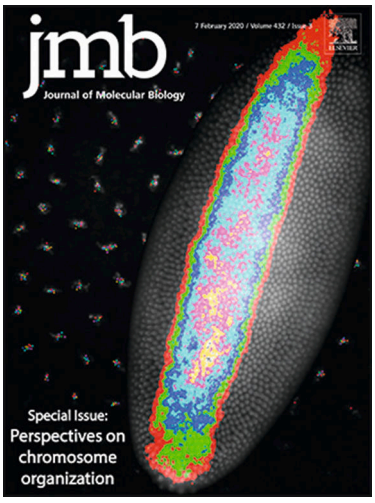
Journal Pre-proofs



Database

CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds

Vaishali P. Waman, Nicola Bordin, Rachel Alcraft, Robert Vickerstaff, Clemens Rauer, Qian Chan, Ian Sillitoe, Hazuki Yamamori, Christine Orengo

PII: S0022-2836(24)00146-3
DOI: https://doi.org/10.1016/j.jmb.2024.168551
Reference: YJMBI 168551

To appear in: *Journal of Molecular Biology*

Received Date: 31 January 2024
Revised Date: 20 March 2024
Accepted Date: 22 March 2024

Please cite this article as: V.P. Waman, N. Bordin, R. Alcraft, R. Vickerstaff, C. Rauer, Q. Chan, I. Sillitoe, H. Yamamori, C. Orengo, CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds, *Journal of Molecular Biology* (2024), doi: https://doi.org/10.1016/j.jmb.2024.168551

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 The Author(s). Published by Elsevier Ltd.

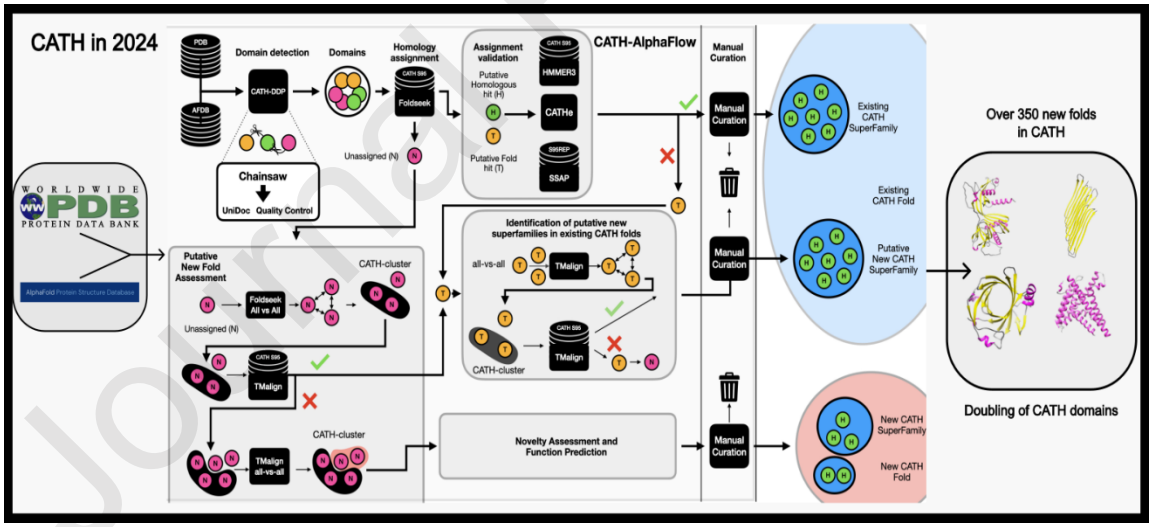
CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds

Vaishali P. Waman[1]*, Nicola Bordin[1]*, Rachel Alcraft[2], Robert Vickerstaff[2], Clemens Rauer[1], Qian Chan[1], Ian Sillitoe[1], Hazuki Yamamori[1], Christine Orengo[1a]

[1] Institute of Structural and Molecular Biology, University College London, London, United Kingdom

[2]Advanced Research Computing Centre, University College London, London, United Kingdom

*Contributed equally, [a] corresponding author c.orengo@ucl.ac.uk



Journal Pre-proofs

Highlights

- New CATH-AlphaFlow workflow classifies experimental and predicted protein structures
- CATH-AlphaFlow doubles the number of structural domains classified in CATH
- Nearly 350 new CATH folds identified (nearly 200 not seen in other resources)

Abstract

CATH (https://www.cathdb.info) classifies domain structures from experimental protein structures in the PDB and predicted structures in the AlphaFold Database (AFDB). To cope with the scale of the predicted data a new NextFlow workflow (CATH-AlphaFlow), has been developed to classify high-quality domains into CATH superfamilies and identify novel fold groups and superfamilies. CATH-AlphaFlow uses a novel state-of-the-art structure-based domain boundary prediction method (ChainSaw) for identifying domains in multi-domain proteins. We applied CATH-AlphaFlow to process PDB structures not classified in CATH and AFDB structures from 21 model organisms, expanding CATH by over 100%.

Domains not classified in existing CATH superfamilies or fold groups were used to seed novel folds, giving 253 new folds from PDB structures (September 2023 release) and 96 from AFDB structures of proteomes of 21 model organisms. Where possible, functional annotations were obtained using (i) predictions from publicly available methods (ii) annotations from structural relatives in AFDB/UniProt50. We also predicted functional sites and highly conserved residues. Some folds are associated with important functions such as photosynthetic acclimation (in flowering plants), iron permease activity (in fungi) and post-natal spermatogenesis (in mice).

CATH-AlphaFlow will allow us to identify many more CATH relatives in the AFDB, further characterising the protein structure landscape.

Keywords: CATH, AlphaFold2, protein domain, fold, superfamily, protein structure prediction

Introduction

Journal Pre-proofs

The number of protein sequences increased exponentially following the advent of next-generation sequencing, and far outpaces the number of experimentally solved 3D protein structures in the Protein Data Bank (PDB). There are ~215,000 experimentally solved protein structures in PDB [1] while UniProtKB contains ~227 million protein sequences. To cope with the massive sequence data, methods for accurate protein structure prediction (e.g. homology modelling and *ab initio* approaches) have been pursued for decades. The Protein Structure Prediction Center hosts a biannual contest via CASP (Critical Assessment of Structure Prediction) which rigorously assesses the performance of these methods [2]. In the latest CASP14, AlphaFold2, a sequence-based AI method (developed by Google's DeepMind) significantly outperformed all other 145 methods and provided good quality models [3].

AlphaFold2 is trained with structures deposited in the PDB, and co-evolution information learned from multiple sequence alignments generated from sequence database searches. AlphaFold2's breakthrough in CASP14 has revolutionised structural biology research by narrowing the protein sequence-structure gap. Subsequently, the AlphaFold Database (AFDB) (https://alphafold.ebi.ac.uk/) was established in a collaboration between Google's DeepMind and EMBL-EBI and the latest release contains 3D-models for 214 million UniProt sequence entries, a major step in bridging the sequence-structure gap [4].

AFDB provides a rich source of information for structural classification and structure-based function prediction. Protein structures can typically reveal more distantly related homologues than sequence alone as structure tends to be more conserved than sequence in evolution. A number of structural classifications have been established (CATH, SCOP, SCOPe, ECOD) which focus on the domain since domains are semi-independent functional units of proteins that can independently evolve and can be arranged in combination with other domains to evolve new or modified protein functions. Both CATH and ECOD have already performed analyses of subsets of the AFDB data [5–7] and shown that there is a substantial amount of good quality predicted structural data which can expand the classification resources significantly.

CATH is a Global Core Biodata Resource (GCBR) which classifies domains into a hierarchy of Class, Architecture, Topology/fold and Homologous superfamily. Class classifies domains based on the type of secondary structure elements (i.e. alpha, beta, alpha/beta, few secondary structures) whilst Architecture classifies domains based on the arrangement of secondary structure elements. Topology (fold) considers both the arrangement and connectivity of secondary structure elements. Homologous superfamily classifies domains that share sufficient structural and sequence similarity to infer homology from a common ancestor. Within the superfamily, CATH provides functionally coherent groups known as CATH-Functional Families (FunFams)[8,9]. In its latest release (version 4.3), CATH classifies 150,885 PDB structures, segregated into 495,811 domains, classified into 5841 homologous superfamilies[10].

CATH has grown steadily since it was established in the early 1990s, but the release of >200 million AlphaFold predicted models is likely to represent a 400-fold or more expansion in domain structures and requires the development of new pipelines to process this data in a timely manner. Since the first release of AlphaFold2, various groups have introduced new computational methods and pipelines for clustering and segmenting AlphaFold2 chains. For example, the Steinegger group has clustered full-chain AlphaFold2 models from AFDB (version 3) into 2.27 million structure clusters and suggested 31% of these do not match any structure deposited in PDB [11]. Durairaj and colleagues analysed AFDB (version 4) dataset and used the structural data to find new families and functional relationships providing novel functional insights into several protein families associated with the dark proteome[12]. Recently, the Grishin group developed DPAM, a novel domain predictor to process AFDB chains and provide structural assignments for domains in 48 proteomes, including human [6,7].

Our group developed a new protocol (CATH-Assign) for processing AFDB models. This includes domain detection, evaluation of model quality and subsequent classification in CATH. CATH-Assign uses a combination of approaches involving new deep-learning methods for protein structure comparison (Foldseek [13]) and protein language model based methods for remote homology detection (CATHe, developed in-house [14]). It was used to perform a preliminary analysis of AFDB models from 21 model organisms, classifying 341,213 domain structures in CATH superfamilies and revealing 25 novel folds with at least one human representative [5].

In this study, we report the extension of CATH-Assign with new deep learning based methods for domain detection (Chainsaw)[15]. To cope with the scale of the data, Chainsaw is faster and more sensitive than previous algorithms used by CATH. We also report the development of a novel pipeline (CATH-AlphaFlow, Figure 1) which encodes major steps of the CATH-Assign protocol in a NextFlow workflow. CATH-AlphaFlow were applied to all novel structures in the PDB not currently classified in CATH. It was also applied to the AFDB structures from the 21 model organisms to refine domain boundaries and improve classification of domains.

Journal Pre-proofs

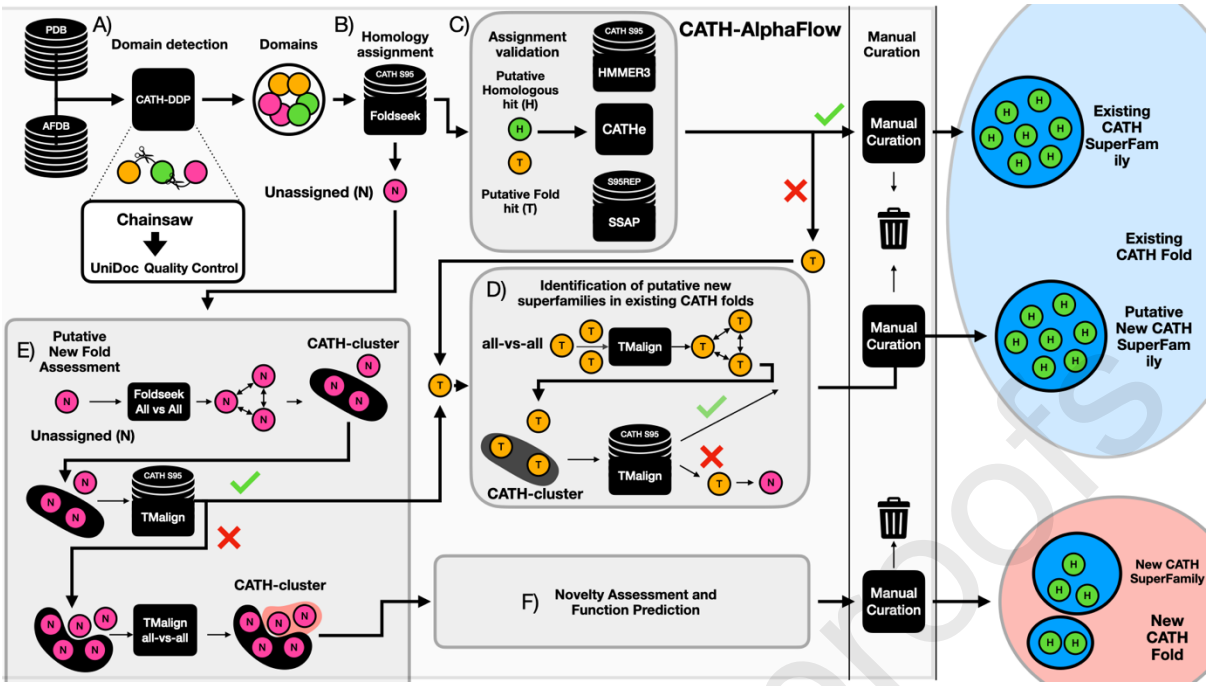


Figure 1. Outline of the CATH classification pipeline.

Application of CATH-AlphaFlow to the PDB and AFDB structures expanded CATH by 112% to 1,060,659 domain structures and brought 349 new folds into CATH (253 from PDB structures and 96 from AFDB). We used various public resources and tools to obtain functional annotations for these.

CATH-AlphaFlow and the functional annotation strategies outlined here are robust and fast and will assist in mining the vast data released by AFDB and related platforms (e.g. 3D-Beacons https://www.ebi.ac.uk/pdbe/pdbe-kb/3dbeacons/). The novel domain assignments and fold groups enabled by CATH-Assign/AlphaFlow will be available from the CATH-beta daily snapshot (ftp://orengoftp.biochem.ucl.ac.uk/).

Materials and Methods

We updated CATH by applying CATH-AlphaFlow, which includes a new domain detection algorithm (Chainsaw [15]) to all the protein structures in the PDB not classified in CATH (v4.3). Our dataset included PDB structures deposited up to September 2023.

We also applied CATH-AlphaFlow with Chainsaw to improve domain boundary assignment and validate the classification of AFDB structures from 21 model organisms.

1. Domain detection

Journal Pre-proofs

Domain boundaries were assigned for both PDB structures and AFDB structures using Chainsaw [15], a novel state-of-the-art deep-learning algorithm benchmarked against other widely used methods (e.g. UniDoc [16]). Low-probability assignments were subsequently validated with UniDoc [16] and manual curation. As expected, Chainsaw gave significantly better domain boundaries than our previous method for detecting domains in UniProt proteins, CATH-resolve-hits (CRH) [17], which relies purely on sequence data. Chainsaw has various advantages over CRH, including better accuracy, higher speed when GPU-accelerated and requires a single PDB or MMCIF file as input. Domain boundaries detected by Chainsaw were subsequently used to extract the domain regions from PDB and AFDB files using the pdb-selres module from pdb-tools [18].

Supplementary Figure 1 shows how the agreement between CRH and Chainsaw boundaries falls as the sequence similarity between the query and the closest relative in CATH falls. Structure-based methods like Chainsaw will also be much better at detecting boundaries for domains with no homologues in CATH.

Removing low quality structures

Domains from the PDB chains or AFDB were assessed for quality using the metrics described in Bordin et al [5]. Well-packed globular domains with an average pLDDT of 70 or more, more than 3 secondary structure elements, less than 30% of the domain residues in Long Unordered Regions, and less than 65% of residues in unordered regions were subsequently processed by CATH-AlphaFlow.

2. Domain Processing with CATH-AlphaFlow

Using CATH-AlphaFlow, segmented domains from AFDB models and PDB structures were assigned to CATH. CATH-AlphaFlow is a series of Python modules created to perform consistent processing of protein chains (either from models or from experimental data), many of which have been orchestrated in NextFlow (https://github.com/UCLOrengoGroup/cath-alphaflow). The steps involved are discussed as follows and highlighted in Figure 1.

CATH assignment using Foldseek

Domains were scanned against CATH superfamilies using Foldseek and tentatively assigned to CATH superfamilies using thresholds benchmarked on a set of curated CATH/SCOP assignments as described in [5]. We consider valid hits at 5% error rate against a library of CATH clustered at 95% sequence identity (S95) from 5,841 Superfamilies in CATH classes 1 to

4, with a 60% overlap between query and target and a bitscore cutoff of 130 for homology (CATH superfamily-level, H) and 98 for fold detection (Topology-level, T) (Figure 1B).

Validation of Foldseek structural matches by structure comparisons with SSAP

Structural matches by Foldseek were confirmed by re-comparing the matched pairs with SSAP [19] against the S95 representative hit in the Foldseek search (Figure 1C). Pairs with a SSAP score >= 70 and a residue overlap of 60% (of the larger domain) were considered valid hits. Higher SSAP scores (>=80) suggest homology which was subsequently validated by sequence based approaches described below.

Validation of superfamily assignment by sequence-based methods

Domain sequences for H-hits were extracted from the PDB or AFDB files using the pdb_tofasta module from pdb-tools and further validated by scanning them against Hidden Markov Models built using HMMER3 [20] from S95 representatives of CATH (Figure 1C). We used thresholds established for CATH-Gene3D[10], an e-value cut-off of 1e-3 for homology assignments with a minimum bitscore cutoff of 25 and overlap of 80%. Query domains matching the same superfamily by HMMER3 scan and Foldseek/SSAP, were assigned to that superfamily.

We also validated homology using CATHe [14], a CATH superfamilies predictor based on embeddings from the ProtT5 protein language model. Structural matches which had CATHe predictions with a probability of 90% were considered valid superfamilies assignments (Figure 1C).

3. Clustering domains assigned to the same CATH fold group into new CATH SuperFamilies

Query domains which matched a particular fold group in CATH were compared against each other all-versus-all using TMalign [21] with a minimum overlap of 60% and TMscore cutoff of 0.7 normalised by the length of the largest domain (as benchmarked in [5]) (Figure 1D). The resulting score matrix was clustered with complete linkage clustering using cath-cluster, from the cath-tools suite (https://cath-tools.readthedocs.io/en/latest/). This gave a set of putative new superfamilies.

4. Clustering domains with no match to a CATH superfamily or fold group to identify putative novel fold groups

Domains not meeting our criteria for inclusion in a CATH fold group or superfamily were clustered by performing an all-vs-all scan with Foldseek with an overlap of 60% and a bitscore cutoff of 130, and the resulting output clustered with complete linkage using cath-cluster. Since TMalign is more sensitive than Foldseek, each cluster representative was subsequently scanned with TMalign against all CATH S95 representatives with an TMscore cutoff of 0.5 and a minimum overlap of 60% to check for fold hits missed by the initial Foldseek scans. Cluster representatives without a CATH fold assignment after this step were compared against each other using TMalign and clustered using cath-cluster with complete linkage, a TMscore cutoff of 0.5 and an overlap of 60%, to give a set of putative novel fold groups (Figure 1E).

FoldCheck: Representatives from putative novel fold groups were superposed to the closest domain in CATH identified by TMalign using cath-superpose (https://cath-tools.readthedocs.io/en/latest/tools/cath-superpose/) and visualised in UCSF Chimera [22] to assess whether they were novel architectures. Fold novelty was further assessed by searching for matches in other classification resources (ECOD, SCOPe) with TMalign (TMscore=0.5, overlap 60%) and SSAP (SSAP score >=70, overlap 60%) followed by extensive manual curation by the CATH curator (Figure 1F).

5. Identification of structural relatives for novel folds/superfamilies in AFDB/UniProt50

Putative novel folds from AFDB and PDB were searched against representatives from AFDB clustered at 50% sequence identity (AFDB/UniProt50) via the Foldseek web server (https://search.foldseek.com/search) to identify all structural relatives. This information was used to determine the size of the structural superfamily and analyse the taxonomic distribution. The multiple sequence alignment generated from all the relatives (TM-score $\geq$ 70) was used to identify conserved sites using the Scorecons program [23].

6. Functional annotation of novel folds

For putative novel folds from AFDB domains we used fold representatives to annotate functions using various approaches described below.

Curation of available functional information

We extracted functional annotations from InterPro, Pfam and UniProt (i.e. GO terms, EC number, information on interactions, functional site information or literature evidence of function).

Assigning predicted functions using sequence and structure based methods

We predicted functional annotations using a number of sequence-based and structure-based methods.

1) DeepFRI, structure-based (threshold score > 0.50)[24],

Journal Pre-proofs

2) PROST, embedding-based method (default parameters)[25].

3) Foldseek scans on the AFDB/UniProt50 database, inheriting GO annotations from structural relatives (TM-Score >0.70)

4) Information on functional partner proteins (from STRING/IntAct).

Predicting functional sites

We predicted functional sites using DeepFRI (score >0.5) and P2RANK [26] (Score >0.50). Additionally, we identified conserved sites by analysing the multiple sequence alignment of structural relatives in AFDB/UniProt50 database using the Scorecons program (threshold score >0.70[23]).

Results

1. Classification of protein structures in the Protein Databank (PDB) currently unclassified in CATH

A total of 108,130 PDB structures, unclassified in CATH, were analysed using CATH-AlphaFlow. CATH-AlphaFlow consists of multiple steps illustrated in Figure 1. Chainsaw identified 212,942 constituent domains within the protein chains. Subsequent scanning of the domain structures against the S95 representatives from the CATHv4.3 superfamilies by Foldseek algorithm gave a total of 151,648 matches to the CATH superfamily (H-level) or fold group (T-level).

Further validation of the Foldseek matches was performed by verifying the structural similarity using TMalign and the in-house SSAP algorithm (see Methods). Assignment to CATH superfamilies was subsequently verified by scanning against CATH S95 HMMs and also by CATHe (see Methods). Overall, 137,193 could be assigned to CATH superfamilies and a further 14,455 to CATH fold groups, whilst the remaining 61,294 are putative novel folds in CATH (see Supplementary Figure 2).

Journal Pre-proofs

These were compared all-against-all using Foldseek and subsequently clustered into 6944 clusters (see Methods 4). Representatives from these clusters were scanned against CATH S95 domain structure representatives using the slower but more sensitive TMalign. This step assigned 3,238 clusters to known fold groups in CATH. Clustering of domains assigned to these fold groups (see Methods 3) gave 1697 additional putative CATH superfamilies.

The remaining 3706 domain representatives were subjected to quality checks, the FoldCheck protocol (see Methods 4) and manual inspection for further evaluation of domain quality and verification of new folds. We removed ~92% which had problematic features (e.g. poor resolution, high proportion of long regions of residues with no secondary structure and poor packing i.e. lacking globularity), multi-domain and synthetic proteins (see supplementary figure 3).

A total of 253 domains were validated as non-problematic globular folds, new to CATH. A significant proportion of these (161/253 i.e. 64%), had already been classified in other resources such as ECOD and SCOPe, giving 92 folds newly identified in our analyses.

As a final check on novelty and to assign architectures, we compared the representatives of these folds with their close structural matches in CATH obtained from the TMalign search. The highest-scoring matches were superposed with the query domain using the cath-superpose tool (https://cath-tools.readthedocs.io/en/latest/tools/cath-superpose/) to confirm the domains were not extremely remote homologues and assess whether there were similarities in architecture. Matches to the core regions were examined particularly carefully as the core (typically >40% of the structure) is typically conserved even between very remote homologues. No representatives were sufficiently structurally similar, nor had functional evidence to suggest an evolutionary relationship with, or same fold as, any CATH domain.

For novel folds where the closest matches had different architectures, new architectures were assigned. Some interesting novel architectures and topologies were identified (see Figure 2), discussed below.

2. Processing AFDB chains from 21 model organisms using CATH-AlphaFlow

Earlier pilot work analysing 369,512 high-quality AF2 domains from 21 model organisms [5] reported that 92.3% (341,213 domains) could be assigned to CATH superfamilies. Unassigned

domains were clustered into 4,235 structural clusters. Preliminary manual analysis of 610 clusters containing human relatives revealed 25 globular domain structures likely to be novel folds. Many of the remainder comprised sequence-based matches to Pfam and CATH families which had problematic domain boundary assignments.

To improve the domain segmentation process, we re-processed all cluster representatives using Chainsaw. The resulting 6,266 domains were then processed using CATH-AlphaFlow modules (e.g. Foldseek, SSAP, CATH-HMM, CATHe) for assignment to CATH (as for the unclassified PDB structures above). The improved domain boundary assignments allowed us to assign 431 domains existing CATH superfamilies and a further 496 to CATH fold groups. The remaining 5,835 domains were scanned in an all-vs-all fashion using Foldseek and clustered into 4,644 structural clusters (see Methods Section 4).

Cluster representatives were scanned against CATH using TMalign which is slower but more sensitive than FoldSeek. 1836 representatives matched to CATH superfamilies. 2,478 representatives matched to CATH fold groups and were clustered into 2,477 putative novel superfamilies. Expanding to all domains, 347,479 AFDB domains could be assigned to CATH superfamilies.

Combining these AFDB domains with the 212,942 experimental domain structures from the PDB which we bring into CATH (see above) represented an increase in CATH domain structures by 560,421 to 1,060,659, a 112% increase. This also gave an increase in the number of CATH superfamilies (i.e. for PDB and AFDB domains with T-hits but not H-hits) from 5,481 to 9,655. We expect these superfamily numbers to reduce in the future (see discussion).

We manually evaluated the remaining 330 putative novel fold clusters. As with the PDB structures, many representatives had problematic features such as remaining segmentation issues, high proportion (or long regions) of residues with no secondary structure, poor packing i.e. lacking globularity. We also used FoldCheck to see whether any were extremely remote homologues of a CATH superfamily. In total, 75 AFDB domains were identified as new folds in addition to the 21 identified in our pilot work, giving a total of 96 novel folds identified in the AFDB dataset of 21 model organisms.

3. Analysis of the novel folds identified in the PDB and AFDB domain structure clusters

We examined all 253 (from PDB) and 96 (from AFDB) novel folds identified. Some (161/253) of the PDB domains had already been classified in SCOPe (version 2.08) and ECOD (version

20230309). A selection of novel folds from PDB and AFDB with highly unusual architectures/topologies are shown in Figure 2.

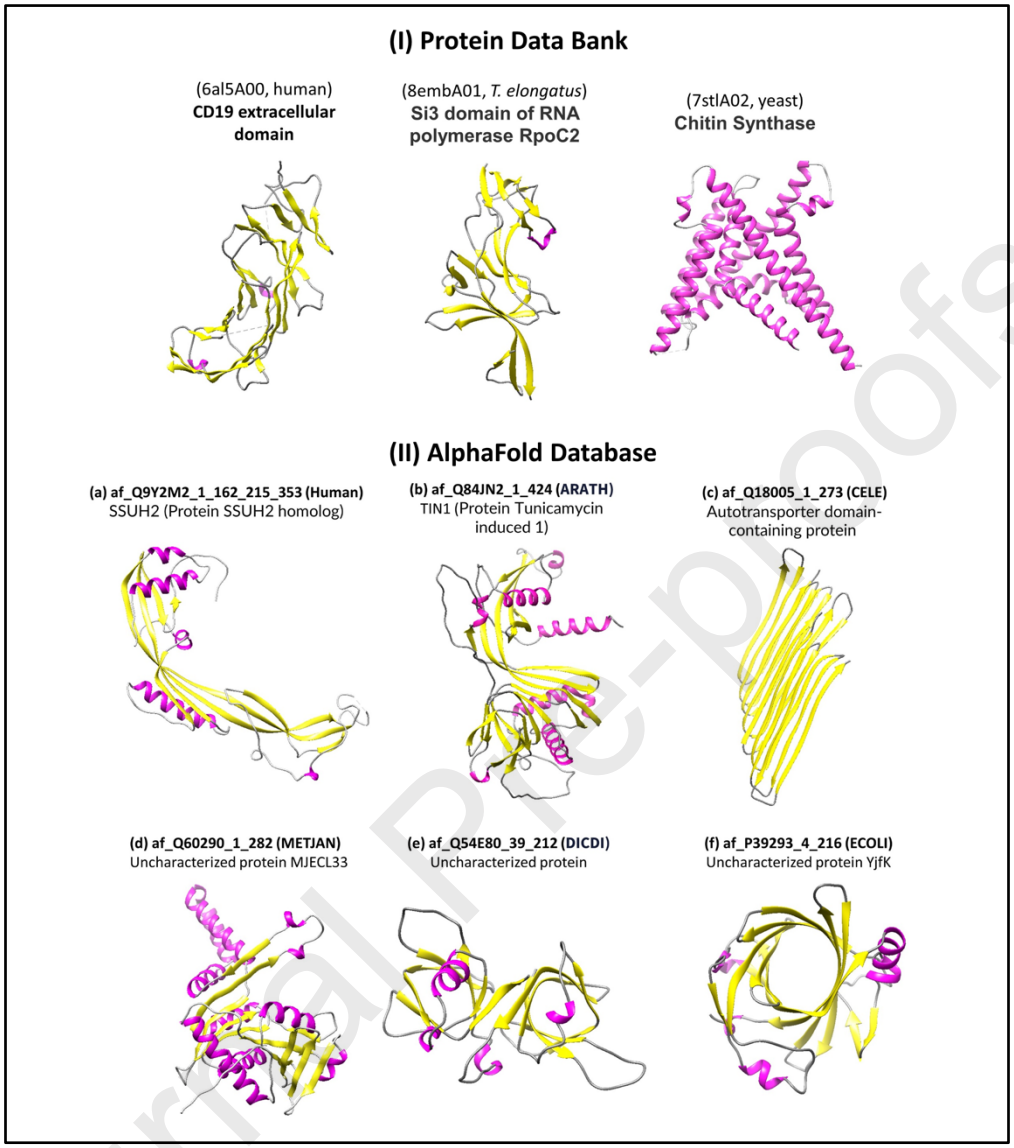


***Figure 2**:* Illustration of novel folds with unusual topologies/architectures. The top panel are PDB domains, lower panel AFDB domains.

Predicting function and mapping functional sites for AFDB domains

Where possible, we obtained functional annotations for the novel folds. Pfam annotations indicate that most are associated with membrane proteins and involved in transmembrane transport (e.g. potassium transporter) iron permease, and oxidoreductase activities (see Supplementary Table-1).

For 63 with no annotations, we predicted their function using the structure-based method DeepFRI, identified putative functional sites using P2RANK/Scorecons/DeepFRI as well as additional annotations by literature and UniProt searches.

Journal Pre-proofs

Their predicted function suggests an association with important biological processes such as spermatogenesis (in mice); photosynthetic acclimation, pollen development and proteasomal degradation in flowering plants; iron permease activity in yeast; odontogenesis in human and female meiosis pathway in Caenorhabditis elegans (Further details in Supplementary Tables 1-3). In bacteria, novel domain folds are involved in functions such as pentosyltransferase activity; 4 iron 4 sulfur cluster binding and cytochrome c oxidase assembly. Supplementary tables:2-3 provides details of annotations from DeepFRI and PROST, and also UniProt GO terms inherited from other structural relatives in the AFDB. Figure 3 shows a selection of examples where the AFDB structures provided useful functional insights.

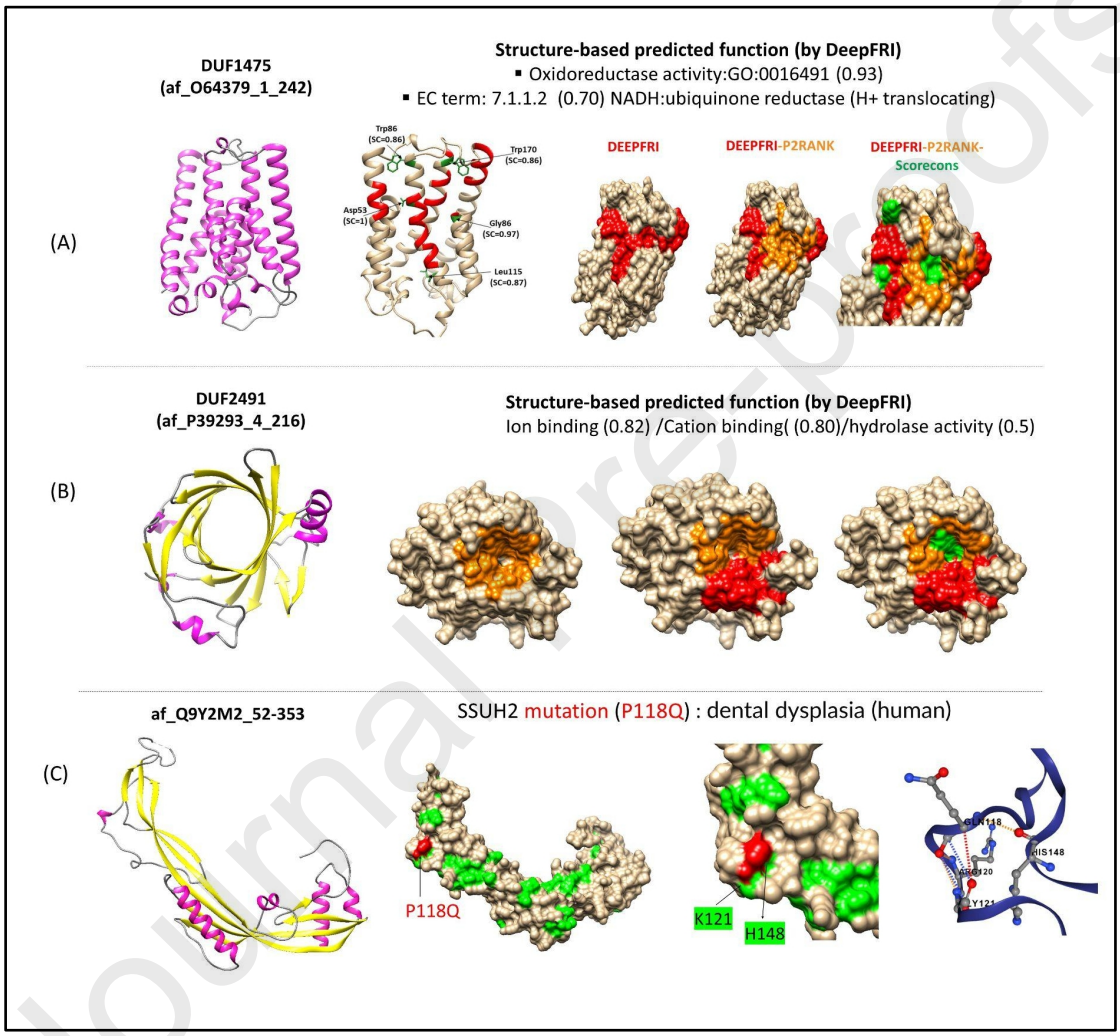


Figure 3: *AFDB annotations using DeepFRI (red), P2RANK (orange) and Scorecons (green). The SSUH2 domain is linked with a human genetic disorder caused by the impact of missense mutation (P118Q) which lies close to a conserved site (shown in C).*

Functional insights into UniProt O64379 - AT1G22750 protein

O64379 (domain 1-242) is from a functionally uncharacterised Pfam family (DUF1475/ PF07343) containing relatives from *Arabidopsis thaliana*. UniProt annotations suggest it is present in plant vacuoles and membranes. Foldseek scans identified 63 structural relatives in

AFDB/UNIPROT50 (TM-score >0.70) from other flowering plants including common tobacco, sunflower, maize, winter squash, peanut, pepper, etc. DeepFRI provides high confidence predictions for ion channel activity (GO:0005216), redox activity (GO:0022900) and electron transport chain (GO:0022900) processes (>0.80), whilst STRING data suggests co-expression with a lipid dehydrogenase having oxidoreductase activity and localised on membranes including vacuoles. Overall, the evidence supports a transmembrane transporter with electron transport chain/oxidoreductase activity in plant vacuoles. Some of the DeepFRI predicted sites (score > 0.80), are also highly conserved (Scorecons>0.85) (Asp 53, Gly86, Trp170, Leu115, Trp46) (see **Figure 3A**) and overlap with P2RANK predicted pocket sites (0.80) suggesting a role in function.

Functional insights into af_P39293_20_216

Domain af_P39293_20_216 (from YjfK protein) is found in pathogenic bacteria including *E.Coli*, Salmonella. IntAct data suggests interaction with yhjD, involved in lipopolysaccharide transport (https://www.ebi.ac.uk/intact/search?query=id:P39293*#interactors). DeepFRI predicts ion binding activity for this protein (0.80). Again there is good correlation of predicted sites among the different methods (P2RANK, DeepFRI and Scorecons) (see **Figure 3B**).

iii. Structural insights into mutations linked to human disease

The SSUH2 gene is associated with a pathogenic missense variant (P118Q), involved in the genetic disorder Dentin Dysplasia Type I (which causes dentin defects in humans [27]). Foldseek scans found 74 structural relatives in AFDB, the alignment of which revealed several high-confidence conserved sites (Score >0.70) (see **Figure 3C)**. The pathogenic mutation lies close to two Scorecons predicted sites (121 and 148, score =0.70). Analyses of the impact using Dynamut2 and mCSM (-0.85 Kcal/mol), suggested that the mutation affects atomic interactions with the proximal conserved sites and is de-stabilising.

Discussion

We applied a novel computational workflow (CATH-AlphaFlow) to classify domains from PDB in CATH, bringing 212,942 PDB domains into CATH and giving 253 novel CATH folds (92 of which are not observed in other classifications). We also applied CATH-AlphaFlow to the AFDB predicted structures for 21 model organisms. This confirmed the previous assignments of 341,213 domain structures to CATH superfamilies [5]. Improved domain assignments by ChainSaw enabled a further 6,266 domains to be accurately detected and assigned to CATH superfamilies. Analysis of the remaining AFDB domains revealed 96 new folds/superfamilies (including the 21 identified in earlier pilot work [5]).

For those folds not annotated in Pfam and other databases, we used a variety of state-of-the-art tools to predict functions and functional site information. High-confidence DeepFRI-predictions were available for 50 novel folds, some of which are associated with important

biological processes in mice (fertility), flowering plants (photosynthetic acclimation, proteasomal degradation), and yeast (iron permease).

Although studies of the PDB in 2012 suggested the domain fold library was nearly complete, the identification of 349 novel folds in this study (increasing CATH folds >25% to 1739) suggests more novelty remains to be elucidated in the AFDB database.

By applying our new classification workflow (CATH-AlphaFlow) we doubled the number of domains in CATH to over one million. 212,942 of these are experimental and mean that CATH is now up-to-date with the September-2023 version of PDB. We are now applying CATH-AlphaFlow to all AFDB entries to expand CATH further by processing all AFDB entries. Assigning AFDB domains to CATH superfamilies will help in bringing functional annotations to the UniProt sequences.

References

[1] D.R. Armstrong, J.M. Berrisford, M.J. Conroy, A. Gutmanas, S. Anyango, P. Choudhary, A.R. Clark, J.M. Dana, M. Deshpande, R. Dunlop, P. Gane, R. Gáborová, D. Gupta, P. Haslam, J. Koča, L. Mak, S. Mir, A. Mukhopadhyay, N. Nadzirin, S. Nair, T. Paysan-Lafosse, L. Pravda, D. Sehnal, O. Salih, O. Smart, J. Tolchard, M. Varadi, R. Svobodova-Vařeková, H. Zaki, G.J. Kleywegt, S. Velankar, PDBe: improved findability of macromolecular structure data in the PDB, Nucleic Acids Res. (2019) gkz990. https://doi.org/10.1093/nar/gkz990.

[2] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)—Round XIV, Proteins Struct. Funct. Bioinforma. 89 (2021) 1607–1617. https://doi.org/10.1002/prot.26237.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[4] M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences, Nucleic Acids Res. 52 (2024) D368–D375. https://doi.org/10.1093/nar/gkad1011.

[5] N. Bordin, I. Sillitoe, V. Nallapareddy, C. Rauer, S.D. Lam, V.P. Waman, N. Sen, M. Heinzinger, M. Littmann, S. Kim, S. Velankar, M. Steinegger, B. Rost, C. Orengo, AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms, Commun. Biol. 6 (2023) 160. https://doi.org/10.1038/s42003-023-04488-9.

Journal Pre-proofs

[6] R.D. Schaeffer, J. Zhang, L.N. Kinch, J. Pei, Q. Cong, N.V. Grishin, Classification of domains in predicted structures of the human proteome, Proc. Natl. Acad. Sci. U. S. A. 120 (2023) e2214069120. https://doi.org/10.1073/pnas.2214069120.

[7] R.D. Schaeffer, J. Zhang, K.E. Medvedev, L.N. Kinch, Q. Cong, N.V. Grishin, ECOD domain classification of 48 whole proteomes from AlphaFold Structure Database using DPAM2, PLOS Comp. Bio. 20(2) (2023) e1011586. https://doi.org/10.1371/journal.pcbi.1011586

[8] S. Das, D. Lee, I. Sillitoe, N.L. Dawson, J.G. Lees, C.A. Orengo, Functional classification of CATH superfamilies: a domain-based approach for protein function annotation, Bioinforma. Oxf. Engl. 31 (2015) 3460–3467. https://doi.org/10.1093/bioinformatics/btv398.

[9] S. Das, H.M. Scholes, N. Sen, C. Orengo, CATH functional families predict functional sites in proteins, Bioinforma. Oxf. Engl. 37 (2021) 1099–1106. https://doi.org/10.1093/bioinformatics/btaa937.

[10] I. Sillitoe, N. Bordin, N. Dawson, V.P. Waman, P. Ashford, H.M. Scholes, C.S.M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S.D. Lam, K. Berka, I.H. Varekova, R. Svobodova, J. Lees, C.A. Orengo, CATH: increased structural coverage of functional space, Nucleic Acids Res. 49 (2021) D266–D273. https://doi.org/10.1093/nar/gkaa1079.

[11] I. Barrio-Hernandez, J. Yeo, J. Jänes, M. Mirdita, C.L.M. Gilchrist, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe, Nature 622 (2023) 637–645. https://doi.org/10.1038/s41586-023-06510-w.

[12] J. Durairaj, A.M. Waterhouse, T. Mets, T. Brodiazhenko, M. Abdullah, G. Studer, G. Tauriello, M. Akdel, A. Andreeva, A. Bateman, T. Tenson, V. Hauryliuk, T. Schwede, J. Pereira, Uncovering new families and folds in the natural protein universe, Nature 622 (2023) 646–653. https://doi.org/10.1038/s41586-023-06622-3.

[13] M. Van Kempen, S.S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C.L.M. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with Foldseek, Nat. Biotechnol. (2023). https://doi.org/10.1038/s41587-023-01773-0.

[14] V. Nallapareddy, N. Bordin, I. Sillitoe, M. Heinzinger, M. Littmann, V.P. Waman, N. Sen, B. Rost, C. Orengo, CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models, Bioinformatics 39 (2023) btad029. https://doi.org/10.1093/bioinformatics/btad029.

[15] J. Wells, A. Hawkins-Hooker, N. Bordin, B. Paige, C. Orengo, Chainsaw: protein domain segmentation with fully convolutional neural networks, Molecular Biology, 2023. https://doi.org/10.1101/2023.07.19.549732.

[16] K. Zhu, H. Su, Z. Peng, J. Yang, A unified approach to protein domain parsing with inter-residue distance matrix, Bioinformatics 39 (2023) btad070. https://doi.org/10.1093/bioinformatics/btad070.

[17] T.E. Lewis, I. Sillitoe, J.G. Lees, cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly, Bioinformatics 35 (2019) 1766–1767. https://doi.org/10.1093/bioinformatics/bty863.

[18] J.P.G.L.M. Rodrigues, J.M.C. Teixeira, M. Trellet, A.M.J.J. Bonvin, pdb-tools: a swiss army knife for molecular structures, F1000Research 7 (2018) 1961.

Journal Pre-proofs

https://doi.org/10.12688/f1000research.17456.1.

[19] C.A. Orengo, W.R. Taylor, SSAP: sequential structure alignment program for protein structure comparison, Methods Enzymol. 266 (1996) 617–635. https://doi.org/10.1016/s0076-6879(96)66038-8.

[20] S.R. Eddy, Accelerated Profile HMM Searches, PLoS Comput. Biol. 7 (2011) e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

[21] Y. Zhang, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309. https://doi.org/10.1093/nar/gki524.

[22] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612. https://doi.org/10.1002/jcc.20084.

[23] W.S.J. Valdar, Scoring residue conservation, Proteins 48 (2002) 227–241. https://doi.org/10.1002/prot.10146.

[24] V. Gligorijević, P.D. Renfrew, T. Kosciolek, J.K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B.C. Taylor, I.M. Fisk, H. Vlamakis, R.J. Xavier, R. Knight, K. Cho, R. Bonneau, Structure-based protein function prediction using graph convolutional networks, Nat. Commun. 12 (2021) 3168. https://doi.org/10.1038/s41467-021-23303-9.

[25] M. Kilinc, K. Jia, R.L. Jernigan, Improved global protein homolog detection with major gains in function identification, Proc. Natl. Acad. Sci. U. S. A. 120 (2023) e2211823120. https://doi.org/10.1073/pnas.2211823120.

[26] R. Krivák, D. Hoksza, P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure, J. Cheminformatics 10 (2018) 39. https://doi.org/10.1186/s13321-018-0285-8.

[27] F. Xiong, Z. Ji, Y. Liu, Y. Zhang, L. Hu, Q. Yang, Q. Qiu, L. Zhao, D. Chen, Z. Tian, X. Shang, L. Zhang, X. Wei, C. Liu, Q. Yu, M. Zhang, J. Cheng, J. Xiong, D. Li, X. Wu, H. Yuan, W. Zhang, X. Xu, Mutation in SSUH2 Causes Autosomal-Dominant Dentin Dysplasia Type I, Hum. Mutat. 38 (2017) 95–104. https://doi.org/10.1002/humu.23130.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Journal Pre-proofs

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *[Journal name]* and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proofs