Large protein databases reveal structural complementarity and functional locality

Paweł Szczerbiak[1,2], Lukasz Szydlowski[1,2], Witold Wydmański[2,3], P. Douglas Renfrew[4], Julia Koehler Leman[5], Tomasz Kosciolek[1*]

1. Sano Centre for Computational Medicine, Kraków, 30-054, Poland
2. Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, 30-387, Poland
3. Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, 30-348, Poland
4. Center for Computational Biology, Flatiron Institute, New York, NY 10010, USA
5. Open Molecular Software Foundation, Davis, CA 95616, USA

*Corresponding author: t.kosciolek@sanoscience.org

Abstract

Recent breakthroughs in protein structure prediction have led to an unprecedented surge in high-quality 3D models, highlighting the need for efficient computational solutions to manage and analyze this wealth of structural data. In our work, we comprehensively examine the structural clusters obtained from the AlphaFold Protein Structure Database (AFDB), a high-quality subset of ESMAtlas, and the Microbiome Immunity Project (MIP). We create a single cohesive low-dimensional representation of the resulting protein space. Our results show that, while each database occupies distinct regions within the protein structure space, they collectively exhibit significant overlap in their functional potential. High-level biological functions tend to cluster in particular regions, revealing a shared functional landscape despite the diverse sources of data. To facilitate exploration and improve access to our data, we developed an open-access web server. Our findings lay the groundwork for more in-depth studies concerning protein sequence-structure-function relationships, where various biological questions can be asked about taxonomic assignments, environmental factors, or functional specificity.

Introduction

The tectonic shift that has occurred in structural and molecular biology thanks to the abundance of high-quality protein 3D structure models calls for a new outlook on protein structures as a source of information about biology and a tool to understand it further. Thanks to decades of work on protein sequences in the field, efficient methods to store, search, cluster or annotate sequences exist and are well-documented (1–6). Since the release of AlphaFold2 and accompanying AlphaFold Protein Structure Database (AFDB), the number of readily available protein structures grew 100-fold (7, 8). These data found its way into UniProt and other repositories, where they are accessible through text-based interfaces. Still, just as protein sequence information became ubiquitous and purpose-made methods to leverage this source of biological data became necessary, there is a need for purpose-built tools and a better understanding of the protein structure space. Since the advent of AlphaFold2 (7) and

accompanying database, i.e. AlphaFold Protein Structure Database (AFDB) (8, 9), comprising more than 200 million of models, the focus has shifted towards developing more efficient tools for protein structure processing, e.g. clustering (10, 11), compression (12, 13) or representation (14–16). Subsequently, the ESMAtlas repository, featuring over 600 million predictions, further expanded the available structural models, bringing the total, including those from the Protein Data Bank (17), to nearly 1 billion structures. Much effort has been put so far into a deeper understanding of the AFDB structure space ranging from novel fold search (18, 19), domain identification (20), clustering (21), and structure space exploration (22). But, we still lack a comprehensive and complementary understanding of the protein structure space and how different sources of structure data – obtained using different methods, each utilizing various underlying sequence databases – contribute to this space.

In this study, we analyze three large protein structure datasets originating from distinct sources. The AlphaFold Protein Structure Database (AFDB) (8), based on UniProt (23), includes protein structures from a wide range of organisms, with a significant representation of eukaryotes. The ESMAtlas (24), derived from MGnify (25), focuses exclusively on proteins from metagenomic studies, and therefore predominantly contains prokaryotic data. The Microbiome Immunity Project (MIP) database (26), based on The Genomic Encyclopedia of Bacteria and Archaea (GEBA) (27), primarily consists of short, single-domain structures ranging from 40 to 200 residues, representing bacterial genomes, and may thus have similarities with ESMAtlas. In contrast, both the AFDB and ESMAtlas datasets include single- and multi-domain proteins, distinguishing them from the MIP dataset. The representative sequences and structures are then projected onto a lower-dimensional space according to their structural features and functionally annotated using an updated version of a structure-based function prediction method deepFRI (28).

In this work, we present a comprehensive analysis of the current protein structure space, integrating data from the most extensive sources available. We show that the protein structure databases are complementary and exhaustive in terms of folds and biological function. We examine the unique characteristics of each database, analyze the distribution of protein functions within this space, and introduce a web server that allows for customized exploration of these structural datasets (https://protein-structure-landscape.sano.science/). This work aims to provide a deeper understanding and contextualization of the current protein structure landscape, enabling further insights into protein function and evolution.

Results

Creating a common sequence and structure set

For all of the major protein structure databases we considered subsets containing only non-redundant sequences: AFDB50, highquality_clust30 being high-quality, representative subset of ESMAtlas, and MIP (see Methods for details). Next, we removed structural redundancy applying the procedure described in reference (21). For AFDB we use results already done in reference (21) including both high- and low-quality models. In contrast, the ESMAtlas-based dataset (highquality_clust30) contains only high-quality predictions. Each of the structure databases was generated using orthogonal structure prediction methods and come from different sequence data source – AFDB is based on UniProt, ESMAtlas based on MGnify

metagenomics data, MIP based on representative putative novel domains from across the bacterial tree of life. This diversity allows us to explore different areas of the structural space more comprehensively e.g. localization of low- and high-quality predictions or single- and multi-domain proteins (see also Methods for additional discussion). Nevertheless, all quantitative analyses have been conducted exclusively on high-quality models.

Structure space is continuous and complementary

Our pipeline is shown in Fig. 1A. We start from three sets of non-redundant sequences that, in the first stage, are clustered independently with Foldseek to remove their individual structural redundancy. Subsequently, in the second stage, all representative structures (excluding AFDB and ESMAtlas singletons – see Methods) were gathered into one set and clustered again with Foldseek to remove structural redundancy between databases (see Figs. S1-S18 and Table S1 in the Supplement for details as well as Table S2 for a summary). Next, we annotate the resulting dataset functionally with deepFRI and create Geometricus representations to elucidate their structural features and analyze cluster heterogeneity (14). Fig. 1B illustrates the key features of the structure space created using PaCMAP (29) with all structural representatives (clusters and singletons) as inputs. For construction details, refer to Figs. S19-S25. We then embedded CATH superfamilies (30) into this space to better understand its organization. To increase the utility of our results, we developed an open-access web server that presents all the aforementioned features in an interactive manner (see Methods for details). We used all proteins from AFDB and ESMAtlas subsets (only MIP comprises mostly single-domain structures) so the CATH covers the space only partially. Even so, we can see a clear separation between CATH classes with the largest density in the alpha/beta region. Apart from two small corners in the structure space, we do not see a gradient of the structure length. We also observe an expected high correlation between structure length and mean pLDDT (Fig. 1B the AFDB pLDDT panel; see also Fig. S26 for dispersion measures). Fig. 1C shows a fraction of structures coming from different databases. For a fair comparison with ESMAtlas, we considered only high-quality AFDB models. We can observe very clear complementarity between AFDB, ESMAtlas and MIP clusters when it comes to the overall coverage of the structure space. The separation between structures originating from different databases cannot be explained by structure length distribution alone, since the majority of AFDB and ESMAtlas structures are of comparable size (see Fig. S17). Interestingly, most of the AFDB light and ESMAtlas proteins lie in the same region of the structure space (see Fig. S27 for absolute values and Fig. S28 for version with all AFDB models) meaning that both databases already saturate the known structural landscape to a high extent (see cluster heterogeneity analysis below for a more detailed discussion). We can also see a large overlap between AFDB light and dark clusters.

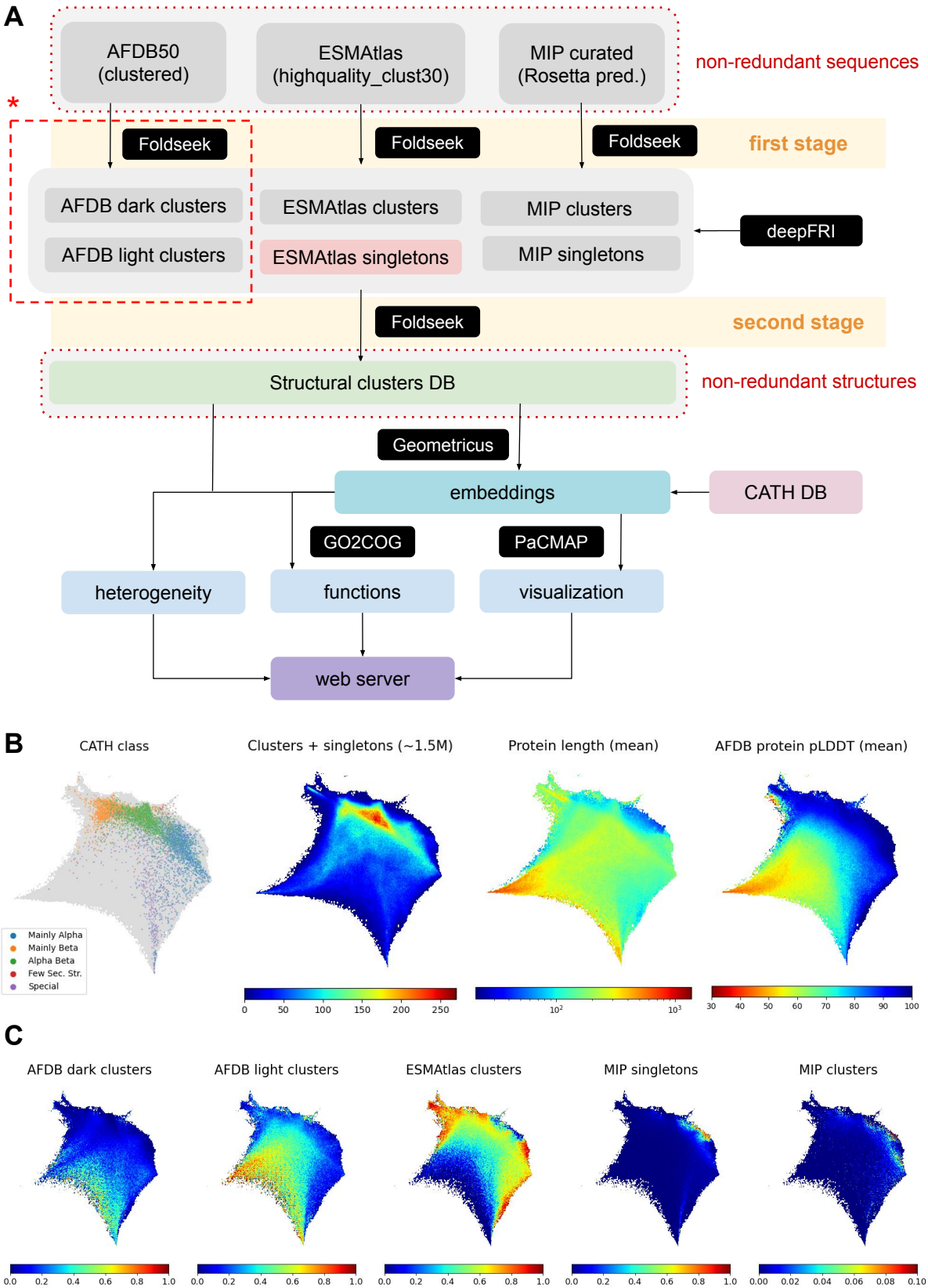

Figure 1: Pipeline and structure space. (A): The diagram illustrates the pipeline from top to bottom. We begin from three sets of proteins that are non-redundant at the sequence level and aim to cluster them at the structure level. Red asterisk indicates the part that has been done in reference (21). Using

a similar methodology, we clustered the remaining two databases and, in the second stage, we clustered the cluster representatives (green box) to eliminate structural redundancy between databases. All structures have been annotated with deepFRI and Geometricus to elucidate its functional potential and visualize the structure space, respectively. Dimensional reduction into 2 dimensions has been performed with PaCMAP. Subsequently, we analyzed cluster heterogeneity. We also created a web server for easy interplay with the final database. **(B):** CATH class (left), number of structures (middle left), protein length (middle right), and pLDDT for AFDB structures (right) distributed across the structure space (we considered all ~1.5 million cluster representatives + singletons). (**C):** Fraction of structures coming from a given database (we included only high-quality models i.e. ~3 million structures - see Table S2). Panels **B** and **C** generated using unnormalized Geometricus representations may be found in the Supplement (see Figs. S29-S32).

All analyzed databases contain similar functional potential

Functional analysis has been performed using deepFRI v1.0 (28) and v1.1 (see Fig. S33). The latter retains the original deepFRI architecture but is trained on AFDB structures. While it generally produces more predictions than version 1.0, it also tends to yield a higher number of false positives (see the Supplement for details). We mapped Gene Ontology (GO) (28, 31) predictions from deepFRI onto Clusters of Orthologous Groups of proteins (COG) (28, 31, 32), and superCOG categories (see Methods for details on how the GO to COG mapping was created). In order to establish the best thresholds on deepFRI score, we validated both versions on *E. coli* proteome – see Figs. S34-S35. When the functionally annotated proteins from six datasets (i.e. the ones from Fig. 1C and MIP new folds as a subset of the entire MIP dataset) are aggregated into superCOG categories, we can notice similar distributions of superCOGs between them with the exception of smaller sets, such as MIP new folds (Fig. 2A). This dataset, representing previously unseen folds (before the advent of AFDB), does not represent unknown functions, as most of the proteins may be annotated (similarly to AFDB dark clusters, where 67-80% of proteins are annotated). In these datasets, functions associated with superCOG 3 (information storage and processing) are less abundant than in others. These functions, related to replication and gene expression processes, are generally conserved throughout living organisms and are relatively well studied (33, 34). Compared to deepFRI v1.0, deepFRI v1.1 has 100% annotation coverage of all proteins. However, it is expected that deepFRI v1.1, being trained on the Uniprot dataset, will predict functions associated with superCOG 3 with less accuracy than deepFRI v1.0, which was trained on Swiss-Prot for the aforementioned reasons. Conversely, superCOG 2 and general function categories agree quite well between v1.0 and v1.1, We can notice complementarity in superCOG 1+3 annotation in the regions without superCOG 1+2 (Fig. 2B). i.e. the lower corner of the plot contains proteins annotated as superCOG 1+3 and not 1+2. Regions not annotated by v1.0 are mostly predicted by v1.1 as superCOG 1+2 (see Fig. S36). Regions representing the general function category lie in the same corner (consisting of mostly low-quality predictions or disordered proteins – compare Fig. 1B), suggesting that generally this cluster is hard to predict, regardless of the annotation tool. Functional overlay with CATH classes reveals that alpha/beta CATH class overlaps mainly with metabolic functions (superCOG 2). This agrees with the higher complexity of proteins that function as enzymes (e.g. substrate-determined structure) (35, 36). Normalized contribution of all functional categories is not randomly distributed according to Moran's I measure (p-value close to zero – see Fig. 2B). The largest COG categories in each dataset are shown in Fig. S43.

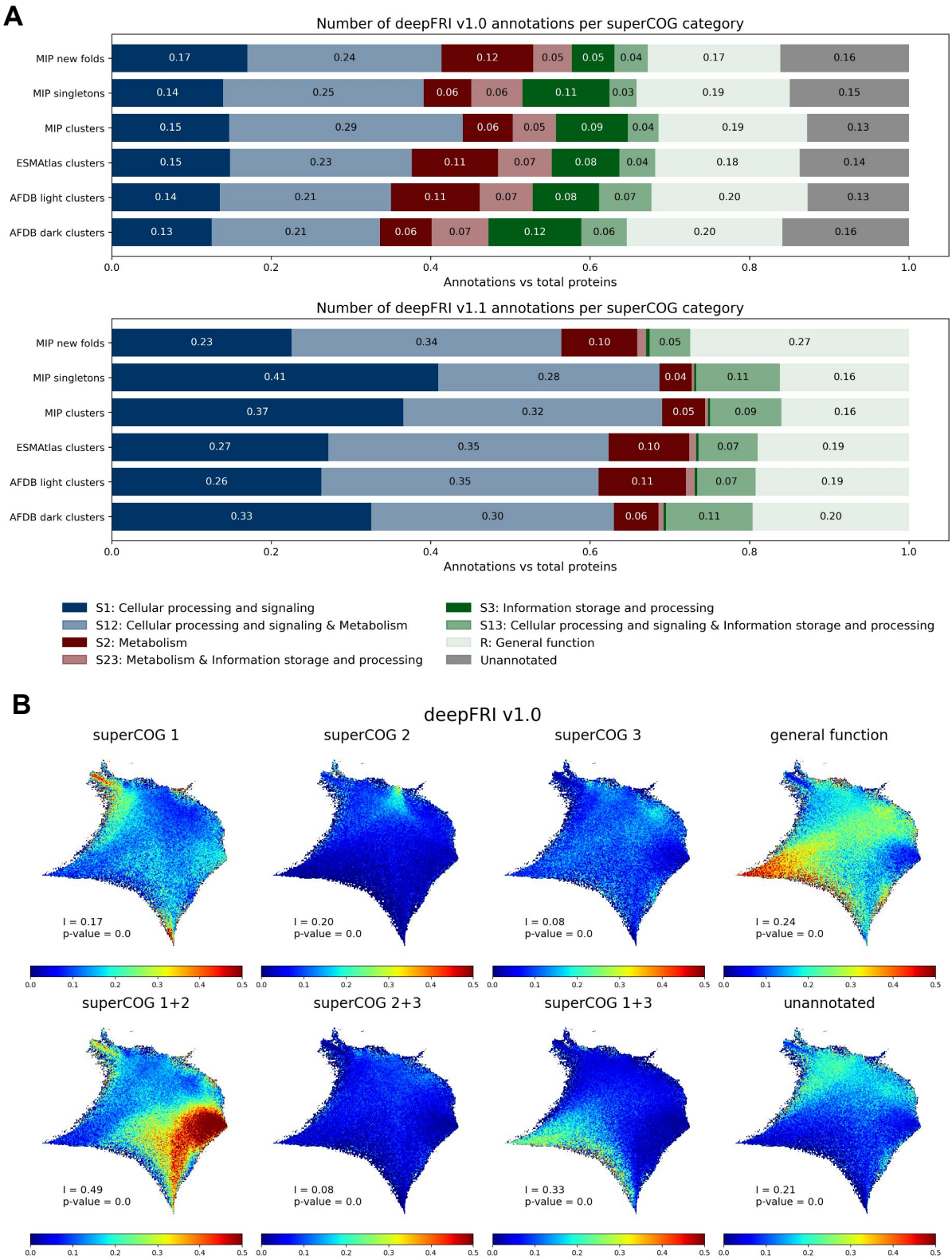

Figure 2: Distribution of functional categories within datasets and the structure space. (A): deepFRI annotation of high-quality structural models in each database based on superCOG aggregation. For deepFRI v1.0 and v1.1, annotations with score ≥ 0.3 and ≥ 0.5 were considered, respectively. **(B):** Normalized contribution of each superCOG category for deepFRI v1.0 (see Fig. S36 for deepFRI V1.1 predictions). All panels have been generated using normalized Geometricus vectors for all ~1.5 million cluster representatives + singletons (versions showing absolute values as well as based on unnormalized Gemetricus representations can be found in the Supplement – Figs. S37-S38

and Figs. S39-S42 respectively). Moran's I coefficient and its corresponding p-value are displayed in each panel.

Novel folds do not have to represent unknown functions

We found a few examples of proteins annotated with high information content (i.e. high information entropy within the GO database; compare (37)) in the MIP new folds clusters (Fig. 3). These include tetrahydromethanopterin S-methyltransferase activity (GO:0030269), intramembrane lipid transporter activity (GO:0140303) and pectinesterase inhibitor activity (GO:0046910). Tetrahydromethanopterin S-methyltransferases form catalytic complexes with cobalamin (vitamin B12) and transfer methyl groups from tetrahydromethanopterin (H4MPT) to coenzyme M. The *mtrA* gene, which encodes the MtrA subunit of the tetrahydromethanopterin S-methyltransferase complex, shows significant phylogenetic diversity (38, 39). Ability to bind complex substrates, such as H4MPT and form complexes with other proteins certainly drives structural diversity (38–41). The protein of interest was derived from a genome of *Blautia* sp., a known methanogen, hence increasing our confidence in this prediction. Foldseek search revealed the highest similarity to the another *Blautia* protein in AFDB database (TM-score = 0.99, Fig. 3A), whereas alignment with other known tetrahydromethanopterin S-methyltransferases resulted in poor alignments (TM score = 0.47). Another example (Fig. 3B) concerns intramembrane lipid transporters which are a diverse group of proteins that move lipids within the plane of the membrane. The structural diversity of these proteins is highlighted by their varying number of transmembrane domains, which can range from a few to over a dozen. They may exhibit different conformations, and mechanisms of action (e.g. flippases, floppases, scramblases) (42, 43). The structural diversity is also determined by the unique, lineage-specific lipid profile of cell membranes. MIP_00055961 sequence has been derived from the genome of *Ectobacillus panaciterrae* DSM 19096, which represents a recently distinguished taxonomic group, therefore its cellular fatty acid profile is indeed different from other *Bacillales*. The closest structural match (TM-score > 0.8, despite only 20% sequence similarity) is to RRXRR domain-containing protein from cyanobacterial organism *Symploca* sp. The third example (Fig. 3C) is related to pectinesterase inhibitors (PMEIs) being extracellular enzymes that act by binding to and inhibiting pectin methylesterases (PMEs). Thus they can regulate the activity of the latter and hence the cell wall modification. PMEIs can act against endogeneous PMEs or exhibit a physiological action toward enzymes derived from different species (i.e. in symbiotic or host-pathogen relationships). The aforementioned factors can greatly influence structural variability (44–46). MIP_00248745 is derived from *Gordonia bronchialis* DSM 43247, which is a known opportunistic human pathogen. Its function as PMEI could therefore explain the opportunistic pathogenicity, as PMEI-like protein could affect glycoprotein formation in human lung mucus. MIP_00248745 shows high similarity to protein from *Gordonia insulae* (TM-score = 0.69, with 48% sequence similarity), but also *Nocardia brasiliensis* and *Mycobacterium tuberculosis* to which *G. bronchalis* is related.

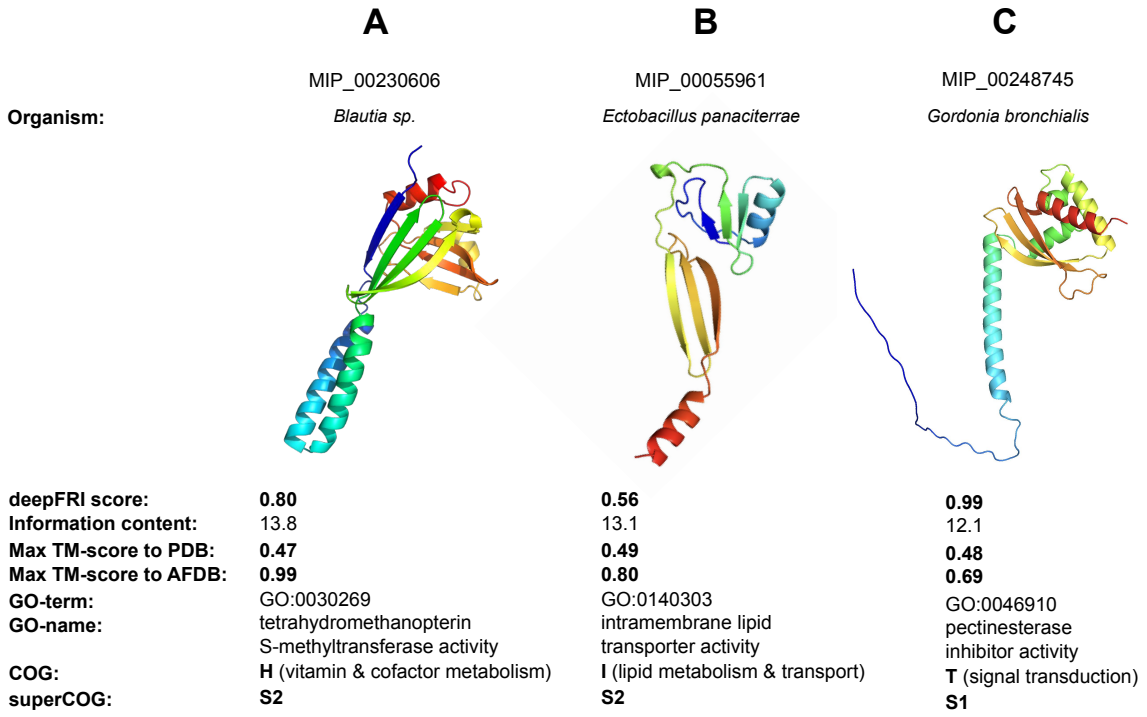

Figure 3: MIP novel fold case study. Functional overview of MIP novel folds. deepFRI v1.1 has been used for function annotations.

Heterogeneous structural clusters constitute most of the protein space

Fig. 4 summarizes different aspects of cluster heterogeneity. In panel A we present a total number of structures in considered databases split by group i.e. singletons, homogeneous clusters, and heterogeneous clusters. Firstly, large parts of AFDB clusters are of low or moderate quality (pLDDT ≤ 70) – for AFDB light and dark this is 38%, and 53% respectively (red bars). This can be explained by the significant structural redundancy in the AFDB, which was eliminated with Foldseek, leaving us with specific high-quality protein building blocks and a large number of either poorly predicted or intrinsically disordered proteins (see also Fig. S18). Consequently, the latter overshadowed the vast number of excellent AlphaFold models in the full AFDB. Interestingly, incorporation of a different procedure for the AFDB (we have done filtering after clustering instead of clustering after filtering like with ESMAtlas) resulted in more-or-less similar proportion of singletons / homogeneous and heterogeneous clusters for both AFDB and ESMAtlas. Also, most of AFDB and ESMAtlas structures mix with each other building heterogeneous clusters (Fig. 4A). Among all high-quality AFDB structures that build heterogeneous clusters with ESMAtlas (or ESMAtlas and MIP), only 13% (or 11%) are annotated as dark. On the other hand, the percentage of dark cluster representatives in all clustered AFDB structures is 31% (and drops to 26% when considering only models with mean pLDDT > 70). Those observations support in a more quantitative way the conclusions that we may observe a higher overlap between light AFDB and ESMAtlas, drawn from structure space analysis in Fig. 1C. In Fig. 4B one can notice that for heterogeneous clusters the percentage of superCOG 3 for representative structures decreases with cluster size and conversely increases for superCOG 1 and 2. Such higher diversity can be supported by the fact that proteins from superCOG 1 and 2 category exhibit fewer constraints than superCOG 3-associated proteins, whose structures are determined largely by nucleic acid and ribosome

conformation. Conversely, an increase in superCOG 3-related proteins in homogeneous ESMAtlas data (Fig. S46) can be attributed to the presence of phage/viral proteins, derived from metagenomic samples. Likewise, the percentage of unannotated proteins decreases with cluster size, suggesting that the largest clusters are better analyzed functionally. Percentage of general functions change differently for AFDB + ESMAtlas and for AFDB + ESMAtlas + MIP clusters but this can partially interfere with the behavior of mixed superCOG categories (not shown in the plot) that may also express more general annotations. In Fig. 4B we compared only cluster centroids. Therefore, we checked if the conclusions above hold also when limiting ourselves to functionally homogeneous clusters only – indeed, the trends are similar (see Fig. S44). We can observe analogous behavior for homogeneous clusters (see Figs. S45-S46) but e.g. superCOG 2 changes less rapidly. The above observations may indicate that all of these databases are big enough to express similar functional potential at a very general level but even at that level one can see a large diversity across specific clusters. In Fig. 4C, for AFDB homogeneous clusters one can notice an increasing trend between cluster size and representative structure length. This is not observed for other databases but similar correlation (albeit much weaker) holds for heterogeneous clusters comprising the AFDB structures (see Figs. S47-S50 for details). Fig. 4D shows that the percentage of singletons rapidly decreases with structure size irrespective of the database. For AFDB, in the tail of long proteins (> 1300 aa) we can also notice a sizable percentage of structures that do not cluster with anything else (this is probably statistically biased i.e. almost no such long structures in ESMAtlas and few in AFDB itself). Fig. 4E depicts a number of structures in clusters built from long proteins. Again, they come almost exclusively from the AFDB. Most of the clusters are small and only 31% have mean pLDDT of their representative structures higher than 70.

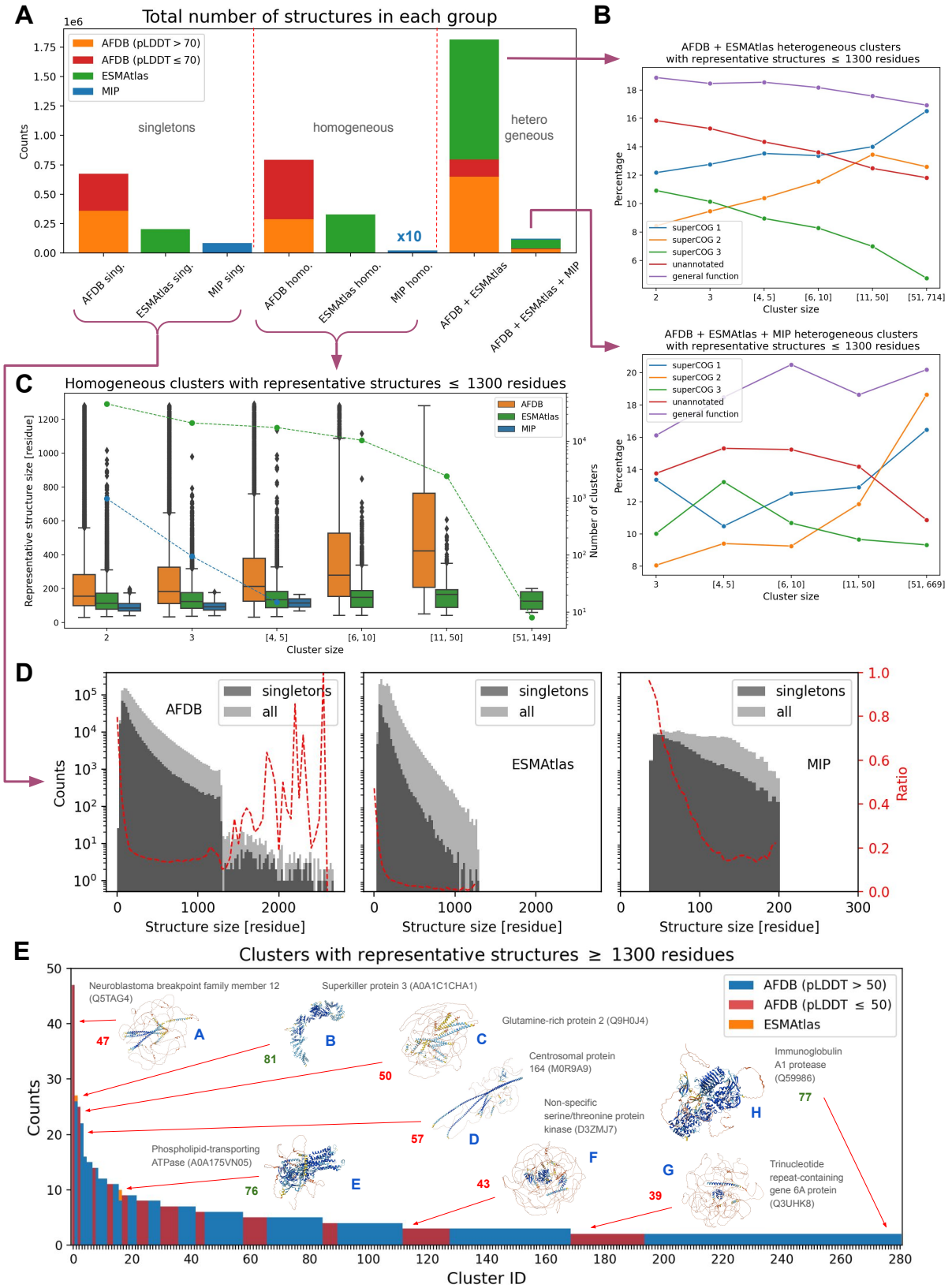

Figure 4: Cluster heterogeneity. (A): Number of structures (clusters + singletons) stratified by database type and cluster group (singletons, homogeneous and heterogeneous clusters respectively). AFDB structures have been split into low- and high-quality groups. **(B):** SuperCOG category of representative structures vs cluster size for the two largest heterogeneous cluster groups i.e. AFDB + ESMAtlas and AFDB + ESMAtlas + MIP (corresponding plots for homogeneous clusters as well as for

all structures building highly functionally homogeneous clusters may be found in the Supplement – Figs. S44-S46). **(C):** Dependance between representative structure length and cluster size for homogeneous clusters (corresponding plots for heterogeneous clusters may be found in the Supplement – Figs. S47-S50). Number of clusters with a given cluster size is plotted as lines on the right vertical axis. **(D):** Number of singletons for each database type versus structure length. Ratio between the number of singletons and the total number of structures is depicted in red (see right vertical axis). **(E):** Distribution of clusters (mostly coming from the AFDB with two exceptions i.e. ESMAtlas structures depicted in orange) with long representative structures (≥ 1300 residues). AFDB structures with mean pLDDT ≤ 50 are depicted in red (see albo numbers attached to each example structure). In **B – D** only high-quality structures have been considered. In **B** and **C** only structure representatives shorter than 1300 residues have been included.

Structural landscape is characterized by high consistency

Fig. 5 presents the landscape of our structure space. We chose several representative structures in different regions with mostly high mean pLDDT (with few exceptions i.e. A0A498NE77 having mean pLDDT equal 69 and five proteins from Fig. 4E). One can notice high consistency and incremental, smooth changes when moving across space (of course, this is not always the rule and some outliers may be found from time to time). Especially interesting region is the one located in the top left, consisting of mostly tube-like proteins with different cross-sections. We can distinguish between square (MGYP000417476362, A0A3G3K1M9), flat (A0A0F6LHQ3, A0A498NE77), irregular (MGYP001600570585, A0A7T9AEB2), and even triangle (P16466) cross-sections that could be responsible for many important functions. Structures clustered in this region, mostly involved in pore-forming activity and transmembrane transport are mostly interesting from the medical perspective, as they are associated with pathogenicity and virulence, e.g. hemolysin (P16466). Other tube-like proteins may also be involved in translation (A0A3G3K1M9) signaling, defense response (A0A0F6LHQ3), interaction with macromolecules (A0A498NE77), kinase activity in response to stimulus (A0A7T9AEB2). Location of structures that build heterogeneous clusters is shown in Figs. S51-S52 – it correlates with the regions with the highest overlap between AFDB light, ESMAtlas and MIP structures (Fig. 1C). Additionally, in Fig. S53 the largest homogeneous and heterogeneous cluster representatives are shown. Similarly to Fig. 5 we can observe a high consistency across the structure space between different databases, although AFDB shows highest divergence among the homogeneous groups. Both heterogeneous groups are located along the top edge, but the heterogeneous AFDB + ESMAtlas group is more spread, including ESMAtlas regions of occurrence. The most numerous structures, as shown in blue and magenta A-F groups, resemble transmembrane or intramembrane proteins, which follows the ubiquity of the structures shared between the databases, as well as the divergence of cellular membranes between different living organisms. In Fig. S54, a quantitative comparison of functions within homogeneous and heterogeneous clusters is provided. SuperCOG 1+2 category is the most ubiquitous (>23%) in the AFDB + ESMAtlas homogeneous clusters; it is also the most numerous in ESMAtlas homogeneous cluster (>23%), whereas in AFDB homogeneous cluster it is second after general function group (>20% and >22%, respectively). For AFDB and ESMAtlas singletons groups, general function is the most ubiquitous, followed by SuperCOG 1+2, which is reversed in MIP singletons. Fig. S54 is accessible in the interactive version on the paper's GitHub repository (see Data Availability section below).

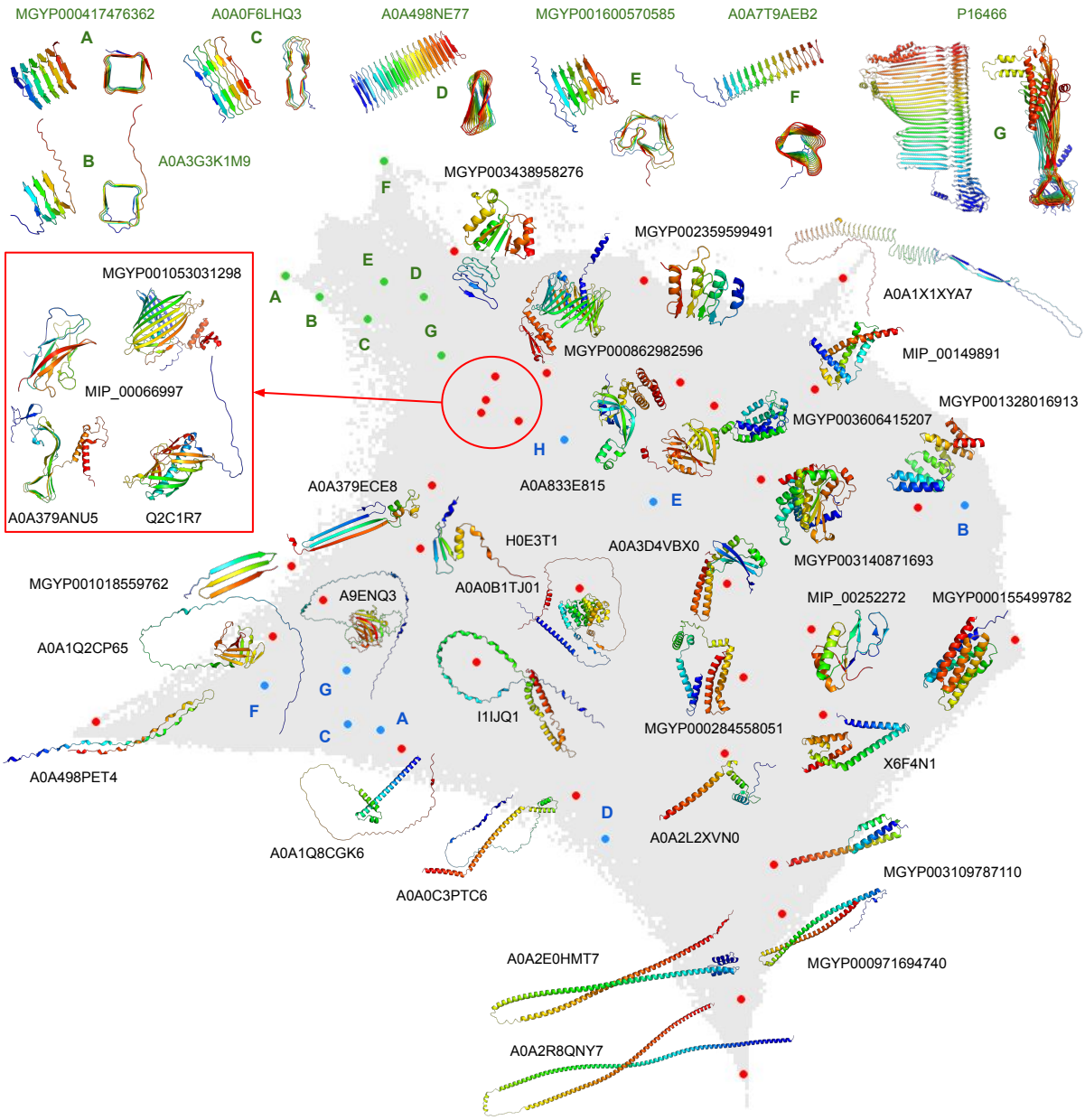

Figure 5: Structural landscape. Example structures of the protein structure space – red and green dots. The latter correspond to tube-like proteins with different kinds of cross-section (we chose examples with high predicted aligned error only to be more certain of their overall geometry). Blue dots depict examples from Fig. 4E. For easier navigation, blue and green dots are labeled with letters.

Discussion

It has been already demonstrated that the protein structure space is continuous and that the Protein Data Bank contains sufficient structural information to understand the protein structure landscape (17, 47–50). This statement is supported by the fact that AlphaFold, trained exclusively on PDB structures, succeeded in predicting structures for the vast majority of the UniProt proteins. In this work, we show that when considering clusters of structures derived from gene-coding proteins in the largest available protein structure databases, the structure space still appears to be continuous (Fig. 1). However, we also observe clear complementarity between these databases: AFDB and ESMAtlas clusters, although greatly overlap, cover

different corners of the structure space on a similar level (Fig. 1C). MIP domains complement ESMAtlas cluster representatives but we might speculate if inclusion of ESMAtlas and AFDB singletons would change the picture. We opted against including singletons from AFDB and ESMAtlas following the work of Barrio-Hernandez et al. (21). Literature reports also highlight the importance of singleton (or individual-specific) genes (51, 52). Discerning putative true or false positive genes lies beyond the scope of this work and represents an exciting direction for future research. Another possible avenue for future work is to split all input structures into domains, evolutionary conserved structural building blocks, before clustering. This approach would provide a different perspective on the protein structure space, highlighting similarities between AFDB and ESMAtlas at the domain level (18, 20).

Functional analysis revealed that all considered datasets possess similar high-level functional potential (on superCOG level – Fig. 2A). Nevertheless, we can observe certain differences on a cluster level (see Fig. 4B). It appears that novel folds do not necessarily indicate unknown functions. To accurately determine their functional capabilities, it is essential to utilize advanced annotation tools that do not rely solely on homology. deepFRI offers a great advantage to homology-independent functional annotation, which was demonstrated in a number of papers (22, 37, 53–55). Furthermore, the differences between deepFRI v1.0 and v1.1 create a complementary advantage in analyzing large databases. While version 1.0 appears more reliable and produces more balanced predictions, version 1.1 proves particularly valuable for studying novel folds (Fig. 3).

Heterogeneity analysis shows that most of the AFDB and ESMAtlas structures build heterogeneous clusters (with at least one representative from each database), meaning that there is a large structural similarity between them. Nonetheless, we can still identify a substantial number of AFDB and ESMAtlas singletons and homogeneous clusters – their nature (including taxonomic assignment) is yet to be fully explored.

By design, we included all AFDB clusters created in reference (21). This approach contrasts with the highquality_clust30 definition but enables us to estimate where low-pLDDT AlphaFold predictions – either low-quality models or intrinsically disordered proteins (IDP) – may reside in the structure space. As expected, deepFRI primarily annotates these as having a general function. However, considering that many of them are likely disordered, this highlights the need for improved functional annotation of intrinsically disordered proteins – IDPs usually fall into the superCOG 1 group, given their function as molecular chaperones (56, 57). Given that deepFRI is a structure-based tool, the annotation of IDPs would require a development of novel approach, e.g. based on dynamic modeling of IDPs under specific physiological conditions.

Methods

Datasets

We considered three large protein structure datasets: AFDB50 (21) i.e. AlphaFold-database clustered at 50% sequence identity level, highquality_clust30 i.e. high quality subset of ESMAtlas (see details here https://github.com/mosamdabhi/esm_meta_protein/blob/main/scripts/atlas/README.md)

with pTM and pLDDT > 0.7 clustered at 30% sequence similarity level, and Microbiome Immunity Project (MIP) (which already consists of structures clustered at 30% sequence identity level) (26). They comprise predictions from AlphaFold (7, 26), ESMFold (24), and Rosetta (24, 58) respectively. By default, we did not filter the AFDB structures by pLDDT as we wanted to observe the distribution of structures with low pLDDT within the structure space (note that some of them may be intrinsically disordered proteins). However, quantitative conclusions are drawn for AFDB structures with mean pLDDT > 0.7 (to be in line with highquality_clust30 construction) i.e. Fig. 1C, 2A, and Fig. 4. We also provide additional plots in the Supplement where the aforementioned constraint is imposed (e.g. the whole heterogeneity analysis section). Similarly to reference (21) we exclude highquality_clust30 singletons. We did include MIP singletons since they may contain important but rare domains from the tree of life coming from a curated set of genomes. See the Supplement for more details.

Structural clustering

Structural clustering has been performed using Foldseek (11) version e99047b2214b20b7c93f36fc0d54394c843c046b. We followed a similar procedure as the one described in reference (21). Structure comparison that outputs TM-score has been done using US-align (21, 59) version 20230609. Parameter tuning and more details may be found in the Supplement.

Function prediction

Function prediction has been done using deepFRI (28) with release v1.0 (https://github.com/flatironinstitute/DeepFRI) and release v1.1 (https://github.com/bioinf-mcb/DeepFRI). deepFRI v1.1 has been trained using exactly the same architecture and training procedure as v1.0 but with a different training and validation sets (see the Supplement for details).

go2cog mapping

Mapping of Gene Ontology (GO) terms (28, 31) into Clusters of Orthologous Groups of proteins (COG) (28, 31, 32) has been performed manually with the use of 1,227 GO-terms, 26 COGs and 3 superCOGs. Thanks to GO-term propagation with GO graph (`go.obo` file available at https://geneontology.org/docs/download-ontology/) we were able to map 4,504 GO-terms for deepFRI v1.0 (95% coverage of all outputted GO-terms), 8,466 GO-terms for deepFRI v1.1 (91% coverage of all outputted GO-terms), and 9,427 GO-terms in total. The mapping is freely available at https://github.com/Tomasz-Lab/go2cog. deepFRI predictions have been mapped to superCOG categories using the following scheme: 1) Use only high-quality deepFRI predictions (threshold put on deepFRI score equals 0.3 and 0.5 for v1.0, and v1.1 respectively – see the Supplement for details). 2) Group predictions by protein. 3) If only R (general function) annotations are present, annotate a given protein as R, but If more than one superCOG category appears, assign them all equally, ignoring R (if present). 4) If all 3 superCOG categories are present, assign them to the R category. For COG categories, deepFRI predictions have been mapped using the same scheme but in point 4) we relaxed the number of different categories from three to at least five.

Structure space construction

Structure embeddings have been generated using Geometricus (14, 60) v0.5.0 with the `ShapemerLearn` class that outputs 1024-dimensional shape-mer representation (see https://turtletools.github.io/geometricus/getting_started.html for details). We computed shape-mers for all representative structures in the final clustered database comprising 1,505,141 proteins (see Fig. 1 and Table S2). The Geometricus representation has been reduced into two dimensions using the Pairwise Controlled Manifold Approximation Projection (PaCMAP) (29) being the current state-of-the-art technique for this purpose. We performed an extensive grid search to tune PaCMAP parameters (see the Supplement for details). By default, we normalize Geometricus representations (to reduce structure length bias), but we also provide results for unnormalized vectors (see Figs. S36-S43). 3D protein structure visualizations have been generated with PyMOL 2.5.0 Open-Source, 2022-11-16.

Web server

The web server enables users to interactively explore the protein structure landscape. Features include filtering by database, superCOG, protein length, and AFDB pLDDT, which may be particularly useful for identifying disordered proteins. The frontend has been created using React (https://react.dev/), while backend is based on Python's FastAPI library (https://fastapi.tiangolo.com/). We have visualized the 2D projections of the embeddings using SciChart (https://www.scichart.com/). The backbone of the visualization consists of 10,000 points randomly sampled from the whole database. Each change in filters or screen location prompts a request to the backend, which responds with 1000 points matching the criteria and displays them on the frontend. The code for both parts is available at https://github.com/wwydmanski/2dPointVis. The web server is accessible at: https://protein-structure-landscape.sano.science/.

Data availability

Lists of clusters and singletons obtained from Foldseek, along with all relevant metadata (including structure space coordinates, functional annotations, and number of structures coming from each dataset) as well as Fig. S54 as an HTML file, can be found in the paper's GitHub repository: https://github.com/Tomasz-Lab/protein-structure-landscape.

Supplementary data

The supplement is available as a separate PDF file.

Author contributions

P.S., T.K., P.D.R. and J.K.L. designed the research. P.S. Ł.S. and T.K. performed analysis. Ł.S. created the go2cog mapping. P.S. and W.W. retrained deepFRI v.1.1. W.W. developed the web server. P.S. T.K. Ł.S. J.K.L. and P.D.R. wrote the manuscript.

Acknowledgements

The computations reported in this paper were (in part) performed using resources made available by the Flatiron Institute. The Flatiron Institute is a division of the Simons Foundation.

We kindly acknowledge the support of World Community Grid team and the community of volunteers who donated their computational resources to the creation of the MIP database. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2022/015689. Special thanks also go to the high-performance cluster at the Małopolska Centre of Biotechnology of Jagiellonian University, and to its dedicated system administrator, Wojciech Pilch, for his invaluable support in our computational endeavors.

Funding

This work was supported by the National Science Centre, Poland grant 2023/05/Y/NZ2/00080 to T.K. and P.S..

Conflict of interest statement

None declared.

References

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**.
2. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L., *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211.
3. Hauser,M., Steinegger,M. and Söding,J. (2016) MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, **32**, 1323–1330.
4. The UniProt Consortium, Bateman,A., Martin,M.J., O'Donovan,C., Magrane,M., Alpi,E., Antunes,R., Bely,B., Bingley,M., Bonilla,C., *et al.* (2016) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
5. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**.
6. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J. and Söding,J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 1–15.
7. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
8. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
9. Varadi,M., Bertoni,D., Magana,P., Paramval,U., Pidruchna,I., Radhakrishnan,M., Tsenkov,M., Nair,S., Mirdita,M., Yeo,J., *et al.* (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.*, **52**, D368–D375.
10. Elofsson,A. (2023) Progress at protein structure prediction, as seen in CASP15. *Curr. Opin. Struct. Biol.*, **80**, 102594.
11. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2023) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.

12. Kim,H., Mirdita,M. and Steinegger,M. (2023) Foldcomp: a library and format for compressing and indexing large protein structure sets. *Bioinformatics*, **39**, btad153.
13. Deorowicz,S. and Gudyś,A. (2023) Efficient protein structure archiving using ProteStAr. *bioRxiv*, 10.1101/2023.07.20.549913.
14. Durairaj,J., Akdel,M., de Ridder,D. and van Dijk,A.D.J. (2020) Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, **36**, i718–i725.
15. Wang,D., Abbas,U.L., Shao,Q., Chen,J. and Xu,D. (2023) S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure. *bioRxiv*, 10.1101/2023.08.06.552203.
16. Derry,A. and Altman,R.B. (2023) COLLAPSE: A representation learning framework for identification and characterization of protein structural sites. *Protein Sci.*, **32**, e4541.
17. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235.
18. Bordin,N., Sillitoe,I., Nallapareddy,V., Rauer,C., Lam,S.D., Waman,V.P., Sen,N., Heinzinger,M., Littmann,M., Kim,S., *et al.* (2023) AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology*, **6**, 1–12.
19. Waman,V.P., Bordin,N., Alcraft,R., Vickerstaff,R., Rauer,C., Chan,Q., Sillitoe,I., Yamamori,H. and Orengo,C. (2024) CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. *J. Mol. Biol.*
20. Lau,A.M., Bordin,N., Kandathil,S.M., Sillitoe,I., Waman,V.P., Wells,J., Orengo,C.A. and Jones,D.T. (2024) Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *bioRxiv*, 10.1101/2024.03.18.585509.
21. Barrio-Hernandez,I., Yeo,J., Jänes,J., Mirdita,M., Gilchrist,C.L.M., Wein,T., Varadi,M., Velankar,S., Beltrao,P. and Steinegger,M. (2023) Clustering predicted structures at the scale of the known protein universe. *Nature*, **622**, 637–645.
22. Durairaj,J., Waterhouse,A.M., Mets,T., Brodiazhenko,T., Abdullah,M., Studer,G., Tauriello,G., Akdel,M., Andreeva,A., Bateman,A., *et al.* (2023) Uncovering new families and folds in the natural protein universe. *Nature*, **622**, 646–653.
23. The UniProt Consortium, Bateman,A., Martin,M.-J., Orchard,S., Magrane,M., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., Bye-A-Jee,H., *et al.* (2022) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
24. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 10.1126/science.ade2574.
25. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale,V., Potter,S.C., Richardson,L.J., *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**.
26. Koehler Leman,J., Szczerbiak,P., Renfrew,P.D., Gligorijevic,V., Berenberg,D., Vatanen,T., Taylor,B.C., Chandler,C., Janssen,S., Pataki,A., *et al.* (2023) Sequence-structure-function relationships in the microbial protein universe. *Nat. Commun.*, **14**, 1–11.
27. Mukherjee,S., Seshadri,R., Varghese,N.J., Eloe-Fadrosh,E.A., Meier-Kolthoff,J.P., Göker,M., Coates,R.C., Hadjithomas,M., Pavlopoulos,G.A., Paez-Espino,D., *et al.* (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.*, **35**, 676–683.
28. Gligorijević,V., Renfrew,P.D., Kosciolek,T., Leman,J.K., Berenberg,D., Vatanen,T., Chandler,C., Taylor,B.C., Fisk,I.M., Vlamakis,H., *et al.* (2021) Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.*, **12**, 1–14.
29. Wang,Y., Huang,H., Rudin,C. and Shaposhnik,Y. (2021) Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and

PaCMAP for Data Visualization. *J. Mach. Learn. Res.*, **22**, 1–73.
30. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH--a hierarchic classification of protein domain structures. *Structure*, **5**.
31. The Gene Ontology Consortium, Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Michael Cherry,J., Davis,A.P., Dolinski,K., Dwight,S.S., *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25.
32. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33.
33. Fullmer,M.S., Ouellette,M., Louyakis,A.S., Papke,R.T. and Gogarten,J.P. (2019) The Patchy Distribution of Restriction–Modification System Genes and the Conservation of Orphan Methyltransferases in Halobacteria. *Genes* , **10**, 233.
34. Hübscher,U. (1983) DNA polymerases in prokaryotes and eukaryotes: Mode of action and biological implications. *Experientia*, **39**, 1–25.
35. Botticelli,S., La Penna,G., Minicozzi,V., Stellato,F., Morante,S., Rossi,G. and Faraloni,C. (2024) Predicting the Structure of Enzymes with Metal Cofactors: The Example of [FeFe] Hydrogenases. *Int. J. Mol. Sci.*, **25**, 3663.
36. Smith,S.P., Bayer,E.A. and Czjzek,M. (2017) Continually emerging mechanistic complexity of the multi-enzyme cellulosome complex. *Curr. Opin. Struct. Biol.*, **44**.
37. Maranga,M., Szczerbiak,P., Bezshapkin,V., Gligorijevic,V., Chandler,C., Bonneau,R., Xavier,R.J., Vatanen,T. and Kosciolek,T. (2023) Comprehensive Functional Annotation of Metagenomes and Microbial Genomes Using a Deep Learning-Based Method. *mSystems*, 10.1128/msystems.01178-22.
38. Wagner,T., Ermler,U. and Shima,S. (2016) MtrA of the sodium ion pumping methyltransferase binds cobalamin in a unique mode. *Sci. Rep.*, **6**, 1–10.
39. Ragsdale,S.W. ed. (2008) Catalysis of Methyl Group Transfers Involving Tetrahydrofolate and B12. In *Vitamins & Hormones*. Academic Press, Vol. 79, pp. 293–324.
40. Nguyen,T.-Q. and Nicolet,Y. (2022) Structure and Catalytic Mechanism of Radical SAM Methylases. *Life*, **12**, 1732.
41. Wang,T., Leibrock,N., Plugge,C.M., Smidt,H. and Zoetendal,E.G. (2023) In vitro interactions between Blautia hydrogenotrophica, Desulfovibrio piger and Methanobrevibacter smithii under hydrogenotrophic conditions. *Gut Microbes*, 10.1080/19490976.2023.2261784.
42. Levental,I. and Lyman,E. (2022) Regulation of membrane protein structure and function by their lipid nano-environment. *Nat. Rev. Mol. Cell Biol.*, **24**, 107–122.
43. Harayama,T. and Riezman,H. (2018) Understanding the diversity of membrane lipid composition. *Nat. Rev. Mol. Cell Biol.*, **19**, 281–296.
44. Jolie,R.P., Duvetter,T., Van Loey,A.M. and Hendrickx,M.E. (2010) Pectin methylesterase and its proteinaceous inhibitor: a review. *Carbohydr. Res.*, **345**, 2583–2595.
45. Wu,H.-C., Bulgakov,V.P. and Jinn,T.-L. (2018) Pectin Methylesterases: Cell Wall Remodeling Proteins Are Required for Plant Response to Heat Stress. *Front. Plant Sci.*, **9**, 412022.
46. Li,Z., Wu,L., Wang,C., Wang,Y., He,L., Wang,Z., Ma,X., Bai,F., Feng,G., Liu,J., *et al.* (2022) Characterization of pectin methylesterase gene family and its possible role in juice sac granulation in navel orange (Citrus sinensis Osbeck). *BMC Genomics*, **23**, 1–18.
47. Lupas,A.N., Ponting,C.P. and Russell,R.B. (2001) On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *J. Struct. Biol.*, **134**, 191–203.
48. Hou,J., Sims,G.E., Zhang,C. and Kim,S.-H. (2003) A global representation of the protein fold space. *Proceedings of the National Academy of Sciences*, **100**, 2386–2390.
49. Hou,J., Jun,S.-R., Zhang,C. and Kim,S.-H. (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National*

Academy of Sciences, **102**, 3651–3656.
50. Osadchy,M. and Kolodny,R. (2011) Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences*, **108**, 12301–12306.
51. Tierney,B.T., Yang,Z., Luber,J.M., Beaudin,M., Wibowo,M.C., Baek,C., Mehlenbacher,E., Patel,C.J. and Kostic,A.D. (2019) The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host Microbe*, **26**, 283–295.e8.
52. Zimmerman,S., Tierney,B.T., Patel,C.J. and Kostic,A.D. (2023) Quantifying Shared and Unique Gene Content across 17 Microbial Ecosystems. *mSystems*, 10.1128/msystems.00118-23.
53. Duller,S., Vrbancic,S., Szydłowski,Ł., Mahnert,A., Blohs,M., Predl,M., Kumpitsch,C., Zrim,V., Högenauer,C., Kosciolek,T., *et al.* (2024) Expanding the cultivated human archaeome by targeted isolation of novel Methanobrevibacter strains from fecal samples. *bioRxiv*, 10.1101/2024.04.10.588852.
54. Szydlowski,L., Ehlich,J., Szczerbiak,P., Shibata,N. and Goryanin,I. (2022) Novel species identification and deep functional annotation of electrogenic biofilms, selectively enriched in a microbial fuel cell array. *Front. Microbiol.*, **13**, 951044.
55. Szydlowski,L.M., Bulbul,A.A., Simpson,A.C., Kaya,D.E., Singh,N.K., Sezerman,U.O., Łabaj,P.P., Kosciolek,T. and Venkateswaran,K.J. (2023) Adaptation to space conditions of novel bacterial species isolated from the International Space Station revealed by functional gene annotations and comparative genome analysis. *bioRxiv*, 10.1101/2023.09.28.559980.
56. Hand,S.C., Menze,M.A., Toner,M., Boswell,L. and Moore,D. (2011) LEA Proteins During Water Stress: Not Just for Plants Anymore. *Annu. Rev. Physiol.*, **73**, 115–134.
57. Chakrabortee,S., Boschetti,C., Walton,L.J., Sarkar,S., Rubinsztein,D.C. and Tunnacliffe,A. (2007) Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proceedings of the National Academy of Sciences*, **104**, 18073–18078.
58. Leman,J.K., Weitzner,B.D., Lewis,S.M., Adolf-Bryfogle,J., Alam,N., Alford,R.F., Aprahamian,M., Baker,D., Barlow,K.A., Barth,P., *et al.* (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods*, **17**, 665–680.
59. Zhang,C., Shine,M., Pyle,A.M. and Zhang,Y. (2022) US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods*, **19**, 1109–1115.
60. Durairaj,J., Akdel,M., de Ridder,D. and van Dijk,A.D.J. (2021) Fast and adaptive protein structure representations for machine learning. *bioRxiv*, 10.1101/2021.04.07.438777.