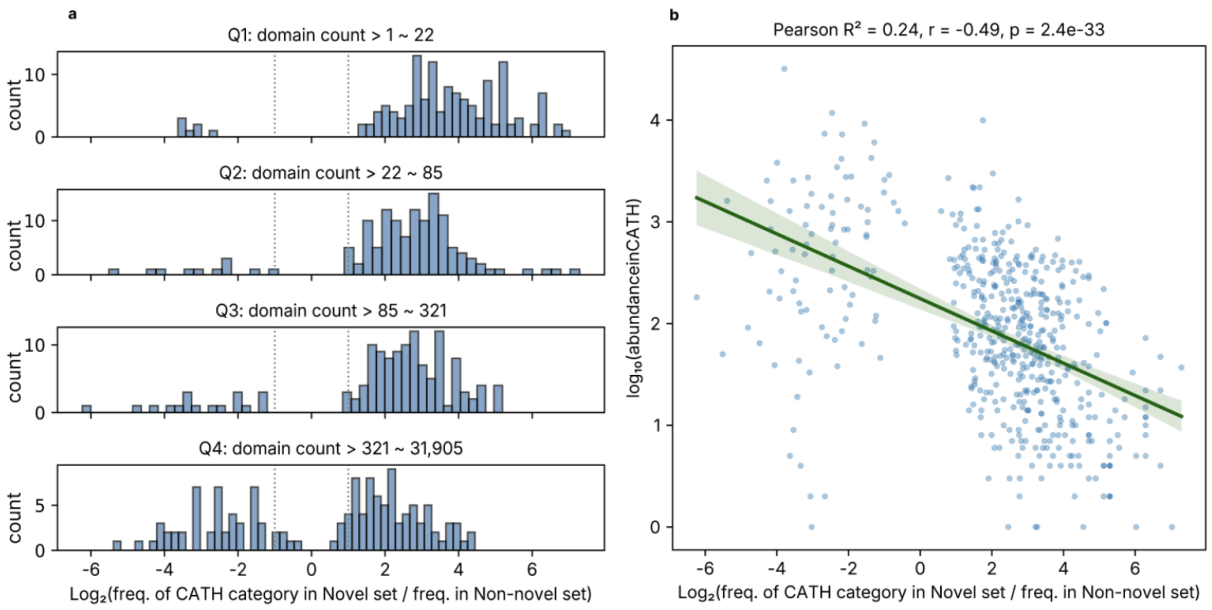
Supplementary Figure 11. a, Quality assessment of domains in ESM-only clusters representatives by pLDDT, number of secondary structure elements and globularity. Domains are deemed of sufficient quality if pLDDT is above 70, they possess 6 or more secondary structure elements and pass both globularity metrics (packing density and normed radius of gyration). **b,** Distribution of globularity (y-axis) and normed radius of gyration (x-axis) for TED (left) and the ESM-only clusters representatives (right) . Domains above the red line and to the left of the blue line (top left quadrant, PDB 95th percentile) are considered globular. **c,** Distribution of pLDDT and $pLDDT_{80}$ by CATH assignment label. Domains with CATH structural matches have higher pLDDT and $pLDDT_{80}$ **d,** Density plot of average pLDDT derived from AF2 versus ESMFold for novel folds in the TED database. **e,** The comparison of pLDDT between the novel domains predicted by ColabFold (y-axis) and ESMFold (x-axis)

Supplementary Table 1. List of 2,906 H-level CATH categories compared between Novel and Non-novel MDPs. For each category, the table includes the CATH identifier, number of occurrences in each set, $\log_2$ (freq. In Novel / freq. in Non-novel), p-value, Benjamini–Hochberg adjusted p-value, differential abundance label (overrepresented, underrepresented, or no significant difference), and presence label (shared, unique to Novel, or unique to Non-novel).

The table is provided as a separate tsv file at: https://afesm.steineggerlab.workers.dev/
File name:
`6-cathID_presenceFlag_nNovel_nNonNovel_statTestMethod_pValue_adjPValue_log2Ratio_ddiffAbundanceFlag_nNovelPartners_cathName.tsv`

Supplementary Table 2. List of 10 most overrepresented CATH H-level categories with significant differential abundance between Novel and Non-novel MDPs, sorted by unique number of novel domain pairing partners. Membrane-associated domains are shown in bold. . For each domain, the table includes the CATH identifier, number of occurrences in both sets, $\log_2$(freq. in Novel set / freq. in Non-novel set), adjusted p-value (Benjamini–Hochberg corrected), and the number of unique novel domain partners.

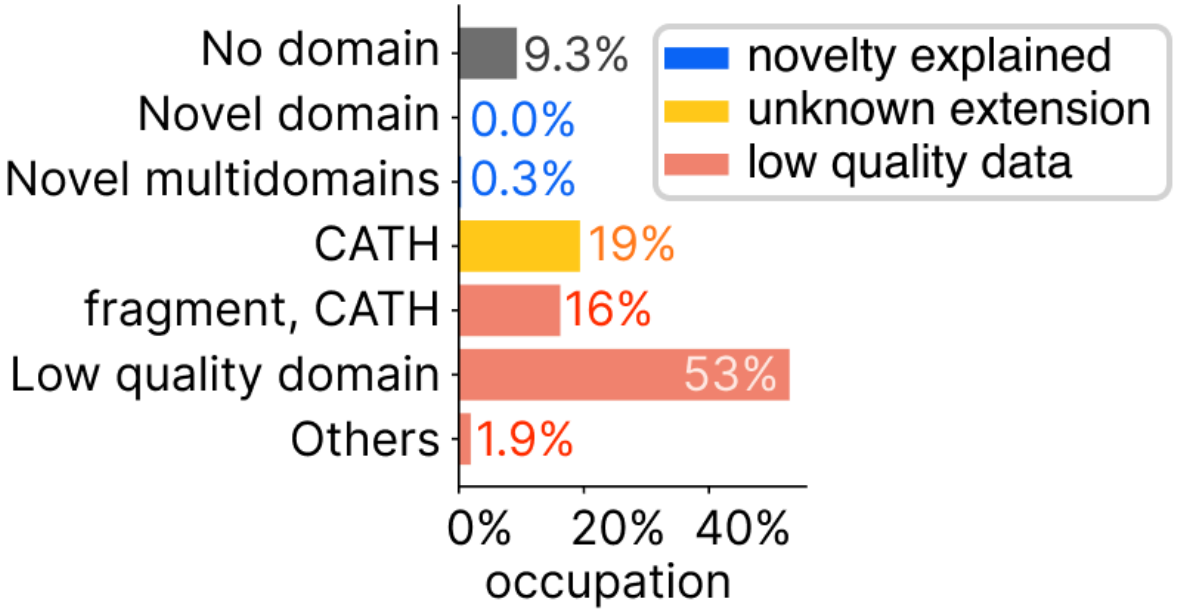
Category	cathname	log2ratio (x)	-log10p_adj (y)	nNovelPair
2.170.130.10	**TonB-dependent receptor, plug domain**	2.2807	60.7733	115
3.65.10.10	Enolpyruvate transferase domain	3.7650	160.8717	103
3.30.2090.10	**Multidrug efflux transporter AcrB TolC docking domain; DN and DC subdomains**	2.9292	75.2153	83
1.10.3720.10	**MetI-like**	2.8088	66.9116	82
3.40.1160.10	Acetylglutamate kinase-like	2.7435	61.8651	82
2.160.10.10	Hexapeptide repeat proteins	1.5259	17.1761	81
2.70.150.10	**Calcium-transporting ATPase, cytoplasmic transduction domain A**	2.6197	58.1307	81
3.40.120.10	Alpha-D-Glucose-1,6-Bisphosphate, subunit A, domain 3	2.4824	50.2820	81
3.90.226.10	2-enoyl-CoA Hydratase; Chain A, domain 1	1.4943	17.5444	79
3.40.710.10	DD-peptidase/beta-lactamase superfamily	1.4637	16.5869	77



Supplementary Figure 12. a, Distribution of $\log_2$(Frequency of CATH H-level categories from Novel set/Frequency of CATH H-level categories from Non-Novel), across quantiles of domain counts in CATH **b**, Negative correlation between $\log_2$ fold change ($\log_2$FC) and $\log_{10}$-transformed abundance of CATH domains. Each dot represents a single protein domain. A linear regression line is shown in green. The Pearson correlation coefficient indicates a moderate negative correlation ($r = -0.49$, $R^2 = 0.24$), which is statistically significant ($p = 2.4 \times 10^{-33}$), suggesting that lower-abundance domains tend to show higher fold changes in expression.

Supplementary Table 3. Top 10 CATH H-level domains by the number of their unique novel pairing partners, ranked by $\log_2$(freq. in Novel set / freq. in Non-novel set).
Each entry includes the domain identifier, enrichment score ($\log_2$ ratio), adjusted p-value, and the number of unique novel domain partners. These domains feature prominently in **Fig. 5e** as part of unexpected or functionally intriguing domain combinations.

Category	cathname	log2ratio (x)	-log10p_adj (y)	nNovelPair	domain_count_CATHDB
3.40.1570.10	HemS/ChuS/ChuX like domains	7.2910	6.0082	5	37
3.50.90.10	YerB-like	7.0280	22.2711	22	1
1.10.700.10	Dioxygenase LigAB, LigA subunit	6.7061	3.3507	3	22
3.30.500.20	BH3703-like domains	6.7061	3.3507	4	4
3.50.70.10	-	6.7061	3.3507	4	47
1.10.3480.10	TorD-like	6.2910	5.3038	5	22
1.10.4010.10	Type II deoxyuridine triphosphatase	6.2910	2.0444	2	8
1.20.1500.10	YheA/YmcA-like	6.2910	2.0444	2	12
2.30.170.40	Ribosomal protein L28/L24	6.2910	2.0444	2	72
3.10.640.10	Restriction endonuclease-like alpha-beta roll domain	6.2910	2.0444	2	5



Supplementary Figure 13. The description of the ESM-only representatives: No domain identified (9.3%), (1) Novelty explained - structurally novel domains (0.0%) or novel multi-domain arrangements (0.3%), (2) Unknown extensions - CATH domains are identified (19%) potentially containing additional regions, (3) Low-quality data — fragments (16%) or predicted domains filtered due to the quality filtering (53%) or other criteria (annotatable domains by iterative clustering and search against PDB or other domain databases, symmetry and quality assessment by Foldclass, 1.9%).