Exploring structural diversity across the protein universe with The Encyclopedia of Domains

Andy M. Lau[1]†‡, Nicola Bordin[2]†, Shaun M. Kandathil[1], Ian Sillitoe[2], Vaishali P. Waman[2], Jude Wells[2,3], Christine A. Orengo[2]* and David T. Jones[1,2]*

Affiliations

[1]Department of Computer Science, University College London, London, WC1E 6BT, UK

[2]Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK

[3]Centre for Artificial Intelligence, University College London, London, WC1V 6BH, UK

*Corresponding authors. Email: c.orengo@ucl.ac.uk, d.t.jones@ucl.ac.uk

†These authors contributed equally to this work.

‡Present address: InstaDeep Ltd, 5 Merchant Square, London, W2 1AY, UK

Abstract

The AlphaFold Protein Structure Database (AFDB) contains over 200 million predicted protein structures, composed of domains: independently folding units found in multiple structural and functional contexts. Identifying domains can enable many functional and evolutionary analyses, but has remained challenging due to the sheer scale of the data. Using deep learning methods, we have detected and classified every domain in AFDB, producing The Encyclopedia of Domains. We detect nearly 365 million domains, over 100 million more than can be found by sequence methods, covering over 1 million taxa. Reassuringly, 80% of these domains are similar to known superfamilies, greatly expanding representation of their domain space. We uncover over 10,000 new structural interactions between superfamilies and thousands of new folds across the fold space continuum.

Main Text

The AlphaFold Protein Structure Database (AFDB) (*1*, *2*) is a groundbreaking initiative which significantly broadened the protein structure universe by expanding 3D representation to over 200 million UniProt sequences. The implications of the AFDB have been profound, not only in academic research in the life sciences but also in the commercial sphere, where the integration of novel data and the advanced technologies used to generate accurate structures is being explored for next-generation structure-based drug discovery (*3*).

Notwithstanding its revolutionary impact, the AFDB is not without its limitations and presents a new set of challenges. The sheer scale of the data makes many traditional tools and pipelines, originally designed for considerably smaller datasets, inadequate for navigating the extensive number of structures and sequences within. This necessitates an evolved strategy, warranting a new perspective on how best to represent and traverse the data and complex relationships within such an expansive database, as well as requiring synergy between new algorithmic methods and computational hardware. Recent studies explored AFDB by partitioning full-length AFDB models into clusters of structurally similar proteins, as well as characterising their functions (*4*, *5*). At a more granular level, Bordin et al. (*6*) and Schaeffer et al. (*7*) interrogated the composition of specific proteomes (the initial AFDB 21 model organism dataset, and 48 proteomes, respectively), cataloguing domains under the CATH (*8*, *9*) and ECOD (*10*) frameworks.

Domain discovery is possible via sequence- and structure-based approaches, with Pfam (*11*, *12*) and Gene3D (*13*) as prime examples of the former. The Pfam database describes collections of protein families, each represented by a multiple sequence alignment (MSA) and a hidden Markov model (HMM) profile (*11*, *12*). In Gene3D, sequences of existing CATH superfamilies, assigned via structure, are leveraged to discover new domains in sequence space via representative profile HMMs (*13*). Sequence-based discovery allows for greater coverage but is limited by HMMs detection capabilities, often failing on remote relatives. Structure-based assignments allow for higher quality domain boundaries and can reveal very remote relatives but have been limited by the low numbers of experimental structures. The task is further complicated by conceptual difficulties when performing structural comparisons, such as the lack of a statistical null model (compared to the use of random sequences for sequence comparisons).

An essential perspective yet to be explored in the context of the AFDB is the comprehensive mapping and analysis of protein domains across various branches of the Tree of Life using structural data. The relationships between protein folds and domains have been extensively highlighted by structural databases such as CATH (*8*, *14*, *15*), ECOD (*10*), SCOP (*16*) and SCOPe (*17*), with differences in criteria for which regions are assigned to domains in proteins. Although early comparisons across domain databases showed agreement for fold level assignments (*10*, *18*, *19*), significant differences exist for the definition of a protein domain. CATH recognises that some structures can be decomposed into further structural units, each with their own repertoire of observed variations, while SCOP takes into account the idea of

fold recurrence - whether a subunit has been observed as reoccuring in another family or only as a single domain (*18*).

Examination of the AFDB through the lens of the CATH framework holds the potential to reveal unprecedented insights that can illuminate subtle yet important connections between structure and function across large swathes of organisms. As such, a structure-based mapping of domains within the AFDB promises not only to massively facilitate the exploration of these relationships, but also provide a foundation for further annotation.

In this study, we present a comprehensive analysis of domain composition for the entirety of the AFDB (version 4). This description encompasses over 364 million putative domains, derived from more than 214 million protein sequences across more than 1 million taxa. The identification of these domain structures is made via the consensus of three automated parsing methodologies: Merizo (*20*), Chainsaw (*21*) and UniDoc (*22*). Further, by employing structural comparison methods such as Foldseek (*23*) and an in-house deep learning method called Merizo-search, more than 251 million domains can be placed on the CATH hierarchy.

A high-throughput strategy for identifying structural domains in the AFDB

Our workflow combines three state-of-the-art domain parsing methods together with structure classification algorithms to identify known domain folds (Fig. 1a-b; Methods) within the AFDB. Using this workflow, we identified a total of 364 million 'TED' domains across the AFDB (Fig. S2, Table S1) - 100 million more domains than found via sequence-based methods (Fig. 1c). TED-100 describes a roughly 42:55 division between single and multidomain proteins (Fig. 1d-i), with the latter up to 20 domains in composition (Fig. S3). Only 2.8% of targets (5.3 million) in TED-100 lack identifiable domains, compared to 33.9% (64.1 million targets) in Gene3D and 26.2% in Pfam (49.4 million targets) (Fig. S4). In TED, these targets either consist entirely of non-domain residues (NDR) or lack any consensus among the three domain segmentation methods employed. Across superkingdoms, the fraction of NDRs varies, with approximately 10% identified across archaeal and bacterial lineages and up to 30% in eukaryotes (Fig. S5).

Analysis of the average predicted local distance difference test (plDDT) scores for TED-100 reveals that the majority of domains fall within the "very high" to "high" bins, with only approximately 2% within the lowest bin (Fig. 1d-ii). Since our domain segmentation methods do not consider residue plDDT when determining domains, this suggests that our domain finding pipeline effectively identifies plausible domains within the well-folded areas of AF2 models.

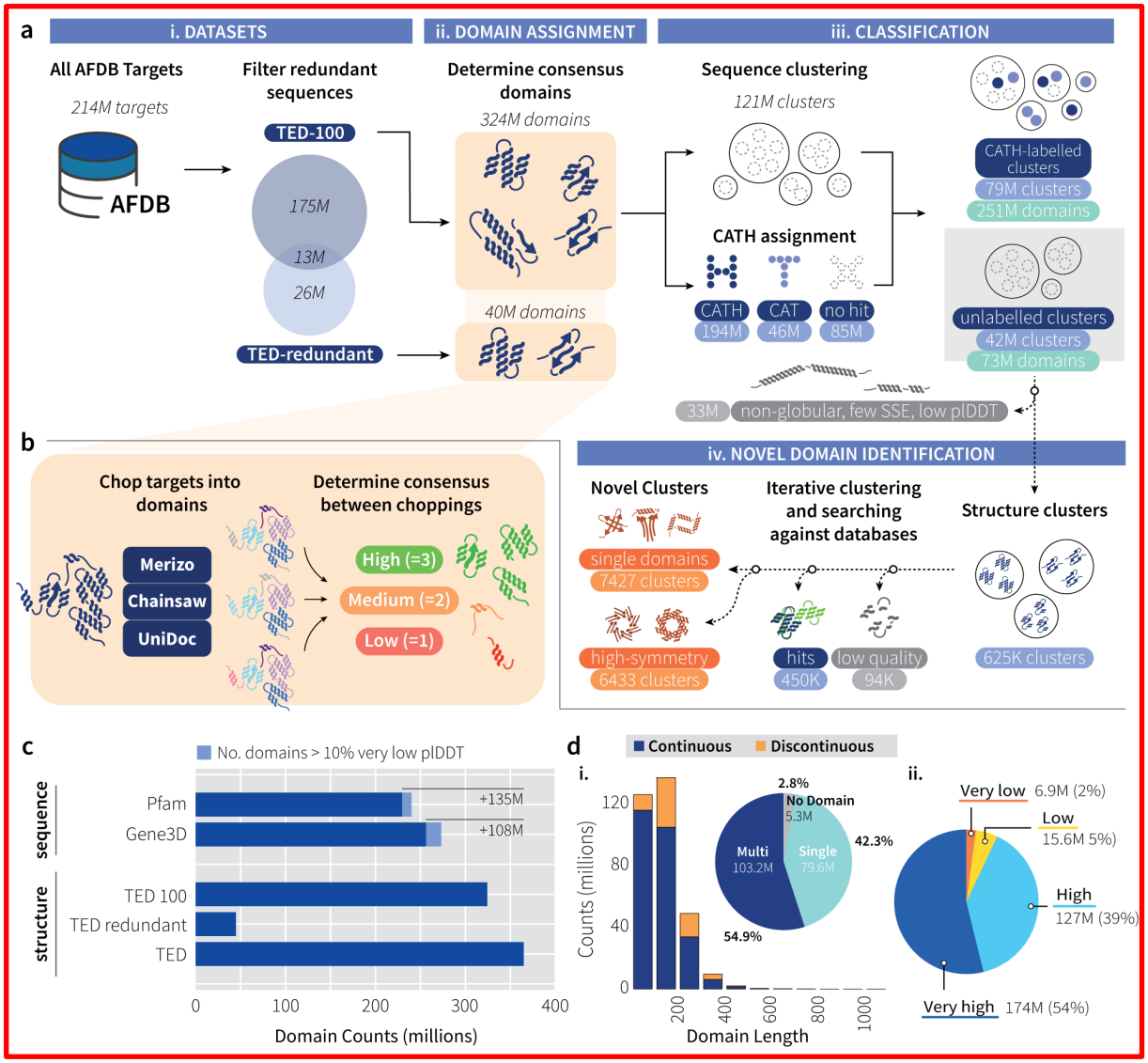

Fig. 1. Overall workflow. (**a**) i. 214 million AFDB target sequences are filtered by 100% sequence identity in order to avoid bias. This identifies 188 million non-redundant targets (TED-100) and a set of sequence-redundant targets (TED-redundant). ii. Both TED-100 and TED-redundant undergo automated domain parsing, with assignments derived from consensus among the three methods. iii. TED-100 domains are processed by MMseqs2, creating over 121 million clusters at 50% identity. Concurrently, domains are matched to CATH domains via Foldseek and Merizo-search, categorised into superfamily (C.A.T.H), topology (C.A.T), or no-matches. Domains found by Merizo-search nearest neighbour matches are considered as topology-level matches. Clusters are annotated with CATH labels, creating partially labelled and unlabelled clusters. Low-quality domains in unlabeled clusters are filtered out. iv. Resultant domains undergo a new workflow for identification, involving clustering and database searches for matches to known structures. Poor quality domains (non-protein-like) are identified using an in-house deep learning method (Methods). Novel domains are additionally scored on internal symmetry using the SymD program (Methods). (**b**) Full-length targets are subjected to automated domain parsing by Merizo, Chainsaw and UniDoc. A consensus is taken by identifying assignments where three (high), two (medium) methods agree or no consensus is found (low). Only high and medium consensus domains are analysed further. (**c**) Comparison of domains identified by sequence (Pfam and Gene3D) versus structure-based methods (TED). The "TED" count combines TED-100 and TED-redundant. (**d**) i. Domain length distribution and proportion of identified continuous (blue) and discontinuous (orange) domains. Inset shows proportion of single, multi-domain and number of targets with no identified domains (n=188,914,411). ii. Average plDDT distribution for TED-100 domains (n=324,389,697) across confidence bins: dark blue/very high (plDDT >= 90), blue/high (90 > plDDT >= 70), yellow/low (70 > plDDT >= 50), and orange/very low (plDDT < 50).

Classification of TED domains into the CATH hierarchy

The 324 million domains in TED-100 were clustered by sequence using MMseqs2 (*24*) and compared against CATH representative domains using fast structure searching methods (Fig. 1a-iii, Methods). This produced approximately 121 million clusters at 50% sequence identity and a minimum coverage of 90% (Fig. S6; Methods). The majority of these clusters comprise just singleton sequences (roughly 81 million), with the largest non-singleton cluster comprising 12,847 domains.

Parallel to sequence clustering, we use a combination of Foldseek (*23*) and Merizo-search (*25*) (an in-house structure search method using domain coordinate embedding) to search all TED-100 domains against CATH SSG5 domains (*9*) (Methods), allowing 194 million domains to be assigned with CATH superfamily (H) labels and 46 million at the topology (T) level (Methods). These labels were further validated by scanning domain sequences against an updated library of HMMs for CATH PDB domains. Approximately 171 million superfamily predictions by Foldseek could be confirmed with exact HMM superfamily matches (88.54%), with an additional 1.8 million domains (0.95%) confirmable at the fold level. Only 4.1 million of the 16 million Foldseek predictions for fold matches on CATH can be validated by HMM scans (25.8%), with 11.8 million fold predictions and 20.3 million superfamily predictions by Foldseek not confirmed by an HMM match, suggesting an expansion by 15.4% in CATH labelled domain coverage using AFDB structures over HMM-based sequence assignments.

By identifying sequence clusters with any CATH label assigned domains, the clusters were partitioned into two categories: 78 million clusters (over 251 million domains, nearly 80% of all TED-100 domains) which contain at least one CATH-labelled member, spanning 148 million proteins, and 26 million proteins having no domain annotations in Pfam and 30 million having none in Gene3D (Fig. S7). The remaining 41 million clusters have no members with CATH labels (approximately 73 million domains; Fig. S6). The absence of similarity to CATH domains in the latter clusters could be attributed to them being novel folds, extremely divergent relatives of existing CATH domains or simply being incorrect models, making them unmatchable to known folds.

Enrichment of Fold Representation by TED

To develop an understanding of how folds are distributed across the AFDB, we assessed the composition of TED using the CATH hierarchy. Fig. 2a shows the top 100 CATH superfamilies of each class (alpha, beta and alpha/beta), which are greatly enriched in TED-100 compared to baseline sequence hits in Gene3D.

These folds and architectures are unevenly distributed across the Tree of Life, however the majority of folds (61%) are shared and reused across all superkingdoms, suggesting essential roles for cellular life. Some folds were found across two superkingdoms (18.5%), while others

are more exclusive, with 0.5%, 9% and 11% of CATH folds found only in archaea, eukarya and bacteria, respectively.

The most abundant superfamilies in TED-100 and CATH are shown in Fig. S8 and Fig. S9. Directly comparing the top superfamilies assigned in TED compared to CATH, in terms of raw domain counts, sees the promotion of several superfamilies into the top 5 of each class, including the MFS general substrate transporter-like domain, translation factors and the FAD/NAD(P)-binding domain (Fig. S8). The set of superfamilies highly enriched in TED include those associated with the archetypal multi-drug efflux pump AcrB. AcrB is a Resistance-Nodulation-Division (RND) transporter and forms part of the AcrAB-TolC efflux pump in bacteria where it is responsible for the export of harmful substances such as antibiotics, contributing to antibiotic resistance (*26*). The constituent domains of AcrB, including the pore domain, transmembrane domain and TolC docking domain, are found greatly enriched in TED compared to the PDB, providing up to nearly a thousand-fold increase in representation for these biologically important superfamilies.

The pore domain, which forms part of the pore selectivity filter in RND transporters, is principally found in bacterial species and in a small minority of archaeal and eukaryotic lineages based on Gene3D hits. However, structure-based searching via TED expands the coverage of the pore domain superfamily into an additional 18 archaeal, 1315 bacterial and 284 eukaryotic lineages unique to TED, which evaded even HMM searches. This broader coverage of organisms, revealed by structural comparisons, may reveal potential evolutionary events, such as lateral gene transfer between bacteria and eukaryotes. This expansion is exemplified in Fig. 2c, where we show that over 1000 CATH superfamilies, described by Gene3D sequence matches as occurring solely in a single superkingdom, could actually be mapped to other branches when structure was taken into account in TED.

TED also enables us to better study the distribution of folds across different branches of the tree of life. Among the 193 million TED domains with superfamily labels, we observed folds that were exclusively localised to specific superkingdoms. These findings are shown in Fig. 2d, where folds are summarised at the CATH topology level, visualised by principal component analysis (PCA), and where points at the vertices represent exclusive occurrences of a topology and all its superfamilies, in a particular superkingdom.

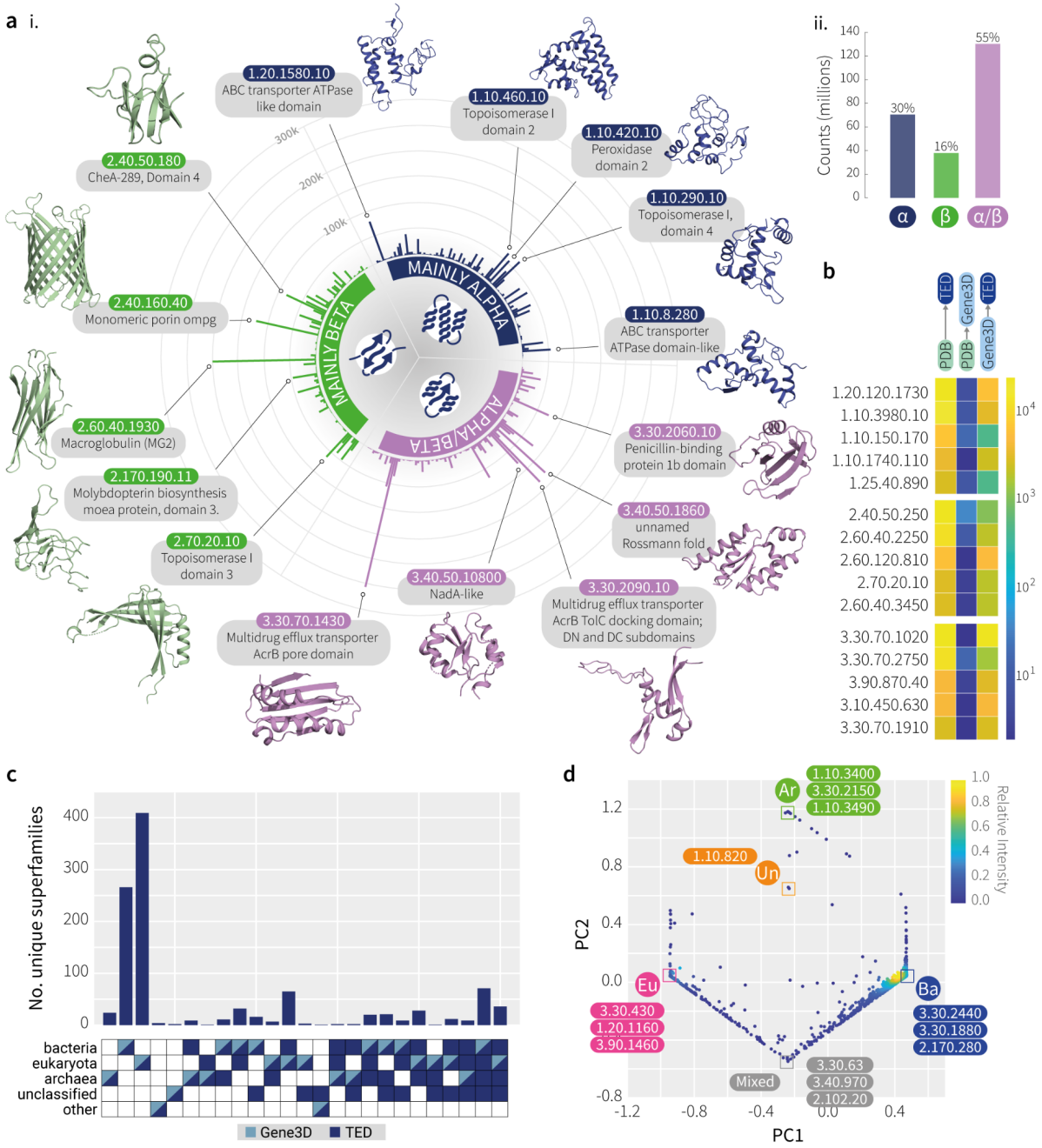

Fig. 2. Classification of TED domains using the CATH hierarchy. (a) i. The top 100 superfamilies in TED-100 for each CATH class where more matches to CATH superfamilies have been identified via structural hits in TED, compared to sequence hits in Gene3D. ii. Proportion of domains matched to CATH classes (n=238,569,631). **(b)** Enrichment of superfamily representation in TED-100 compared to PDB and Gene3D. The top 5 superfamilies of each CATH class are shown, where enrichment in TED-100 compared to PDB is the greatest. Colour scale represents fold-change in superfamily representation in PDB and Gene3D compared to Gene3D and TED. A full list of fold names corresponding to the CATH superfamily codes can be found in Table S2. (**c**) Expansion of CATH superfamilies to new superkingdoms in TED. Plot shows the number of unique superfamilies found in each superkingdom (across the 653,460 taxa of TED-100) according to Gene3D and TED assignments. Each column along the horizontal axis depicts the number of superfamilies that are exclusive to a single superkingdom when only considering Gene3D assignments, but are expanded into one, two or three additional superkingdoms in TED. Only superfamilies where Gene3D domains are exclusive to a single superkingdom are shown (n=1061). **(d)** Exclusivity of CATH topologies across superkingdoms. PCA of normalised CATH topology counts across five superkingdoms: eukarya (Eu), bacteria (Ba), archaea (Ar),

unclassified and other sequences (Un). The 'mixed' category comprises topologies found in roughly equal proportions in Eu/Ba domains. Examples of superkingdom-exclusive topologies are shown for each category.

Novel high-symmetry architectures

From our TED workflow (Fig. 1), we identified 41 million sequence clusters which could not be linked to CATH superfamilies. Representatives of these clusters were subjected to a workflow aimed at identifying novel domain folds (Fig. 1a-iv; Methods).

While reviewing these clusters, it was apparent that we would need to treat repeat architectures with high internal symmetry separately. A good example of domains in this class are the various WD40 beta propellers, which are considered distinct domain architectures in their own right, but clearly comprise repeats of domain-like units. To identify similar domains in our workflow, we calculated Z-scores using the SymD program (*27*), sequestering any cluster representatives with a symmetry Z-score of greater than 9, into a new category of 6,433 highly symmetric novel fold clusters (Fig. S10).

Within these clusters, we find new architectures such as an 11-bladed beta-propeller, a closed alpha ring-like 11-helix propeller and 6-helix propeller which have not been seen before (Fig. 3). Various other propeller arrangements including the "beta-flower" domain shown by Durairaj et al. (*4*), were also identified, with some visually striking examples shown in Fig. 3.

More curiously, we find a broad category of architectures which are composed of cyclic repeats, extruded along an axis to form highly repetitive and symmetric structures which we call "extruded repeats". Several studies (*28*, *29*) and databases such as the Database of Structural Repeats in Proteins (DbStRiPs) (*30*) have curated repeat architectures from the PDB, including alpha- and beta-solenoids, and horseshoes formed of alpha-repeats such as HEAT, ankyrin and armadillo. Fig. 3 showcases some complex examples of these structures, with some featuring highly varied and unstructured loops between repeating units (Fig. S11). Many of these domains architectures resemble other solenoid folds that were recently reported in a systematic study of beta-solenoid fold space (*29*).

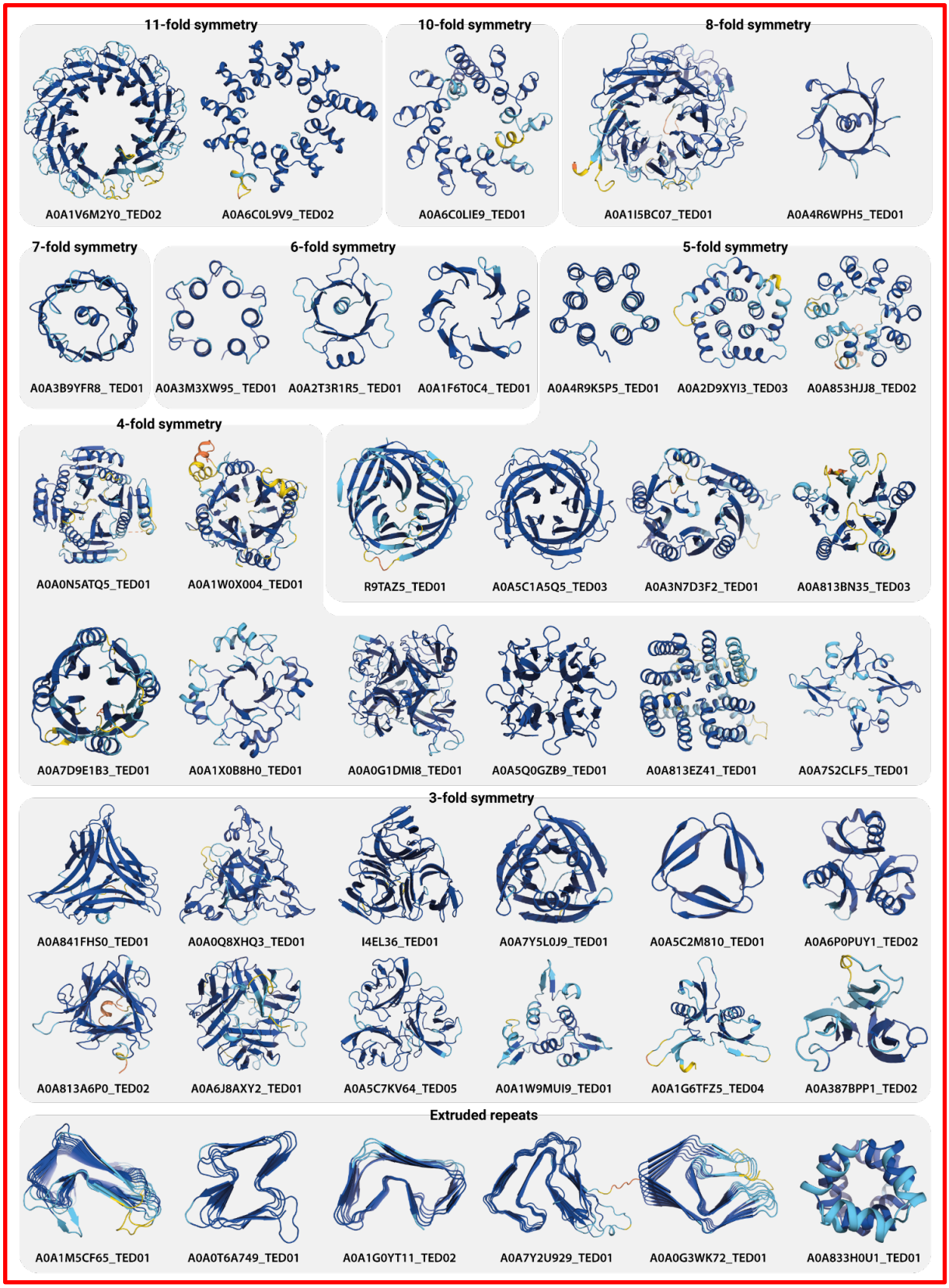

Fig. 3. Examples of high-symmetry domains and extruded repeats. Domains are identified as part of the novel domain identification pipeline and are identified as domains with high internal symmetry via scoring with the SymD program (Methods). Extruded repeats are domains with a high number of ordered cyclical repeats projecting along one axis. Colouration follows plDDT confidence bins as per the AFDB (dark blue/very high: plDDT >= 90, blue/high: 90 > plDDT >= 70, yellow/low: 70 > plDDT >= 50 and orange/very low: plDDT < 50).

Novel Domains and their distribution across the Tree of Life

The remaining low-symmetry clusters were assessed on domain quality, using a variant of the Foldclass network (*25*) trained to identify poor quality domain choppings (Supp. Methods), and novelty by further matching against known structure libraries. The final output of our workflow produced 7,427 clusters of putative domains which appear to be well-folded but dissimilar to any known domain fold. Although there is no exact boundary between novel and just highly divergent examples of known folds, by applying a density-based anomaly detection algorithm (Methods), we could at least rank these domains by novelty relative to known domains.

Fig. 4 showcases examples of novel domains identified by our workflow. Over a quarter of clusters corresponded to singletons at the sequence cluster level (1930 domains), which are uniformly distributed in terms of the Foldclass novelty score. Most novel domains identified are from bacterial proteins, which is unsurprising given that the latter comprise most proteins in the AFDB and in our TED-100 dataset. Overall, these domains are distributed evenly across different phyla, but several bacterial phyla were found overrepresented when compared to baseline counts across all domains in TED-100, specifically in the PVC group, Myxococcota, Spirochaetota, Bdellovibrionota, Nitrospinae/Tectomicrobia group and Calditrichota (Fig. S12), suggesting that species within these phyla may be underrepresented in terms of domain coverage in the current PDB.

Ranked first by novelty, is a curious archaeal domain found as a sequence singleton from *Candiatus Poseidoniales archaeon* (TED: A0A7M3WA57_TED05; novelty score: 77.2; Fig. 4c-i). This protein is not documented in InterPro and no GO terms are available. The structure is composed of paired beta-strands, in a closed, twisted hairpin with both termini adjacent to one another. Reviewing the context of the full-chain from the AFDB shows that the domain forms part of an extended loop, protruding from the middle of a immunoglobulin-like domain (TED: A0A7M3WA57_TED04). The hairpin topology of A0A7M3WA57_TED05 is mirrored by another identified novel domain represented by E1Z635_TED02 shown in Fig. 4d-ii, but is alpha-helical in nature. This domain is found only in eukaryotes, primarily in species belonging to the Viridiplantae phyla (97 species) but also in a minority of Opisthokonta (14 species) and Amoebozoa (1 species), suggesting an evolutionary link between these lineages.

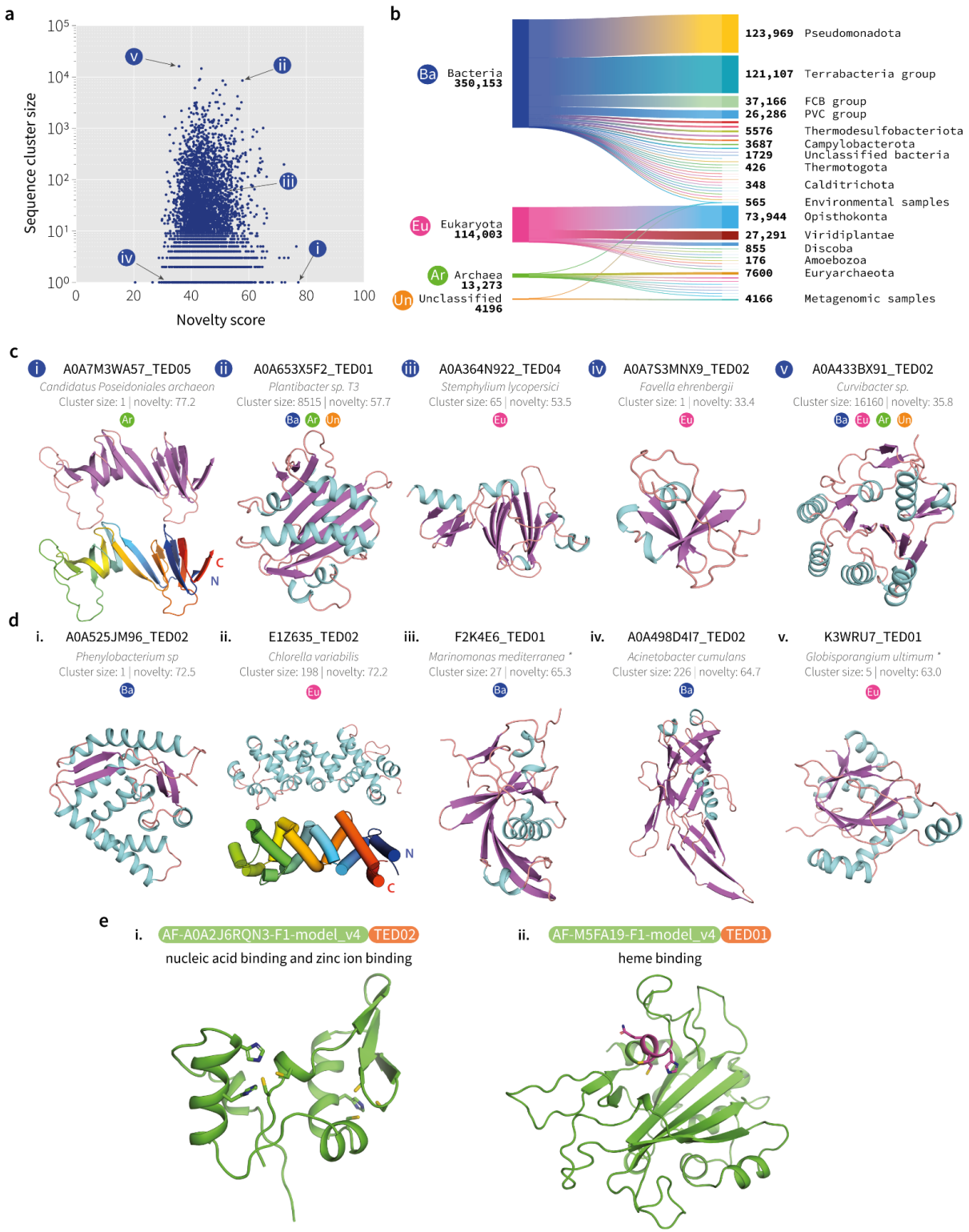

Fig. 4. Examples of novel domain clusters identified in TED. (**a**) Comparison of domain novelty score versus sequence cluster size (n=7427). Novelty scores are predicted by the Foldclass algorithm where novel domains are ranked with a score close to 100. (**b**) Taxonomic distribution of novel domain clusters (for all sequence cluster members; n=483,732). Largest common phyla are shown across superkingdoms along with the number of domains in sequence clusters assigned to each level of the hierarchy. (**c**) Subpanels i-v correspond to labels shown in panel (a). In panel i, the bottom sub-panel shows the arrangement of strands that form the coiled hairpin loop from the N-terminus (blue) to C-terminus (red). The quoted cluster size represents the number of identified homologues at the sequence cluster level. Labels denoting superkingdoms correspond to panel (b) and represent the

superkingdom that all cluster members belong to. The cluster is distributed across multiple superkingdoms when multiple labels are shown. (**d**) Examples of high-novelty structures. In panel ii, the bottom sub-panel shows the arrangement of helices that form the coiled hairpin loop from the N-terminus (blue) to C-terminus (red). Asterisks denote where organism names have been shortened: iii. *Marinomonas mediterranea* (strain ATCC 700492 / JCM 21426 / NBRC 103028 / MMB-1), v. *Globisporangium ultimum* (strain ATCC 200006 / CBS 805.95 / DAOM BR144) (*Pythium ultimum*). (**e**) Novel folds with predicted functions. i. Example of a domain predicted to have nucleic acid and zinc binding properties. Potential zinc binding site residues are highlighted as sticks. The left-hand site is composed of 2 Cys and 2 His residues, whereas the right-hand site has 3 Cys and 1 His in a tetrahedral arrangement. ii. Example of a heme binding domain. The residues of the heme *c* binding motif are highlighted.

Sequence-based function prediction for novel fold and repeat domains

To see if any functions could be assigned to the domains with novel folds and repeats, we use a sequence-based deep learning model (Supp. Methods) to predict Gene Ontology (GO) terms. This analysis shows that 1321/7427 (18%) of the domains in the putative novel fold set, and 1419/6433 (22%) of the repeat set can be assigned high confidence ($p < 10^{-4}$) Molecular Function GO term labels. The top 20 GO terms predicted for the two sets of domains are shown in Table S4 and Table S5. Manual inspection of the domains predicted to have zinc binding and nucleic acid binding functions reveals that many of the domains contain plausible zinc binding sites (*31*), most containing 2 Cys and 2 His residues arranged in tetrahedral fashion, including as part of zinc finger-like supersecondary structure motifs. Fig. 4e-i shows one example containing two zinc binding sites and a possible nucleic acid binding alpha-helix, but which lacks the canonical zinc finger supersecondary motif. We also consider the set of domains predicted to have heme binding properties and find that most of these contain the canonical heme *c* binding motif (CXXCH) (*32*). Inspection of the three-dimensional structures reveals that each of these domains has one or more heme binding sites in a plausible conformation; one example is shown in Fig. 4e-ii, with the residues of the heme *c* binding motif highlighted. The His residue that binds the heme iron is in a conformation compatible with placement of the heme group in the pocket, which is primarily hydrophobic. The presence of clear sequence motifs and structural features consistent with the assigned functions suggests that these novel domains may indeed have the predicted functions, and further work is needed to validate the remainder of the predicted functions for these novel domains.

Novel Interactions Between Domain Pairs

Unlike sequence-based domain annotations such as Gene3D, the availability of full-chain AF2 models for multidomain proteins allows us the unique opportunity to interrogate and compare domain pair packing interactions in TED and CATH. TED contains a total of 27,280,057 instances of interacting domains categorised into 13,771 Interacting Superfamily Pairs (ISPs). In contrast, the set of interacting domains in CATH consists of just 196,234 instances across 5,111 ISPs. The relative enrichment in the number of instances of ISPs common to CATH and TED is shown in Fig. 5a, which shows that most of these ISPs have many more members in TED. Assessing the diversity of interaction geometry for ISPs common to TED and CATH (Methods and Fig. 5b-i-ii), we find that for most ISPs, the diversity of interaction geometries

in TED is consistent with that seen in CATH, indicating that on average, AF2 tends to recapitulate inter-domain geometries already seen in the PDB.

A small proportion (5.4%) of these ISPs show enhanced diversity in interaction geometries in TED, as measured by an increase in Conservation of Interaction Orientation (CIO) score (Methods) of 0.3 or more. A smaller proportion of ISPs (2.3%) are more diverse in CATH. We find no strong dependence of difference in CIO value on whether the ISP in question is homotypic (2 copies of domains from the same superfamily) or heterotypic (Welch two-sample *t*-test, p=0.141). Most of the interacting superfamily pairs (ISPs) in the TED set (10,701 out of 13,771) are unique to TED, i.e. they are not observed in CATH, and CATH contains 2,041 ISPs not seen in TED (Fig. S13).

That some interactions in CATH are not captured in TED is not entirely surprising, firstly due to the possibility of domain parsing and classification errors. Secondly, TED excludes any sequence not modelled in the AFDB; these include sequences from viruses (around 5,304,757 sequences), and longer sequences, which we estimate contain between 50 and 61 million domains (Supplementary Methods). The AFDB will also not contain experimental constructs comprising domain combinations not observed in natural sequences. Lastly, instances of ISPs in TED are filtered based on whether the constituent domains are in contact in the full-length AF2 structure, and whether they have a favourable inter-domain Predicted Aligned Error (PAE; Methods), both of which depend on the availability of homologous sequences and template structures that ideally span both domains. For these reasons, we do not expect that TED ISPs will be a strict superset of CATH ones. Nevertheless, we find that TED essentially doubles the set of known domain interactions at the superfamily level (10,701 novel TED interactions versus 5,111 known interactions in CATH).

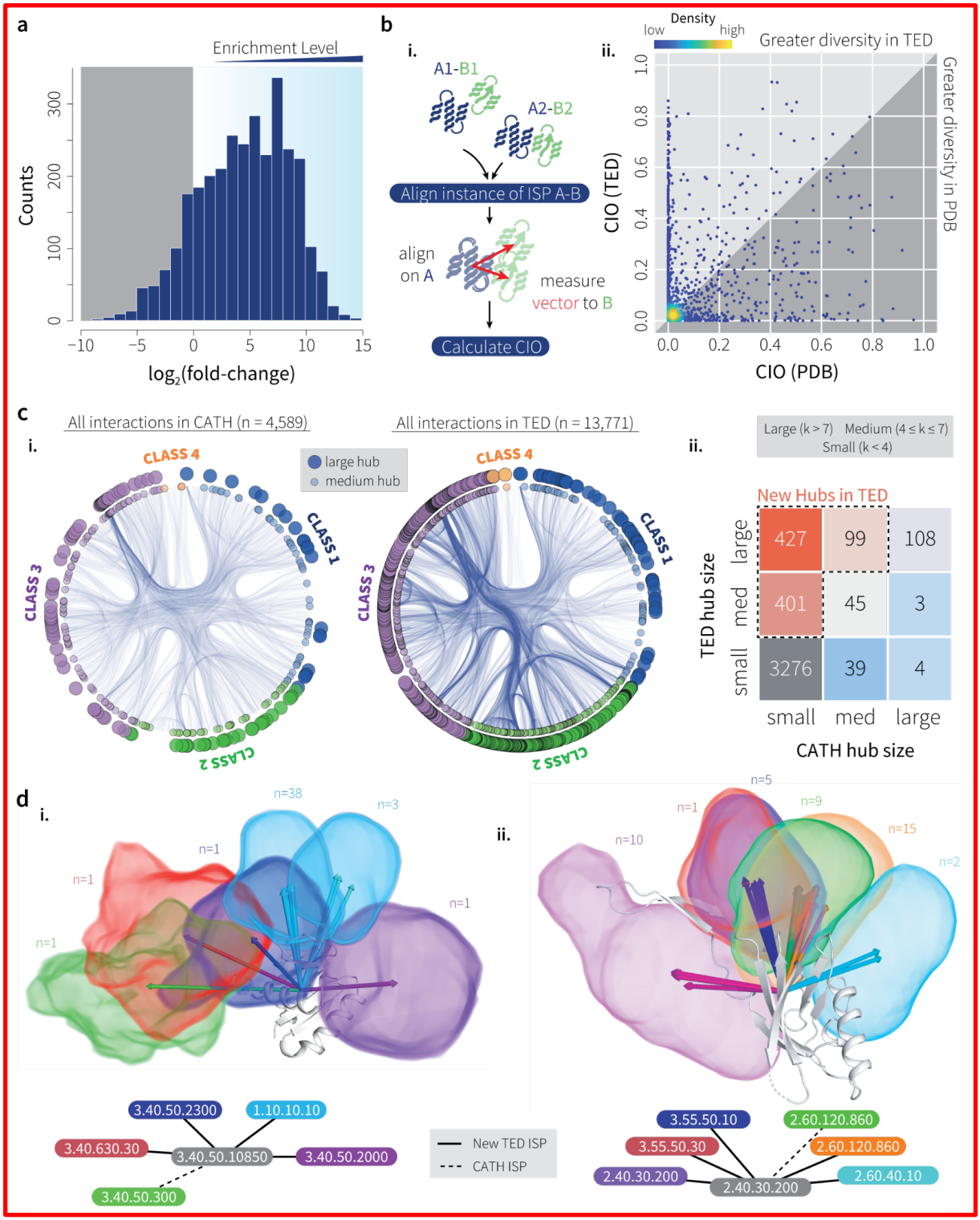

Fig. 5. Interacting superfamily pairs (ISPs). (**a**) Enrichment of the number of instances of ISPs common to the CATH and TED datasets, expressed as $\log_2$(fold change) (n=3070). (**b**) i. Alignment procedure used to compute CIO values for an ISP. One domain in each instance of each ISP is used as a reference and aligned to a designated 'master' reference domain structure. The rotation and translation from each alignment is applied to the second, 'tag-along' domain to bring all domain pairs into a common frame. Vectors are then computed between the centres of mass of each pair of domains, and used to compute the CIO measure (see Methods). ii. Comparison of CIO values for ISPs common to CATH and TED. Most ISPs show a high degree of conservation in interaction patterns. (**c**) i. Hierarchical edge bundling plots illustrating differences in domain superfamily interaction patterns between CATH (left) and TED (right). Curves in the plots connect interacting superfamilies. Hubs are marked by medium

(4-7 connections) and large circles (>7 connections) on the outer rim. ii. Comparison of hub domains in CATH and TED. The heatmap compares CATH superfamilies in CATH and TED as hubs, categorised as small (<4 connections), medium (5–7), and large (≥8). Hub thresholds used are from Ekman et al. (*33*). (**d**) Two examples of new hub superfamilies in TED, with groups of domains for interacting superfamilies placed in a common frame and represented as volumes, alongside chains involved in each group and a graph representation of each hub. The sets of interactions for superfamily 3.40.50.10850 (NtrC-like protein domain) are disjoint between CATH and TED, whereas the set of TED interactions for superfamily 2.40.30.200 (Distal tail protein domain) include that seen in CATH (orange and green). A decomposed view of d.i. appears in Fig. S15.

A visual illustration of ISP sets can be seen in Fig. 5c-i, in which a path is drawn between two superfamilies if at least one interaction is observed between domains assigned to those superfamilies. This representation uses the CATH hierarchy as a guide to 'bundle' paths drawn between superfamilies in related parts of the CATH hierarchy (Methods; Fig. S13), and shows visually that a very large number of new interactions are seen, especially between superfamilies in CATH classes 2 and 3 (all-beta and alpha-beta classes, respectively). As mentioned above, the majority of the interactions seen in the TED set are unique to TED, and a visual comparison of the set of TED-unique interactions to all known interactions in CATH (Fig. S13) reveals a huge expansion in the set of known interactions. The network of superfamily interactions also allows us to identify superfamilies that can be considered hubs on account of their ability to interact with many other superfamilies. As shown in Fig. 5c-i-ii, a large number of superfamilies in CATH have their hub status 'promoted' due to new interactions seen in TED. In Fig. 5d-i-ii we show two examples of superfamilies that have been promoted in this way. Notably, the set of interactions observed in TED for these superfamilies show that they can contain only novel (in the case of 3.40.50.10850) or, in some cases, a mixture of novel and already-observed interaction modes (for 2.40.30.200).

The interaction data shows that TED greatly expands the set of known interactions and interaction modes between domains. Future work will aim to delve deeper into the interaction data and determine the functional roles that these interactions (particularly those not seen before) might play in cellular processes, in particular, taking into account the multi-domain architecture of individual proteins and their evolutionary relationships.

Structures of redundant sequences in the AFDB

Of the 214 million structures in the AFDB, nearly 39 million are exact sequence duplicates of other proteins in the database (13 million unique sequences within this set). The 26 million redundant proteins comprise our TED-redundant set (Methods).

Remarkably, the AFDB models for these sequence-redundant proteins often diverge from one another, with approximately 42% of clusters (5.6 million) showing a maximum cluster RMSD of greater than 1Å (Fig. S16). The very largest RMSDs in the distribution tend to relate to changes in domain packing, but even at the domain fold level, changes can be observed. Fig. S16 shows the distribution of maximum pairwise RMSD for each cluster of identical

sequences. We found structural variation at the chain-level (Fig. S16a), as well as in the PAE maps generated by AF2 (Fig. S16a-ii).
One explanation for our observations here is that we are looking at alternative conformers of the protein chains relating to different MSAs. However, in these cases, the input MSAs should be identical given the described modelling protocol. The larger differences are also far too large to attribute to the short relaxation step that follows the model prediction step. This is most evident in the example shown in Fig. S16a-ii, where two AFDB models for identical sequences deviate by nearly 65Å and show clear differences in the PAE map.

To further investigate structural diversity within these sequence-redundant clusters, we subjected the TED-redundant targets to our domain parsing workflow, identifying approximately 40 million domains across the set. In addition, we also derived a consensus across all structures in each identical sequence cluster (Methods). This allowed us to investigate domain-level changes in conformation between identical sequences, and identified many cases where the consensus domains were dramatically different (Fig. S16b).

Discussion and Conclusions

What we have shown so far in developing TED is a way in which structural data in the AFDB can be augmented, by carefully breaking down structures into their component domains, allowing them to be classified through the CATH framework. This initiative not only drives forward the associations that we can make between structure and function, but as shown in our study, can be used to discover and reclaim the dark areas of fold space that are not accessible to sequence-based discovery.

Comparing TED with a recent study on the 21 model organisms dataset (*6*) (released prior to the 214 million release), already shows that the TED workflow identifies not only a greater number of domains, but that the domains are also of higher quality and capture many more remote homologies (Table S3). TED currently annotates domains for over 1 million taxa, of which 600,000 are currently mapped to CATH domains in TED-100, and such an extended mapping of domains to superfamilies will enable many more evolutionary discoveries. A good example of how such expansion of CATH superfamilies enhances understanding of evolutionary processes and aids inheritance of functional properties, even towards drug repurposing is shown in Fig. S17.

The coverage of TED dwarfs sequence-based assignment methods (namely Gene3D and Pfam), identifying over 100 million more domains in the AFDB compared to the latter. The proportion of proteins that each method can find domains within is shown in Fig. S4, which illustrates the advantage of considering structural domains (TED finds domains in 40-50 million proteins that Gene3D and Pfam cannot). Interestingly, prior studies based purely on sequence comparisons have suggested that between 40-65% of prokaryotic proteins are composed of multiple domains, with a higher proportion proposed in eukaryotes (*34–36*).

These values are comparable to those seen in TED, where we find a roughly 42:55% split between single and multidomain proteins in the AFDB (Fig. S4). Compared to TED, the proportion of multidomain proteins is much lower in Gene3D (29%) and Pfam (24%) assignments (Fig. S4).

Although most structures in the AFDB are undeniably of high quality, the sheer scale of the data means that errors and anomalies are inevitable, and these should be discussed in order to assess the overall robustness of the data. Some limitations have already been pointed out by other studies (*37*, *38*). One such idiosyncrasy that appeared during our development of TED was the observation that models of 100% identical sequences were sometimes dramatically different. To reframe the issue, the implication of this is that for a given protein, a user may be able to find a better model or domain within the AFDB (and one that AF2 clearly places higher confidence in) if sequence-redundant copies are considered. This may mean that some alternative structures could only be detected when redundant copies are available to compare against. Given that the vast majority of AFDB entries (175/214 million) do not have duplicates, the implication here is that an unknown (but probably large) number of low-quality structures or regions in AFDB might be improvable by resampling the input MSAs and reassessing the quality of the models. Even though this might be too heavy a task to do across the whole of AFDB, users of the database should certainly bear this in mind when looking at specific domains of interest.

The obvious explanation for the existence of divergent models for a given sequence must be that the models were generated using significantly different MSA information, and we note that explicit pre-sampling of the MSAs given to AF2 has been explored recently as a way to persuade the network to generate alternate conformations (*39*, *40*). As the structure generation pipeline in AF2 includes possible random MSA resampling during the forward pass, it could be that a borderline MSA could result in very divergent predictions on different runs due to this resampling (fixed random seeds were not used in making the AFDB). Another possibility is that the sequence data banks used for different instances of the same target sequence were inadvertently updated to newer versions if they were predicted at later times. We were not able to investigate these possibilities further ourselves as there is no information on the MSAs provided in the AFDB data, but we do suggest that it is a topic worthy of further investigation by the AFDB developers.

Fortunately, we find many examples that reiterate the notion that plDDT is an invaluable discriminator of the confidence that a user should place on a model. Fig. S18 shows an example whereby an AFDB entry has been generated within the very low confidence bin. This model features a significant number of structural defects not typical of AF2 models. “Re-folding” the sequence in ColabFold (*41*) produces a number of visually striking structures across the five AF2 models, the variation of which, along with their unusually low plDDT scores, strongly indicates that AF2 has hallucinated these folds and that they should likely be disregarded.

However, overall, the large proportion of domains mapping to CATH evolutionary families gives confidence in the quality of the models, capturing well the structural features of folds and

preserving their distinctive structural characteristics. In this context, the performance of the domain segmentation algorithms deployed here has also been key, as early pilot work on the 21 model organisms suggested a much higher proportion of problematic AF2 models largely caused by the poor segmentation of full-length proteins into domains by the sequence-based methods. As well as confirming earlier hypotheses that the majority of domain structural families had already been characterised experimentally (*42*, *43*), our study has also revealed some intriguing and beautiful new domain architectures and folds, especially some highly symmetric repetitive structures. Future development of TED will aim to comprehensively analyse these new folds and incorporate them into the CATH hierarchy, which will require extensive manual curation. Additionally, we recognize the need to optimise the current TED workflow for better detection of repeat structures, such as the different extruded repeats highlighted in our work. This optimisation will necessitate the development and integration of new tools and analysis methods specifically designed to identify these types of structures.

Throughout our study, we had to make a number of algorithmic decisions which were primarily motivated by the monumental amount of data we had to process in the AFDB. As such, we intend for TED to be an ongoing development, which will evolve as the data and the needs of its users do. Our aim is to provide the community with the most comprehensive summary and breakdown of the structures within the AFDB. We expect TED to be used as a starting point for a whole host of analyses, including providing a comprehensive dataset to train and test a new generation of deep learning based applications in structural biology.

References and Notes

1. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

2. 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.*, doi: 10.1093/nar/gkad1011 (2023).

3. N. Borkakoti, J. M. Thornton, AlphaFold2 protein structure prediction: Implications for drug discovery. *Curr. Opin. Struct. Biol.* **78**, 102526 (2023).

4. 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**, 646–653 (2023).

5. 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**, 637–645 (2023).

6. 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**, 160 (2023).

7. R. Dustin 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 Comput. Biol.* **20**, e1011586 (2024).

8. 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**, D266–D273 (2021).

9. A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, C. A. Orengo, The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* **37**, D310–4 (2009).

10. H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, N. V. Grishin, ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).

11. A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, E. L. L. Sonnhammer, The Pfam Protein Families Database. *Nucleic Acids Res.* **30**, 276–280 (2002).

12. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).

13. J. Lees, C. Yeats, J. Perkins, I. Sillitoe, R. Rentzsch, B. H. Dessailly, C. Orengo, Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* **40**, D465–71 (2012).

14. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).

15. T. E. Lewis, I. Sillitoe, N. Dawson, S. D. Lam, T. Clarke, D. Lee, C. Orengo, J. Lees, Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **46**, D1282 (2018).

16. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).

17. N. K. Fox, S. E. Brenner, J.-M. Chandonia, SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–9 (2014).

18. C. Hadley, D. T. Jones, A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **7**, 1099–1112 (1999).

19. R. Day, D. A. C. Beck, R. S. Armen, V. Daggett, A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.* **12**, 2150–2160 (2003).

20. A. M. Lau, S. M. Kandathil, D. T. Jones, Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat. Commun.* **14**, 8445 (2023).

21. J. Wells, A. Hawkins-Hooker, N. Bordin, B. Paige, C. Orengo, Chainsaw: protein domain segmentation with fully convolutional neural networks, *bioRxiv* (2023)p. 2023.07.19.549732.

22. K. Zhu, H. Su, Z. Peng, J. Yang, A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics* **39** (2023).

23. 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.*, doi: 10.1038/s41587-023-01773-0 (2023).

24. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

25. S. M. Kandathil, A. M. Lau, D. W. A. Buchan, D. T. Jones, Foldclass and Merizo-search: embedding-based deep learning tools for protein domain segmentation, fold recognition and comparison, *biorXiv* (2024). https://www.biorxiv.org/content/10.1101/2024.03.25.586696.

26. D. Du, Z. Wang, N. R. James, J. E. Voss, E. Klimont, T. Ohene-Agyei, H. Venter, W. Chiu, B. F. Luisi, Structure of the AcrAB–TolC multidrug efflux pump. *Nature* **509**, 512–515 (2014).

27. C.-H. Tai, R. Paul, K. C. Dukka, J. D. Shilling, B. Lee, SymD webserver: a platform for detecting internally symmetric protein structures. *Nucleic Acids Res.* **42**, W296–300 (2014).

28. A. V. Kajava, A. C. Steven, "β-Rolls, β-Helices, and Other β-Solenoid Proteins" in *Advances in Protein Chemistry* (Academic Press, 2006)vol. 73, pp. 55–96.

29. S. Mesdaghi, R. M. Price, J. Madine, D. J. Rigden, Deep Learning-based structure modelling illuminates structure and function in uncharted regions of β-solenoid fold space. *J. Struct. Biol.* **215**, 108010 (2023).

30. B. Chakrabarty, N. Parekh, DbStRiPs: Database of structural repeats in proteins. *Protein Sci.* **31**, 23–36 (2022).

31. M. Laitaoja, J. Valjakka, J. Jänis, Zinc coordination spheres in protein structures. *Inorg. Chem.* **52**, 10983–10991 (2013).

32. T. Li, H. L. Bonkovsky, J.-T. Guo, Structural analysis of heme proteins: implications for design and prediction. *BMC Struct. Biol.* **11**, 13 (2011).

33. D. Ekman, S. Light, Å. K. Björklund, A. Elofsson, What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol.* **7**, 1–13 (2006).

34. G. Apic, J. Gough, S. A. Teichmann, Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).

35. S. Batey, A. A. Nickson, J. Clarke, Studying the folding of multidomain proteins. *HFSP J.* **2**, 365–377 (2008).

36. X. Zhou, J. Hu, C. Zhang, G. Zhang, Y. Zhang, Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15930–15938 (2019).

37. Y. Hou, T. Xie, L. He, L. Tao, J. Huang, Topological links in predicted protein complex structures reveal limitations of AlphaFold. *Commun Biol* **6**, 1098 (2023).

38. D. T. Jones, J. M. Thornton, The impact of AlphaFold2 one year on. *Nat. Methods* **19**, 15–20 (2022).

39. H. K. Wayment-Steele, A. Ojoawo, R. Otten, J. M. Apitz, W. Pitsawong, M. Hömberger, S. Ovchinnikov, L. Colwell, D. Kern, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).

40. D. Del Alamo, D. Sala, H. S. Mchaourab, J. Meiler, Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11** (2022).

41. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

42. N. Bordin, I. Sillitoe, J. G. Lees, C. Orengo, Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds. *Front. Mol. Biosci.* **8**, 668184 (2021).

43. C. Chothia, One thousand families for the molecular biologist, *Nature Publishing Group UK* (1992). https://doi.org/10.1038/357543a0.

44. A. Lau, N. Bordin, S. Kandathil, I. Sillitoe, V. Waman, J. Wells, C. Orengo, D. T. Jones, The Encyclopedia of Domains (TED) structural domains assignments for AlphaFold Database v4, Zenodo (2024); https://doi.org/10.5281/ZENODO.13236614.

45. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

46. W. R. Taylor, Protein structural domain identification. *Protein Eng.* **12**, 203–216 (1999).

47. Conservation of Orientation and Sequence in Protein Domain–Domain Interactions. *J. Mol. Biol.* **345**, 1265–1279 (2005).

48. D. Holten, Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. *IEEE Trans. Vis. Comput. Graph.* **12**, 741–748 (2006).

49. T. L. Pedersen, ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. [Preprint] (2024). https://ggraph.data-imaginist.com.

50. T. E. Lewis, I. Sillitoe, J. G. Lees, cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. *Bioinformatics* **35**, 1766–1767 (2018).

51. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).

52. N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, L. Davis, T. Dogan, V. Atalay, A. S. Rifaioglu, A. Dalkıran, R. Cetin Atalay, C. Zhang, R. L. Hurto, P. L. Freddolino, Y. Zhang, P. Bhat, F. Supek, J. M. Fernández, B. Gemovic, V. R. Perovic, R. S. Davidović, N. Sumonja, N. Veljkovic, E. Asgari, M. R. K. Mofrad, G. Profiti, C. Savojardo, P. L. Martelli, R. Casadio, F. Boecker, H. Schoof, I. Kahanda, N. Thurlby, A. C. McHardy, A. Renaux, R. Saidi, J. Gough, A. A. Freitas, M. Antczak, F. Fabris, M. N. Wass, J. Hou, J. Cheng, Z. Wang, A. E. Romero, A. Paccanaro, H. Yang, T. Goldberg, C. Zhao, L. Holm, P. Törönen, A. J. Medlar, E. Zosa, I. Borukhov, I. Novikov, A. Wilkins, O. Lichtarge, P.-H. Chi, W.-C. Tseng, M. Linial, P. W. Rose, C. Dessimoz, V. Vidulin, S. Dzeroski, I. Sillitoe, S. Das, J. G. Lees, D. T. Jones, C. Wan, D. Cozzetto, R. Fa, M. Torres, A. Warwick Vesztrocy, J. M. Rodriguez, M. L. Tress, M. Frasca, M. Notaro, G. Grossi, A. Petrini, M. Re, G. Valentini, M. Mesiti, D. B. Roche, J. Reeb, D. W. Ritchie, S. Aridhi, S. Z. Alborzi, M.-D. Devignes, D. C. E. Koo, R. Bonneau, V. Gligorijević, M. Barot, H. Fang, S. Toppo, E. Lavezzo, M. Falda, M. Berselli, S. C. E. Tosatto, M. Carraro, D. Piovesan, H. Ur Rehman, Q. Mao, S. Zhang, S. Vucetic, G. S. Black, D. Jo, E. Suh, J. B. Dayton, D. J. Larsen, A. R. Omdahl, L. J. McGuffin, D. A. Brackenridge, P. C. Babbitt, J. M. Yunes, P. Fontana, F. Zhang, S. Zhu, R. You, Z. Zhang, S. Dai, S. Yao, W. Tian, R. Cao, C. Chandler, M. Amezola, D. Johnson, J.-M. Chang, W.-H. Liao, Y.-W. Liu, S. Pascarelli, Y. Frank, R. Hoehndorf, M. Kulmanov, I. Boudellioua, G. Politano, S. Di Carlo, A. Benso, K. Hakala, F. Ginter, F. Mehryary, S. Kaewphan, J. Björne, H. Moen, M. E. E. Tolvanen, T. Salakoski, D. Kihara, A. Jain, T. Šmuc, A. Altenhoff, A. Ben-Hur, B. Rost, S. E. Brenner, C. A. Orengo, C. J. Jeffery, G. Bosco, D. A. Hogan, M. J. Martin, C. O'Donovan, S. D. Mooney, C. S. Greene, P. Radivojac, I. Friedberg, The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).

53. UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).

54. C. A. Orengo, W. R. Taylor, SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635 (1996).

55. R. A. Laskowski, J. Jabłońska, L. Pravda, R. S. Vařeková, J. M. Thornton, PDBsum: Structural summaries of PDB entries. *Protein Sci.* **27**, 129–134 (2018).

56. W. S. J. Valdar, Scoring residue conservation. *Proteins* **48**, 227–241 (2002).

57. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

58. T. Biswas, J. L. Houghton, S. Garneau-Tsodikova, O. V. Tsodikov, The structural basis for substrate versatility of chloramphenicol acetyltransferase CATI. *Protein Sci.* **21**, 520–530 (2012).

Acknowledgements

We would like to acknowledge continuous support from the University College London Computer Science Technical Support Group (TSG). We also thank Sameer Velankar and Martin Steinegger for helpful discussions.

Funding

This work was funded by BBSRC grant BB/T019409/1 (A.M.L. and D.T.J.), BB/W008556/1 (S.M.K. and D.T.J.) and BB/W018802/1 (I.S), Wellcome Trust grant 221327/Z/20/Z (N.B., V.P.W). J.W. acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (218529/Z/19/Z).

Author Contributions

A.M.L., N.B., I.S. and D.T.J. designed and created the datasets. A.M.L. and I.S. designed and executed the domain chopping workflow. D.T.J. designed the Foldseek domain assignment process, and N.B. carried out the Foldseek domain assignment, HMM searches and sequence clustering analysis. D.T.J. designed and A.M.L. performed the Foldclass domain assignments. D.T.J. designed and executed the novel fold workflow analysis. A.M.L, D.T.J. and V.P.W. identified novel folds. A.M.L. conducted the TED-redundant analysis. S.M.K. performed the domain-domain interaction analysis. S.M.K. and D.T.J. performed GO term analysis. A.M.L. and J.W. carried out the coverage comparison analysis. D.T.J. and C.A.O. supervised the project. A.M.L., N.B., S.M.K., V.P.W., C.A.O. and D.T.J. wrote the manuscript. A.M.L, S.M.K, V.P.W and N.B produced figures. All authors contributed to research design and manuscript revision.

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

The Encyclopedia of Domains (TED) structural domain assignments for AlphaFold Database v4, as well as associated code will be available as a Zenodo deposition at https://zenodo.org/records/13236614 (DOI: 10.5281/zenodo.13236614) upon publication (*44*). The deposition contains domain assignments for TED, PDB files for novel folds and individual domain assignments from Chainsaw, Merizo and UniDoc to facilitate further benchmarking efforts. Individual protein annotations can also be browsed from the TED website: https://ted-dev.cathdb.info.

The code for calculating consensus domain chopping, domain quality, Foldclass embedding, search, and GO term analysis is available via the PSIPRED github repository (https://github.com/psipred/ted-tools). The code for globularity prediction is part of CATH-AlphaFlow (https://github.com/UCLOrengoGroup/cath-alphaflow). The code for the TED website is available at TED-web (https://github.com/UCLOrengoGroup/ted-web).