Sequence-structure-function relationships in the microbial protein universe

Julia Koehler Leman[#*1,2], Pawel Szczerbiak[#,3], P. Douglas Renfrew[#,1,2], Vladimir Gligorijevic[1,14], Daniel Berenberg[1,4,5,14], Tommi Vatanen[6,8], Bryn C. Taylor[9,16], Chris Chandler[1], Stefan Janssen[10,15], Andras Pataki[11], Nick Carriero[11], Ian Fisk[11], Ramnik J. Xavier[6,7], Rob Knight[9,10,12,13], Richard Bonneau[1,2,4,5,14], Tomasz Kosciolek[#*3]

Affiliations
1 Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY
2 Department of Biology, New York University
3 Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland
4 Center for Data Science, New York University, New York, NY 10011, USA
5 Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY, USA
6 Broad Institute, Cambridge, MA, USA
7 Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, MA 02139
8 Liggins Institute, University of Auckland, New Zealand
9 Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
10 Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA 92093, USA
11 Scientific Computing Core, Flatiron Institute, Simons Foundation, New York, NY, USA
12 Department of Bioengineering, University of California, San Diego
13 Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.
14 present affiliation: Prescient Design, a Genentech accelerator, NY, NY 10010, USA
15 present affiliation: Algorithmic Bioinformatics, Justus Liebig University Giessen, Germany
16 present affiliation: In Silico Discovery and External Innovation, Janssen Research and Development, San Diego CA 92122

* corresponding
# These authors contributed equally

Abstract / Summary

For the past half-century, structural biologists relied on the notion that similar protein sequences give rise to similar structures and functions. While this assumption has driven research to explore certain parts of the protein universe, it disregards spaces that don't rely on this assumption. Here we explore areas of the protein universe where similar protein functions can be achieved by different sequences and different structures. We predict ~200,000 structures for diverse protein sequences from 1,003 representative genomes[1] across the microbial tree of life, and annotate them functionally on a per-residue basis. Structure prediction is accomplished using the World Community Grid, a large-scale citizen science initiative. The resulting database of structural models is complementary to the AlphaFold database, with regards to domains of life as well as sequence diversity and sequence length. We identify 148 novel folds and describe examples where we map specific functions to structural motifs. We also show that the structural space is continuous and largely saturated, highlighting the need for shifting the focus from obtaining structures to putting them into context, to transform all branches of biology, including a shift from sequence-based to sequence-structure-function based meta-omics analyses.

Introduction

Structural biology follows the sequence-structure-function paradigm, which states that the sequence of a protein determines its structure, which in turn, determines its function[2–5]. Experimental structure determination efforts were unable to keep up with the exponential growth of available sequences, yet recent breakthroughs in protein structure prediction and renewed focus on machine learning approaches, through methods like AlphaFold2[6], now allow for closing the sequence-structure gap. While disordered sequences, large complexes, multiple chains, and protein-protein interactions remain to be addressed, the large number of available protein structures and models has drastically shifted the perspective in the field. Here, we predict the structures of ~200,000 metagenomic sequences leveraging a citizen-science approach. We annotate these models in terms of protein function[7], specifically providing residue-specific annotations, and analyze the features of the resulting protein structure-function universe, including fold novelty and structure-function relationships. Our work demonstrates how to integrate massive structural datasets into a sequence and function context and motivates a shift in perspective to include structurally informed functional annotations as the starting point to understand biological questions.

Methods

Here we performed large-scale structure prediction on representative protein domains from the Genomic Encyclopedia of Bacteria and Archaea (GEBA1003) reference genome database across the microbial tree of life[1]. A summary of our workflow is shown in Fig. 1a. From a non-redundant GEBA1003 gene catalog we extracted protein sequences without matches to any structural databases and which produced multiple-sequence alignments deep enough for robust structure predictions using Rosetta[8] or DMPfold[9] (N_eff > 16, see supplement). For computational tractability we prioritized sequences according to their length and exhaustively sampled all putative novel domains between 40 and 200 residues. For each sequence we generated 20,000 Rosetta *de novo* models[8] using World Community Grid (formerly IBM) via the Microbiome Immunity Project and up to 5 models per sequence using DMPfold[9]; unless otherwise stated, we use Rosetta models for the figures in this manuscript. We then curated the initial output dataset (*MIP_raw*) of about 240,000 models to arrive at high-quality models comprising about 75% of the original dataset (*MIP_curated*). All analyses in this paper are either on *MIP_curated* or a subset. Putative new folds were identified by comparing our models against representative domains in CATH[10] and the PDB, using a TM-score cutoff[11,12] of 0.5. Putative novel folds were also verified by AlphaFold2. To contextualize our structural findings, we projected 42-dimensional graphlet vector representations of each model in a representative subset encompassing 10,000 models from *MIP_curated* onto a 3D space, along with CATH representative structures, using UMAP dimensionality reduction. Functional annotations of the entire dataset were created using structure-based Graph Convolutional Network embeddings from DeepFRI[7].

Results & Discussion

Recent advances in the availability of predicted protein structures, including the AlphaFold database and the MIP database presented here, change the view on protein sequence-structure-function relationships from a relative paucity of structural information to a relative abundance of it. This puts us in a position to start answering fundamental questions previously out of reach. How much of the protein structure and fold space is still unexplored? And can we learn anything new about the sequence-structure-function universe of microbial proteins? Here, we try to answer some of these and other questions by large-scale structure prediction efforts that we relate to the sequence space and residue-specific function prediction.

The MIP database is orthogonal to existing databases

Fig 1b shows the length distributions of protein structures in various databases. The baseline is the PDB90, which are sequences from the Protein Data Bank with a pairwise sequence identity ≤ 90%. CATH superfamilies are a non-redundant subset of the PDB90, covering over 6,000 folds (v4.3.0). The AlphaFold protein structure database[6,13] contains almost 1 million protein models, vastly increasing the known structure space, and covers a wide range of organisms and sequence lengths, primarily from Eukaryotes. Our MIP database is distinct from the other databases because it consists of proteins from Archaea and Bacteria, whose protein sequences are generally shorter than Eukaryotic[14,15]. MIP models drastically increase the available structure space of smaller proteins and domains from 40 to 200 residues (Fig 1b). We further split the sequences into domains before structure prediction, unlike structures in the AlphaFold database. Also, only about 3.6% of structures in the AlphaFold database belong to Archaea and Bacteria, indicating that AlphaFold and MIP databases are complementary.

The Rosetta models in our MIP database generally contain fewer coil residues than the DMPfold models (Figure S4), yet the quality of the DMPfold models is higher for larger proteins (see Supplement). Further, the model quality assessment score (average TM-score of 10 lowest energy models from Rosetta and the raw score from DMPfold) correlates with the TM-score[16] between the Rosetta and DMPfold models (Supplement section MIP dataset curation), indicating that models that agree better between Rosetta and DMPfold, are generally of higher quality.

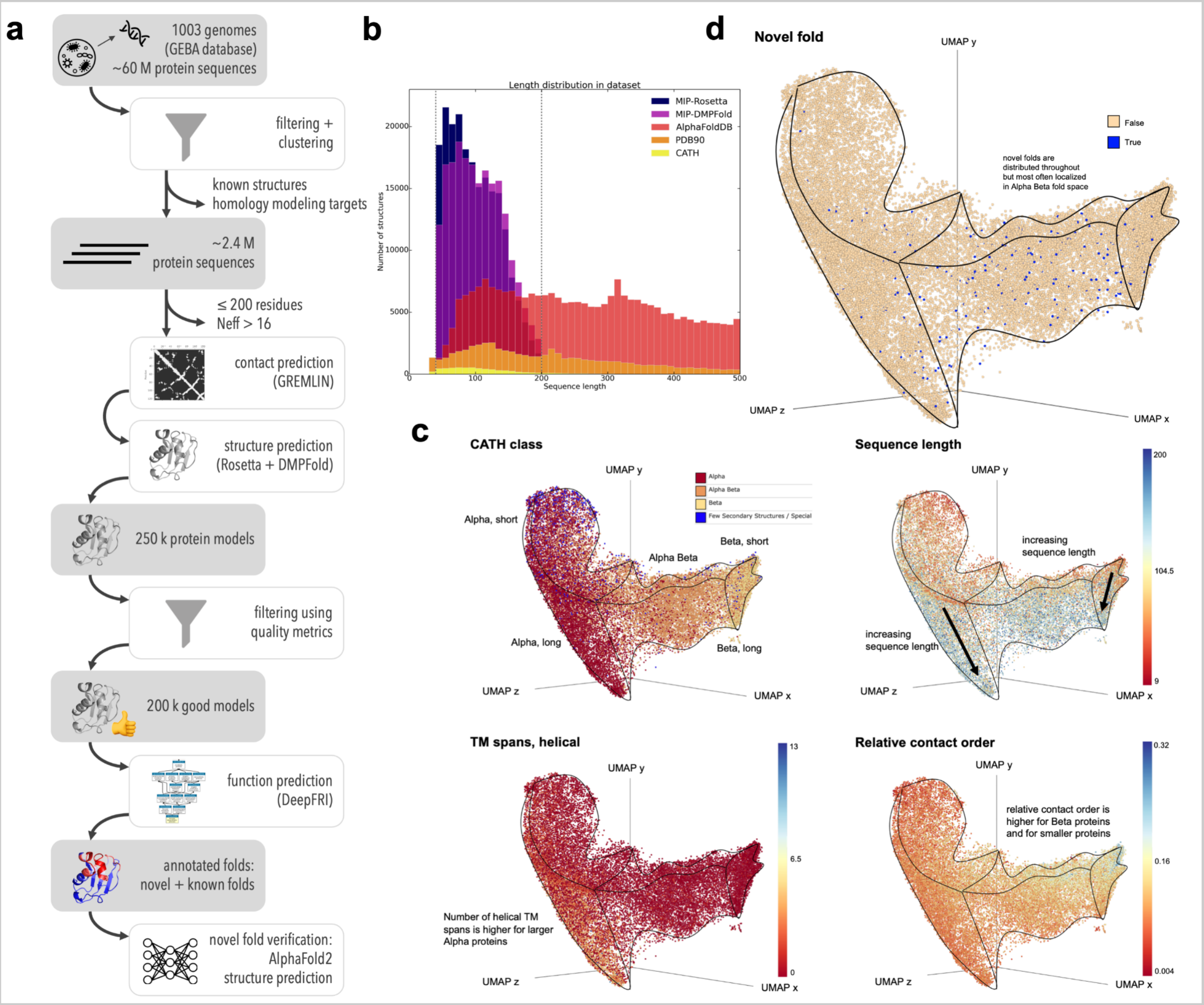


Fig 1: The fold space covered by the microbial protein structure universe is continuous. (a) Flowchart of our process to arrive at ~200,000 *de novo* protein models covering a diverse sequence space. (b) The sequence length distribution shows that our sequences are shorter than many of the proteins in the PDB, CATH or AlphaFold databases. Our proteins are between 40 and 200 residues long, which is in agreement with the fact that microbial protein sequences are often shorter than eukaryotic sequences. (c) The protein structure universe in UMAP space is color-coded according to features, such as similarity to CATH classes, sequence length, number of helical transmembrane spans, and relative contact order. (d) Novel folds (blue dots) are spread throughout the fold space with fewer representatives in the purely α-helical and purely β-sheet folds.

The microbial protein universe maps into a continuous fold space

We wanted to contextualize the MIP dataset in relation to existing structures and to investigate the features of a more complete and less biased protein structure universe[17–20]. Visualization was created by generating a 42-dimensional graphlet vector representation[21] for each model in the vizualization dataset and CATH superfamilies and mapping these vectors into 3D space using UMAP dimensionality reduction (Fig 1c and d). Visualization was done in Emperor[22]. The surfaces of the 3D structure cloud are outlined in black. We investigated several features in this mapping, including sequence length, relative contact order, number of transmembrane spans, and mapping to a CATH class. The 3D mapping of the protein universe allows to distinguish different sequence lengths, the number of helical transmembrane spans and the relative contact order of the protein folds, as different shadings show in Fig 1c. The visualization further illustrates that the protein universe space is continuous, indicating that folds may evolve along a trajectory where small changes in the tertiary structure can eventually lead to a different fold, which is in agreement with prior work[23,24]. In contrast, a discrete fold space would display distinct clusters of folds that require larger conformational changes to interconvert between them. We identify 438 previously unseen structures in our MIP dataset that cluster into 148 distinct, novel folds (46 clusters with multiple proteins and 102 singletons). Fig 1d shows that the majority of novel folds are distributed throughout α/β fold space (compare with Fig 1c) with few novel folds in α or β fold space.

MIP dataset explores the sequence-structure-function universe

We computed pairwise similarities between 3,052 random sequences in the curated dataset (*MIP_random5000_curated* - see Table S3) in terms of sequence identity, structural similarity (TM-score) and functional similarity (cosine similarity score). This was compared against a PDB baseline of 1000 protein chains, covering pairwise sequence similarities between 0 and 100%.

By design, protein sequences in the MIP dataset are dissimilar (30% sequence identity cutoff). When correlating sequence similarities to structural similarities for pairs of proteins, the vast majority of dissimilar sequences fold into different structures. However, there are a fair number of proteins that have vastly different sequences and still fold into similar structures (Fig 2b). The PDB baseline that covers sequence similarities across all ranges from 0 - 100% confirms this expected trend, and it also confirms the general notion that similar sequences fold into similar structures (Fig 2a).

When correlating sequence identity and functional similarity, the majority of sequences have different functions, but still a fair number of dissimilar sequences have similar functions. This originates in the multiplicity of biological systems (Fig 2d). i.e. achieving the same functional outcome by different pathways (for example [25,26]). The PDB baseline gives the same trend and has an additional known population where similar sequences achieve similar functions (Fig 2c).

When correlating structural similarity (of dissimilar sequences!) to functional similarity, we find 4 populations (Fig 2f): (a) the largest population following expectations of dissimilar structures having different functions - quadrant III, (b) the 2nd largest population of dissimilar structures having similar functions - quadrant I, (c) the third largest population of similar structures having different functions - quadrant IV, and (d) the smallest population following expectations of similar structures having similar functions. Quadrants I and IV are the most interesting ones with examples shown in Fig 4. The PDB baseline covers all sequence similarities and follows mainly known expectations of quadrants II and III (Fig 2e).

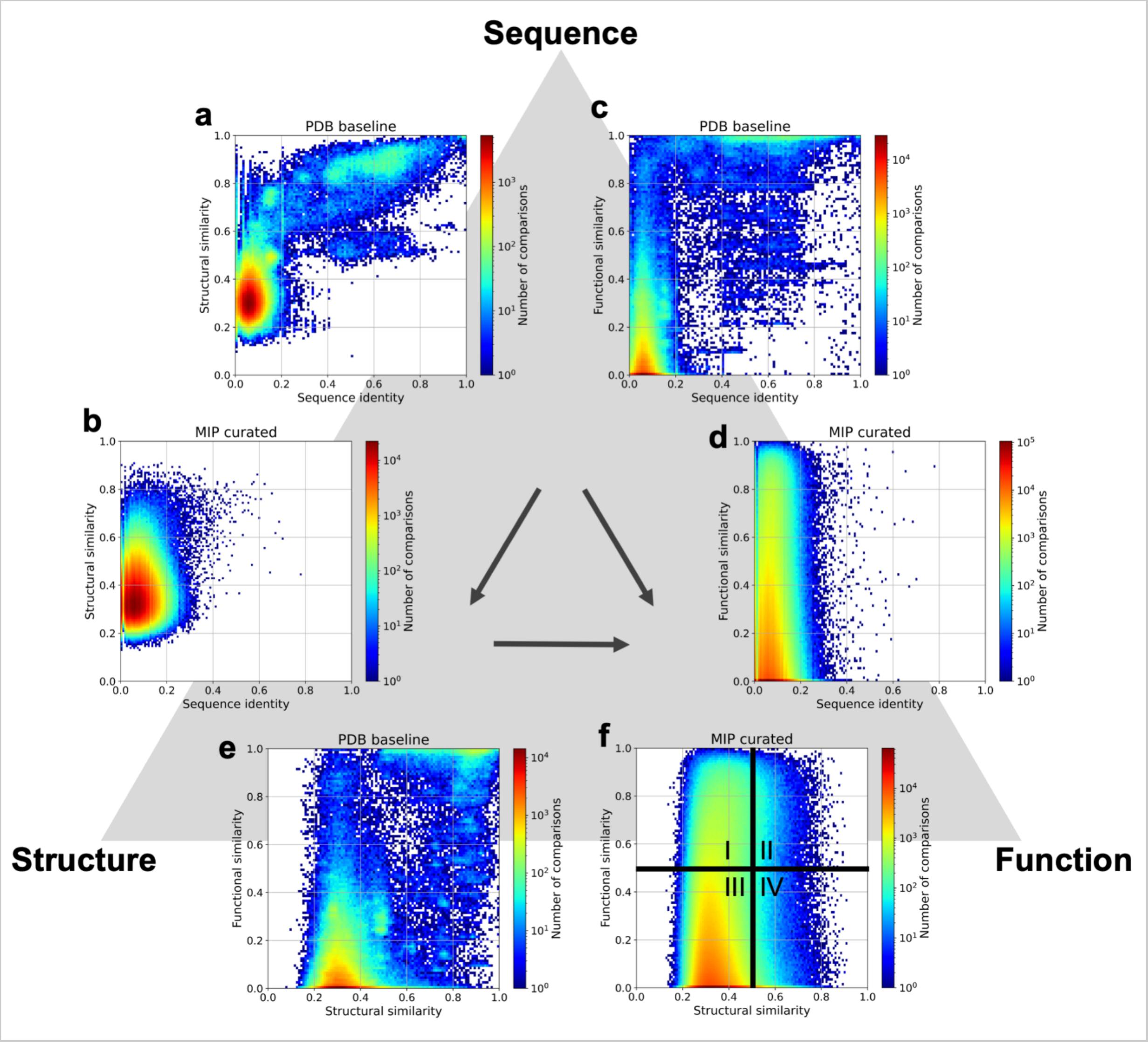


Fig 2: Sequence-structure-function relationships in both PDB and the MIP dataset. Pairwise comparisons of protein sequences (using sequence identity), structures (TM-score), and functions (cosine similarity between DeepFRI output vectors) for two datasets: a baseline from the PDB and the MIP_random5000_curated dataset, containing 3,052 Rosetta generated models (see Table S3). The PDB baseline dataset contains 1000 chains covering pairwise sequence similarities between 0 and 100% while the MIP dataset is a non-redundant set with mostly dissimilar sequences (sequence identity < 30% threshold was imposed before sequential domain splitting). Analyses of these two datasets in this way lead us to the following conclusions: sequence identity correlates with structural similarity (a), yet high structural similarity can be achieved by low sequence identity (b). High sequence identity leads to high functional similarity (d), yet high functional similarity can be achieved by proteins with low sequence identity (d). Structural similarity often correlates with functional similarity ((e) and quadrants II and III in (f)). However, there are plenty of examples where low structural similarity can be seen in proteins with high functional similarity (quadrant I in (f)), and highly similar structures can exhibit different functions (quadrant IV in (f)).

Most functions are produced by the same structural motifs, even for dissimilar sequences

For each of the 148 novel fold structural clusters, we compared functional similarities for each protein pair by computing the cosine similarity for the function vectors; this is shown as a heatmap in Fig 3. We then picked several proteins for each structural cluster and mapped selected top-scoring functions onto the predicted structures (right panel in Fig 3). Residues that our function prediction network DeepFRI predicts to have high importance to achieve a particular function are highlighted in red, whereas blue residues are not involved in generating that particular function. Many examples of this analysis are outlined in the supplement, some of which are shown in Figs 3 and 4. We find that the majority of functions in those structural clusters map to the same residues in the structure ("structural motif") as shown for the largest cluster 161 in Fig 3. However, we also find more complicated structure-function relationships in these clusters as shown in Fig 4 and discussed in the next section. Note that the sequences in each structural cluster (and in the MIP dataset) are dissimilar to each other and neither structural nor functional prediction could be inferred by sequence identity for these proteins due to lack of homology.

Per-residue functional annotations reveal a more complex picture of protein structure-function relationships

Some of the structure-function relationships map to quadrants I and IV in Fig 2f, where similar structures can have different functions or different structures can have similar functions. To a first approximation this is not surprising. Similar structures can achieve different functions due to the fact that the gene ontology database is organized in a hierarchical manner and that parent or child functions are related but still different[27]. Different structures can generate similar functions due to the multiplicity of functional pathways[25,26] as a back-up plan for organisms to survive. However, a closer look at some of the structure-function disparities reveals some surprises.

Figs 4a and b show two proteins that use the same structural motif for different functions. While the overall sequence identity between these proteins is low (~30 and 25% for panels A and B, respectively), a short sequence motif underlies the structural motif, which in turn has different functions. Fig 4a shows two proteins where the terminus of the central helix is involved in phosphatase activity, where the same motif in a different protein is involved in actin binding. The sequence motif for this region is GGWDXP. In Fig 4b, the N-terminus of one protein is involved in zinc ion binding and 'catalytic activity, acting on a protein', whereas the N-terminus of another protein of that structural cluster is involved in DNA binding and 'identical protein binding'. The underlying sequence motif for this structural motif is CXCCG.

Figs 4c, 4d, and 4e show examples where a different structural motif in the same protein fold achieves the same function. This seems unusual and doesn't seem to rely on a short sequence motif. In the first example (Fig 4c), transferase activity either maps to a beta-sheet or a C-terminal short helix in two different proteins. In the second example transmembrane transporter activity maps to either two helices or a beta-sheet (Fig 4d). Fig 4e shows that ion transmembrane transporter activity maps to different structural motifs for different proteins. This entire structural cluster (cluster 146) has very high similarity across predicted functions, indicated by the heatmap showing mostly yellow hues.

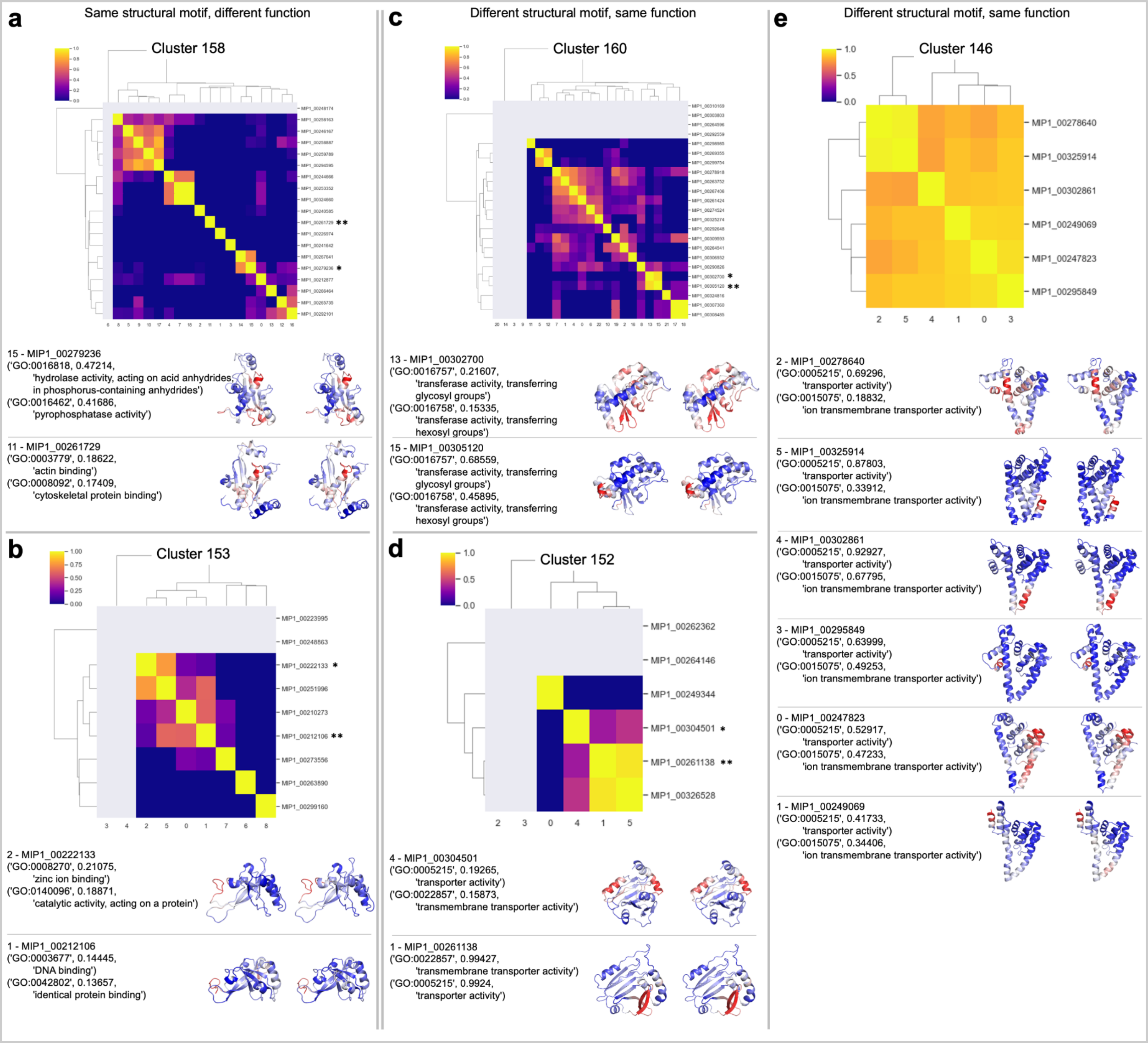


Fig 3: Functional diversity of proteins with the same structure.
We show examples from several structural clusters (Rosetta models) that exhibit novel folds. The heatmaps show functional similarity (cosine similarity of the function vectors) of protein pairs within the cluster. Proteins that have predicted functions with scores < 0.1 are shown in gray in the heatmaps. Asterisks highlight the examples shown below. (a) and (b) show cases where the same structural motif in two different proteins produces different, unrelated functions. (c), (d), and (e) show cases where the same function is generated by different structural motifs in different proteins, even though the proteins have the same fold.

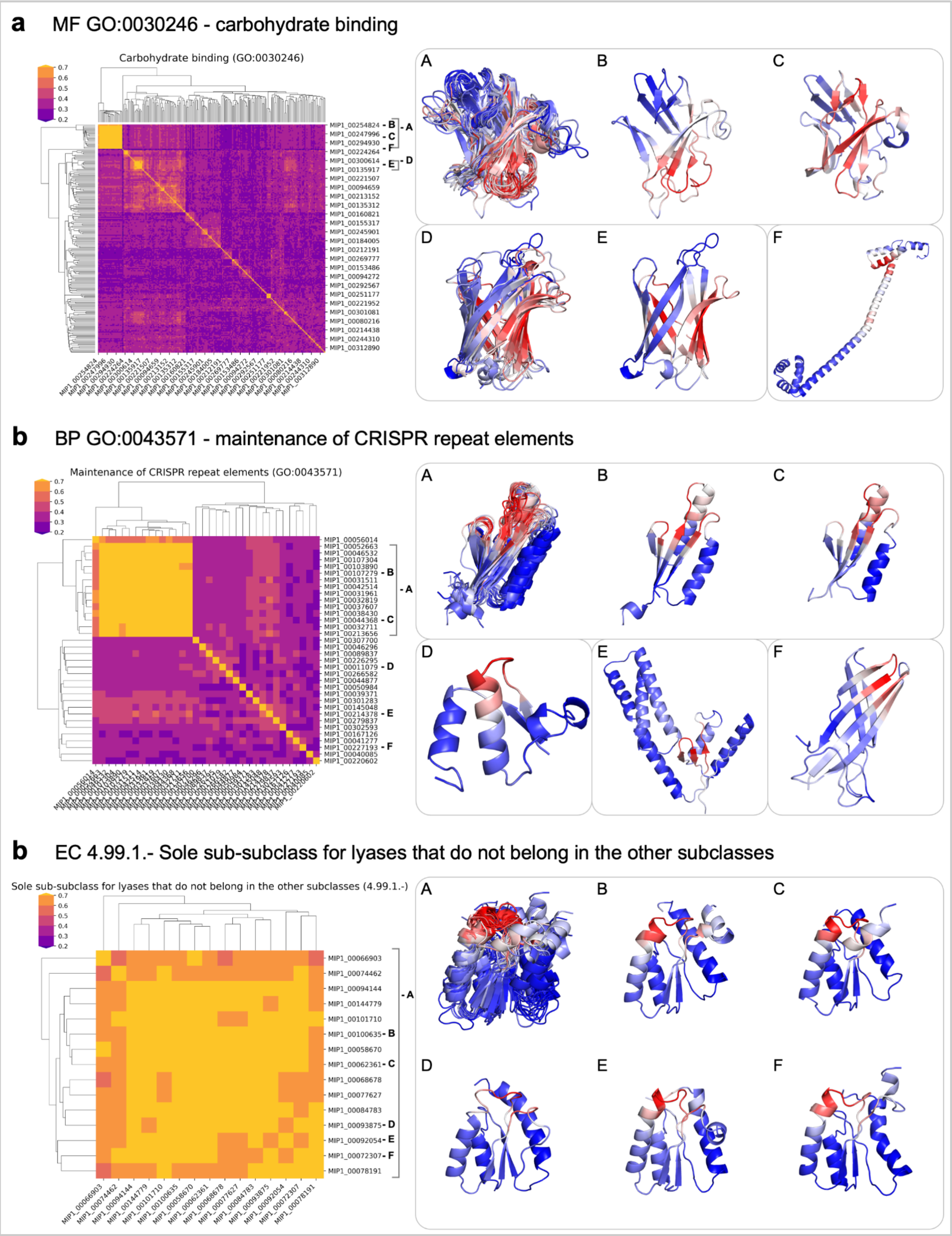
Higher functional specificity is carried out by fewer possible folds

We investigated different protein functions and examined the structures with these functions (Fig. 4 and Figs. S88 – S94). Some protein functions are sufficiently general such that they can be achieved by different folds, examples are 'carbohydrate binding' (GO:0030246), 'protein tyrosine kinases' (EC 2.7.10.), and 'mitigation of host immune response by virus' (GO:0030683). More specific functions are accomplished by fewer folds. Examples of specific functions with a single fold in our MIP dataset are 'thymidine kinase' (EC 2.7.1.21) and 'sole sub-class for lyases that do not belong in the other subclasses' (EC 4.99.1.).

The functional cluster for carbohydrate binding (Fig. 4a) covers many different folds with high β-sheet propensity, including β-barrels, twisted sheets, and stacked sheets. This class contains a single helical protein, indicated by the single blue line in the heatmap in Fig. 4a with the structure shown in (F). The largest structural cluster in this functional category corresponds to the largest novel-fold cluster (yellow square in the heatmap) and the salient residues in this cluster show a high degree of overlap.

Fig. 4b shows the function 'maintenance of CRISPR repeat elements'. CRISPR repeats are short DNA sequences in bacteria and archaea. They derive from DNA fragments of bacteriophages that previously infected those organisms and allows them to identify recurring invaders. Hence, the CRISPR-Cas system functions like a microbial immune system[28]. Cas1 and Cas2 identify the site in the bacterial genome where viral DNA is inserted and ultimately cleaved by Cas9[29]. The structural cluster in Fig. 4b (A) overlays with part of Cas2 (PDB ID 5sd5 or 5xvp, chains EF) and the predicted salient residues bind DNA in the structure. Cluster (D) in Fig. 4b is similar to parts of Cas1 (PDB ID 5sd5 or 5xvp, chains ABCD) but doesn't overlay perfectly.

Fig. 4c shows the function 'sole sub-class for lyases that do not belong in the other subclasses'. Lyases are enzymes that catalyze the breaking of chemical bonds by means other than hydrolysis or oxidation. None of the lyases in the other classes (EC. 4.1 - EC.4.6) have the same fold as our predicted MIP models, even though there are structural similarities. Our models have an (αβ)x3 fold with sequential strand connections – the other lyases have various (αβ)xN folds but their strand connections are non-sequential.



a MF GO:0030246 - carbohydrate binding

b BP GO:0043571 - maintenance of CRISPR repeat elements

b EC 4.99.1.- Sole sub-subclass for lyases that do not belong in the other subclasses

Fig 4: Structural diversity of proteins with the same function.

We examine proteins that have the same function and plot the TM-score as a measure of structural similarity as a heatmap, with larger numbers (more yellow) representing more similar structures. We also map the residue-specific function predictions onto the structures on the right, where residues in red are responsible for the functions. (a) Gene ontology molecular function carbohydrate binding with GO number GO:0030246. Except for the protein shown in (F) which has high helical propensity, the proteins in this functional cluster have high β-sheet content. The largest cluster in the heatmap in yellow is also the largest novel-fold cluster. The salient residues responsible for this function overlay nicely across the proteins in this cluster. (b) Gene ontology biological process function 'maintenance of CRISPR repeat elements' with GO number GO:0043571. The largest cluster highlighted in yellow superimposes with Cas2 and the salient residues in red interact with DNA. (c) Enzyme commission number EC 4.99.1. with the function 'Sole sub-class for lyases that do not belong in the other subclasses'. All structures in this functional cluster have the same fold and the salient residues responsible for this function overlay onto the same structural motif in the protein. More details in the text.

Conclusion

In this study, we used a citizen-science approach to predict ~200,000 protein structures for non-redundant microbial sequences across the tree of life. Structures were predicted by two state-of-the-art independent methods (Rosetta and DMPfold) and evaluated by quality metrics to indicate model quality. Functional annotations give us a unique look at the microbial protein universe in terms of sequence, structure, and function. Our database is orthogonal to the AlphaFold database in terms of domains of life, sequence diversity and sequence length. We predicted 148 novel folds which were verified by AlphaFold. With functional annotations, we can more closely relate sequence-structure-function relationships in this universe, that go beyond the main homology assumption of similar sequences, structures and functions. In fact, we frequently see that these dissimilar sequences fold into similar structures, indicating that the sequence diversity is much greater than the structural diversity. From a structure prediction standpoint, this highlights the importance of distant homology detection and fold recognition methods for dissimilar sequences. Moreover, we provide examples that challenge our classic understanding of biological sequence-structure-function relationships. We hope that this research inspires the scientific community to advance our understanding of site-specific protein function by developing experimental and computational tools for high-quality measurements and predictions. Only these new tools can lead to a more complete understanding of the complexities of how proteins fold, function, evolve and interact, to address questions related to health, disease and engineering applications to solve some of the world's biggest problems.

Contributions

RK, RB, RJX, TK conceived and initiated the project; TK, JKL coordinated the project; PDR, TK prepared and supervised World Community Grid project execution; VG developed the methodology and performed functional annotations; DB, PS developed, implemented and performed low-dimensional representation of the protein space; PS, JKL, TK, PDR analyzed the data; JKL, TK, PDR, PS wrote the manuscript with input from all authors; CC, PDR prepared computational framework for data aggregation and analysis; TK, PDR, JKL, BCT, TV performed preliminary data analysis; SJ, TK implemented sequence processing pipeline; AP, IF, NC provided HPC cluster and computational support.

Conflicts of interest

RB, VG and DB are currently working at Genentech and no explicit conflicts of interest result from this change in affiliation.

Online Methods

Sequence clustering of GEBA dataset

The MIP dataset is constructed on the basis of GEBA1003 representative bacterial and archeal genomes from across the tree of life[1]. The dataset includes environmental samples from soil, ocean water, human gut microbiome and was designed to sample the microbial tree of life evenly. For each genome, we generated a list of predicted genes using Prodigal[30]. The raw gene catalog was processed using an incremental clustering approach, similar to the one employed by UniClust[31]. First, redundancy in the dataset was removed by using `linclust` (ie. clustering at 100% sequence identity), as implemented in MMSeqs2[32,33]. Then, the dataset was further clustered into 90%, 70% and 30% sequence identity clusters, with the last step (70% to 30% clustering) executed using the MMSeqs2 `clust` module. The resulting dataset was sorted according to sequence length, sampling the entirety of sequences between 40 and 200 residues.

Rosetta structure prediction

The structure of each MIP sequence was predicted using version 2016.32.58837 of the Rosetta Macromolecular Modeling Suite, modified to run on the IBM World Community Grid. Residue-residue contacts from sequences closely-related to the target sequence were inferred using GREMLIN[34] (version 2.0.1) and incorporated as constraints during the folding protocol. For each MIP sequence, 20,000 models were generated. Models were ranked using the REF2015 energy function[35] and the lowest energy model was used for further analysis. Details of the fragment selection, contact prediction, and Rosetta model building can be found in the supplement.

DMPfold structure prediction

We additionally predicted the structures of all MIP sequences using DMPFold[9]. The same multiple sequence alignments used for contact prediction in the Rosetta structure prediction pipeline were used, instead of DMPfolds default method of generating an MSA from the uniclust30 database. All other parameters were kept to their default values. Details of the DMPfold model building can be found in the Supplement.

Quality metrics: pairwise sequence identity, TM-score, cosine similarity

Model quality measures and construction of the MIP curated is discussed in detail in the Supplement (see section *MIP dataset curation*). MQA score for AlphaFold2 predictions is the mean pLDDT (in some places we also provide pTM values). Pairwise sequence identity and structural similarity (TM-score) were generated using TM-align. Two structures were identified as similar (including novel fold identification) if the corresponding TM-score >= 0.5 (unless otherwise stated). Pairwise function similarity was computed as a cosine similarity between concatenated DeepFRI output vectors (which comprise scores for 6315 GO terms/EC numbers) with threshold 0.1 i.e. scores < 0.1 were replaced with 0 (we noticed that using full output introduces too much noise).

Protein universe visualization

For every structure in the MIP visualization dataset (comprising 10,000 Rosetta and corresponding DMPfold models plus 6,631 CATH 4.3.0 superfamily structures - see Tables S2 and S3), we generated a contact map (representing residues closer than < 6Å from each other), which was then transformed into graph. Such graph representation was subsequently used to form a 42-dimensional graphlet vector[21]. The collection of graphlet vectors (26,631 x 42 matrix) was then projected onto a 3D space using two standard dimensionality reduction methods i.e. UMAP and PCA. For UMAP we used the following set of parameters (which provided reasonable spread of the data with clear CATH class separability): `n_neighbors = 100`, `min_dist = 0.001`, `N_components = 3`, `metric` = `cosine`. Visualizations were created with Emperor[22]. An overview of the pipeline is depicted in Fig. S36.

Meta-data

MIP models were superimposed against all CATH 4.3.0 superfamilies (http://www.cathdb.info/, accession January 5, 2021) using TM-align in order to filter out novel folds (see below) and annotate them using CATH classification i.e. the most similar CATH structure to a given protein (with the highest TM-score normalized by MIP model size) is used as a template. The annotation quality drops with decreasing TM-score (which is important for novel folds) but we noticed that in general the quality is high (especially at the class level). Proteins were annotated as alpha transmembranes if their OCTOPUS output contained at least one “M“ segment. Similarly, proteins were annotated as beta transmembranes if their BOCTOPUS output contained at least eight “pL“ segments. All the other metadata are discussed in the Supplement.

Novel fold identification

To identify new folds we started from high quality MIP models i.e. the MIP curated dataset. First, we performed a TM-align structural superposition against CATH 4.3.0 superfamilies (see above). For the models without any significant structural similarity to CATH (TM-score < 0.5), we performed a superposition against representative structures from the PDB90 (time stamp January 15, 2021). A putative novel fold is a high quality (i.e. MIP curated) single domain predicted by both Rosetta and DMPfold with satisfactory confidence (agreement TM-score >= 0.5 between Rosetta and DMPfold predictions) with a maximum TM-score < 0.5 against CATH and the PDB90. Note that when comparing MIP and CATH/PDB structures we use the TM-score normalized by the MIP sequence length. The output set contained 452 structures grouped into 161 clusters. AlphaFold2 verification found 14 false positives which resulted in 438 novel structures grouped into 148 novel fold clusters. See the Supplement for more information.

DeepFRI prediction

DeepFRI computes saliency maps for each predicted GO term[7]. These maps identify residues that are important for this function but that does not mean that these are active residues, they could be important for protein stability or short sequence motifs away from the active site to identify the function of this protein. Heatmaps in Fig. 4 were generated based on curated MIP (Rosetta) models with DeepFRI score >= 0.2. Models were then grouped by GO-term and pairwise structural similarity was computed as the maximum TM-score of two superimposed MIP models (i.e. the larger of TM-scores normalized by the first and second sequence lengths was chosen).

Data availability

All sequence, structure and function data generated in the project, along with relevant metadata are deposited on Zenodo with the DOI 10.5281/zenodo.6477242 and on Github at https://github.com/microbiome-immunity-project/protein_universe. Information on the directory structure and how to search the database are available both on Zenodo.

Code availability

TM-align v.20190822 (https://zhanggroup.org//TM-align/) was used for computing TM-scores and sequence identities of aligned structures[11]. Structure visualizations were created in Pymol v.2.4.0 (https://github.com/schrodinger/pymol-open-source). Secondary structure assignments were generated using Stride v.20021022[36]. Alpha-helical transmembrane annotations were generated using OCTOPUS (as a part of TOPCONS2 software[37]; singularity image downloaded on July 17, 2020, dependencies: Blast v.2.2.26, Uniref90 v.20200119, Pfam 20191204). Beta-strand transmembrane annotations were generated using BOCTOPUS2[38] (zip downloaded on August 8, 2020; dependencies: HH-suite v.2.0.16, Blast v.2.2.26, Uniprot20 v.20160226). Absolute and relative contact order was computed from definition[39]. For disordered sequence identification we used MobiDB-lite[40] v.1.0 (March 2016) and DISOPRED3[41] (zip downloaded on September 16, 2021; dependencies: Blast[42] v.2.2.26, Uniref90 v.20210731). Putative new fold clusters were computed using Python package NetworkX v.2.7.1. For putative new fold verification, we used AlphaFold2 with "preset" flag set to `full_pdb` (repository downloaded on August 16, 2021; reference databases which includes the PDB downloaded on July 31, 2021). A cosmetically modified version of the Rosetta Macromolecular Modeling Suite[43,44], based on release 2016.32.58837, was used for protein structure prediction on the World Community Grid. The fragment picking pipeline[45] is also part of the standard Rosetta distribution. Both are obtainable from the Rosetta Commons (https://www.rosettacommons.org/). Residue-residue pair constraints were obtained using GREMLIN[34] version 2.0.1. DMPfold[9] (https://github.com/psipred/DMPfold, downloaded September 2019) was used to predict the structures of all MIP sequences.

Acknowledgments

We kindly acknowledge the wonderful support and encouragement of the IBM World Community Grid team (Caitlin Larkin, Juan A Hindo, Al Seippel, Erika Tuttle, Jonathan D Armstrong, Kevin

Reed, Ray Johnson, and Viktors Berstis), and the community of 790,000 volunteers who donated 140,661 computational years since Aug 2017 their computer time over the course of the project. This research was also supported in part by PLGrid Infrastructure. The authors thank Hera Vlamakis and Damian Plichta from the Broad Institute for helpful discussions.

The research reported in this publication was supported by the Flatiron Institute as part of the Simons Foundation to JKL, PDR, VG, DB, CC, AP, NC, IF, and RB. This research was also supported by grants NAWA PPN/PPO/2018/1/00014 to PS and TK, PLGrid to PS, and NIH - DK043351 to TV and RJX.

References - 30-50

1. 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., Yoshikuni, Y., Visel, A., Whitman, W. B., Garrity, G. M., Eisen, J. A., Hugenholtz, P., Pati, A., Ivanova, N. N., Woyke, T., Klenk, H. P. & Kyrpides, N. C. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol. 2017 357* **35,** 676–683 (2017).
2. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (80-.).* **181,** 223–230 (1973).
3. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nat. 1970 2255232* **225,** 563–564 (1970).
4. Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S. M. Q., Roodveldt, C. & Tawfik, D. S. The 'evolvability' of promiscuous protein functions. *Nat. Genet. 2004 371* **37,** 73–76 (2004).
5. Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **18,** 394–402 (2008).
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596,** 583–589 (2021).
7. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K. & Bonneau, R. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun. 2021 121* **12,** 1–14 (2021).
8. Koehler Leman, J., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., Correia, B. E., Coventry, B., Das, R., De Jong, R. M., DiMaio, F., Dsilva, L., Dunbrack, R., Ford, A. S., Frenz, B., Fu, D. Y., Geniesse, C., Goldschmidt, L., Gowthaman, R., Gray, J. J., Gront, D., Guffy, S., Horowitz, S., Huang, P. S., Huber, T., Jacobs, T. M., Jeliazkov, J. R., Johnson, D. K., Kappel, K., Karanicolas, J., Khakzad, H., Khar, K. R., Khare, S. D., Khatib, F., Khramushin, A., King, I. C., Kleffner, R., Koepnick, B., Kortemme, T., Kuenze, G., Kuhlman, B., Kuroda, D., Labonte, J. W., Lai, J. K., Lapidoth, G., Leaver-Fay, A., Lindert, S., Linsky, T., London, N., Lubin, J. H., Lyskov, S., Maguire, J., Malmström, L., Marcos, E., Marcu, O., Marze, N. A., Meiler, J., Moretti, R., Mulligan, V. K., Nerli, S., Norn, C., Ó'Conchúir, S., Ollikainen, N., Ovchinnikov, S., Pacella, M. S., Pan, X., Park, H., Pavlovicz, R. E., Pethe,

M., Pierce, B. G., Pilla, K. B., Raveh, B., Renfrew, P. D., Burman, S. S. R., Rubenstein, A., Sauer, M. F., Scheck, A., Schief, W., Schueler-Furman, O., Sedan, Y., Sevy, A. M., Sgourakis, N. G., Shi, L., Siegel, J. B., Silva, D. A., Smith, S., Song, Y., Stein, A., Szegedy, M., Teets, F. D., Thyme, S. B., Wang, R. Y. R., Watkins, A., Zimmerman, L. & Bonneau, R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17,** 665–680 (2020).

9. Greener, J. G., Kandathil, S. M. & Jones, D. T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun. 2019 101* **10,** 1–13 (2019).
10. Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J. & Orengo, C. A. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49,** D266–D273 (2021).
11. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33,** 2302–2309 (2005).
12. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26,** 889 (2010).
13. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50,** D439–D444 (2022).
14. Zhang, J. Protein-length distributions for the three domains of life. *Trends Genet.* **16,** 107–109 (2000).
15. Gong, X., Fan, S., Bilderbeck, A., Li, M., Pang, H. & Tao, S. Comparative analysis of essential genes and nonessential genes in Escherichia coli K12. *Mol. Genet. Genomics* **279,** 87–94 (2008).
16. Zhang, Y. & Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. **710,** 702–710 (2004).
17. Hou, J., Jun, S. R., Zhang, C. & Kim, S. H. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 3651–3656 (2005).
18. Hou, J., Sims, G. E., Zhang, C. & Kim, S. H. A global representation of the protein fold space. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 2386–2390 (2003).
19. Levitt, M. & Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 5913–5920 (1998).
20. Osadchy, M. & Kolodny, R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12301–12306 (2011).
21. Faisal, F. E., Newaz, K., Chaney, J. L., Li, J., Emrich, S. J., Clark, P. L. & Milenković, T. GRAFENE: Graphlet-based alignment-free network approach integrates 3D structural and sequence (residue order) data to improve protein structural comparison. *Sci. Rep.* **7,** (2017).
22. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: A tool for visualizing high-throughput microbial community data. *Gigascience* **2,** 1–4 (2013).
23. Holm, L. & Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26,** 316–319 (1998).
24. Taylor, W. R. Evolutionary transitions in protein fold space. *Curr. Opin. Struct. Biol.* **17,** 354–361 (2007).
25. Chiesa, S., Mingueneau, M., Fuseri, N., Malissen, B., Raulet, D. H., Malissen, M., Vivier,

E. & Tomasello, E. Multiplicity and plasticity of natural killer cell signaling pathways. *Blood* **107,** 2364–2372 (2006).

26. Guillén, D., Sánchez, S. & Rodríguez-Sanoja, R. Carbohydrate-binding domains: multiplicity of biological roles. *Appl. Microbiol. Biotechnol.* **85,** 1241–1249 (2010).
27. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat. Genet. 2000 251* **25,** 25–29 (2000).
28. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of Bacteria and Archaea. *Science (80-.).* **327,** 167–170 (2010).
29. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117,** 119–128 (2015).
30. Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 1–11 (2010).
31. Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Soding, J. & Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45,** D170–D176 (2017).
32. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun. 2018 91* **9,** 1–8 (2018).
33. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol. 2017 3511* **35,** 1026–1028 (2017).
34. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins Struct. Funct. Bioinforma.* **79,** 1061–1078 (2011).
35. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., Dimaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13,** 1–35 (2017).
36. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23,** 566–579 (1995).
37. Tsirigos, K. D., Peters, C., Shu, N., Käll, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43,** W401-7 (2015).
38. Hayat, S. & Elofsson, A. BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. *Bioinformatics* **28,** 516–522 (2012).
39. Shi, Y., Zhou, J., Arndt, D., Wishart, D. S. & Lin, G. Protein contact order prediction from primary sequences. *BMC Bioinformatics* **9,** 1–9 (2008).
40. Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33,** 1402–1404 (2017).
41. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31,** 857–863 (2015).
42. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
43. 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., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., Correia, B. E., Coventry, B., Das, R., De Jong, R. M., DiMaio, F., Dsilva, L., Dunbrack, R., Ford, A. S., Frenz, B., Fu, D. Y., Geniesse, C., Goldschmidt, L., Gowthaman, R., Gray, J. J., Gront, D.,

Guffy, S., Horowitz, S., Huang, P. S., Huber, T., Jacobs, T. M., Jeliazkov, J. R., Johnson, D. K., Kappel, K., Karanicolas, J., Khakzad, H., Khar, K. R., Khare, S. D., Khatib, F., Khramushin, A., King, I. C., Kleffner, R., Koepnick, B., Kortemme, T., Kuenze, G., Kuhlman, B., Kuroda, D., Labonte, J. W., Lai, J. K., Lapidoth, G., Leaver-Fay, A., Lindert, S., Linsky, T., London, N., Lubin, J. H., Lyskov, S., Maguire, J., Malmström, L., Marcos, E., Marcu, O., Marze, N. A., Meiler, J., Moretti, R., Mulligan, V. K., Nerli, S., Norn, C., Ó'Conchúir, S., Ollikainen, N., Ovchinnikov, S., Pacella, M. S., Pan, X., Park, H., Pavlovicz, R. E., Pethe, M., Pierce, B. G., Pilla, K. B., Raveh, B., Renfrew, P. D., Burman, S. S. R., Rubenstein, A., Sauer, M. F., Scheck, A., Schief, W., Schueler-Furman, O., Sedan, Y., Sevy, A. M., Sgourakis, N. G., Shi, L., Siegel, J. B., Silva, D. A., Smith, S., Song, Y., Stein, A., Szegedy, M., Teets, F. D., Thyme, S. B., Wang, R. Y. R., Watkins, A., Zimmerman, L. & Bonneau, R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods 2020 177* **17,** 665–680 (2020).

44. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3,** 171–6 (1999).

45. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M. & Baker, D. Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS One* **6,** e23294 (2011).