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

Andy M. Lau, Nicola Bordin, Shaun M. Kandathil, Ian Sillitoe, Vaishali P. Waman, Jude Wells, Christine A. Orengo, David T. Jones

https://www.science.org/doi/10.1126/science.adq4946

Burak Suyunu

LifeLU Reading Group

26.12.2024

Introduction

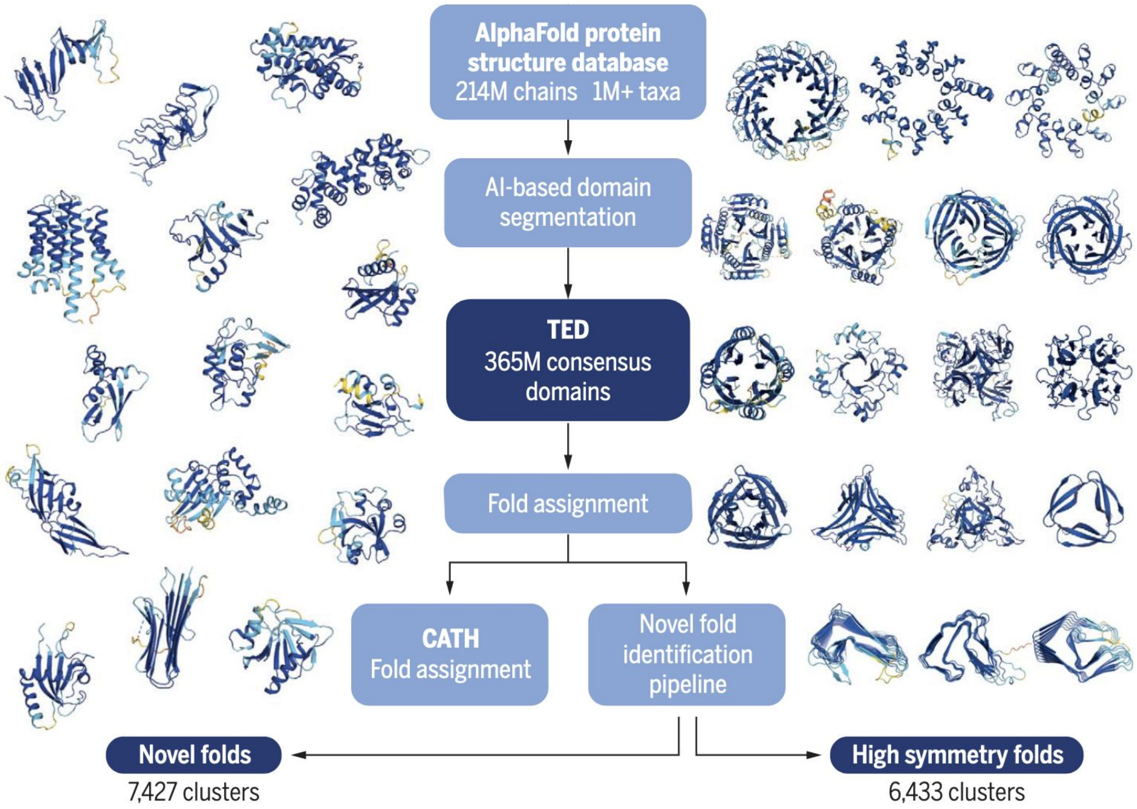
Analysis of domain composition for the entirety of the AFDB (version 4).

365 million putative domains derived from >214 million protein sequences across >1 million taxa.

Consensus of three automated parsing methodologies: Merizo, Chainsaw, and UniDoc.

Structural comparison methods such as Foldseek and an in-house deep learning method called Merizo-search

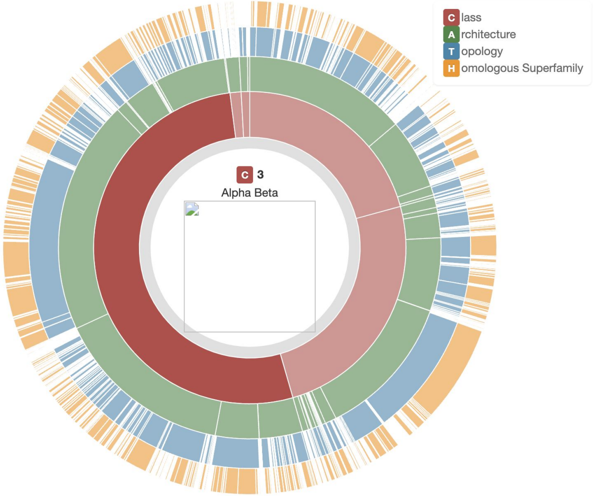
>251 million domai nsplaced on the CATH hierarchy.



TED workflow. TED consists of 365 million domains identified from the AFDB by taking a consensus from three state-of-the-art domain segmentation methods. Nonredundant domains are assigned with CATH labels, and unassigned domains are subjected to a novel fold identification pipeline. Overall, we found several thousand novel folds, as well as novel high-symmetry fold clusters.

What is CATH?

The CATH (Class, Architecture, Topology, Homology) database stands as one of the most important resources in structural biology, providing a hierarchical classification system for protein domains.



CATH Hierarchy

The name CATH reflects its four main levels of classification:

- **Class (C):** The most basic level, describing the protein domain's secondary structure composition
  - Mainly Alpha (α), Mainly Beta (β), Alpha/Beta (α/β), Few Secondary Structures
- **Architecture (A):** The general shape formed by the secondary structures
  - Barrel shapes, Sandwiches, Roll structures, Irregular structures
- **Topology (T):** The specific folding pattern of the secondary structures
- **Homology (H):** Groups of proteins with evidence of evolutionary relationship

CATH Examples

Example 1: Hemoglobin (CATH ID: 1.10.490.10)

- Class (1): Mainly Alpha - The protein is dominated by α-helices
- Architecture (10): Orthogonal Bundle - α-helices arranged at roughly right angles
- Topology (490): Globin-like - Specific arrangement of helices characteristic of globins
- Homology (10): Globin family - Shows clear evolutionary relationship to other globins

This classification tells us that hemoglobin belongs to a well-characterized family of proteins that share not just their structural features but also their evolutionary history and function of oxygen transport.

CATH Examples

Example 2: Triose Phosphate Isomerase (CATH ID: 3.20.20.70)

- Class (3): Alpha/Beta - Contains both α-helices and β-sheets
- Architecture (20): TIM Barrel - Named after this very enzyme
- Topology (20): TIM Barrel - Specific arrangement of 8 α-helices and 8 β-strands
- Homology (70): Aldolase superfamily - Related to other enzymes that catalyze similar reactions

This example shows how a protein's structure can become the archetypal representative of an entire architectural class (the TIM barrel).

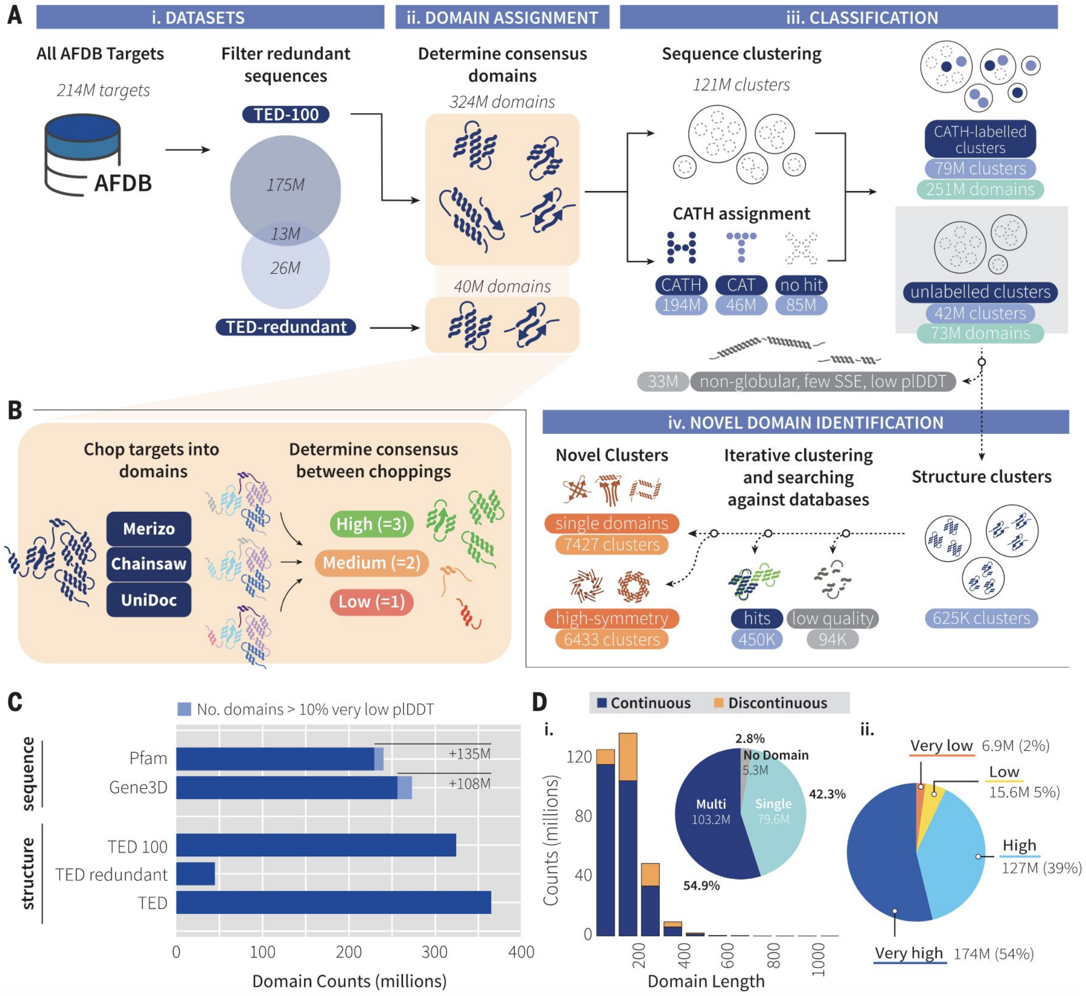
CATH Examples

Example 3: Immunoglobulin (CATH ID: 2.60.40.10)

- Class (2): Mainly Beta - Dominated by β-sheets
- Architecture (60): Sandwich - Two β-sheets packed face-to-face
- Topology (40): Immunoglobulin-like - Specific connectivity pattern of β-strands
- Homology (10): Immunoglobulin - Part of the antibody protein family

This classification demonstrates how structural arrangements can be optimized for specific functions, in this case, antigen recognition.

Workflow Of TED



Deriving a unified assignment of domains in the AFDB

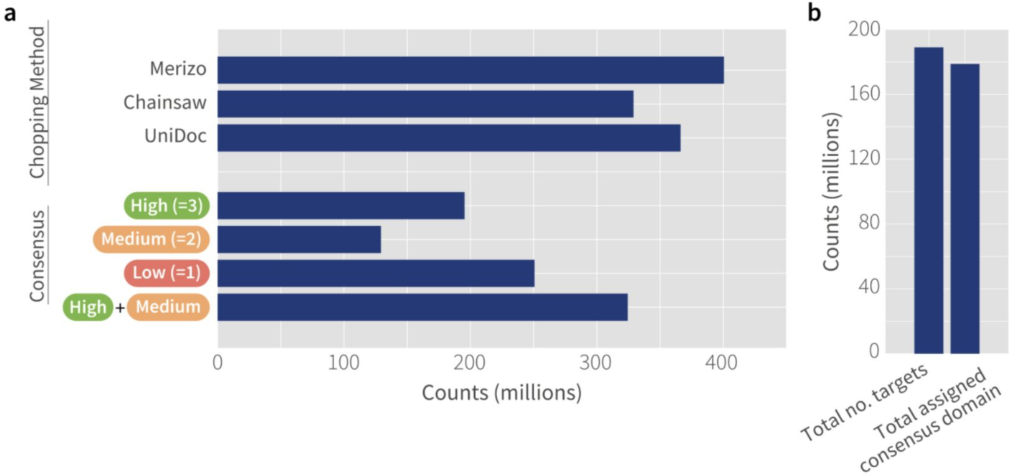


Fig. S2. Domain counts identified from consensus chopping methods in TED-100. (a)

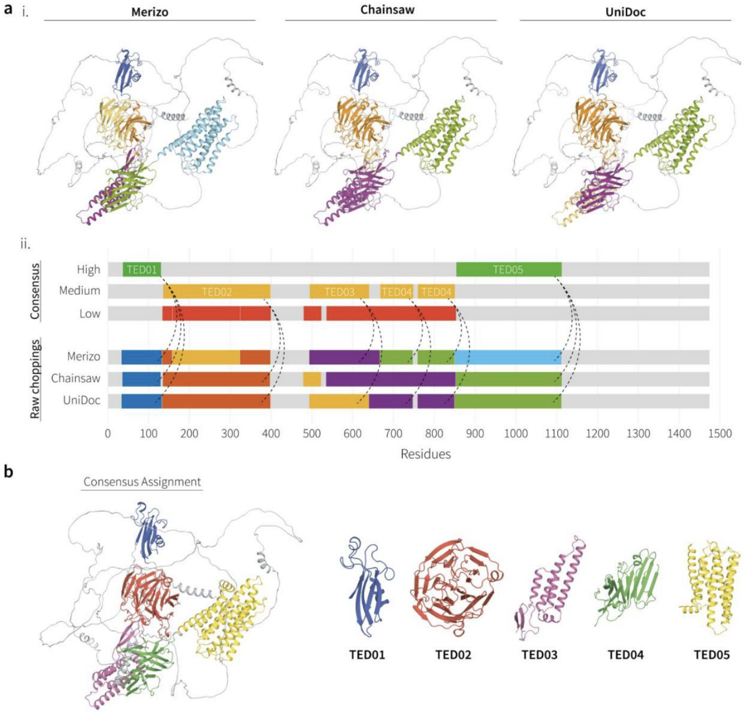


Fig. S14. Example of consensus domain derivation. **(a)** i. Example of predicted domain choppings from Merizo, Chainsaw and UniDoc for target AF-O94910-F1-model_v4 (Adhesion

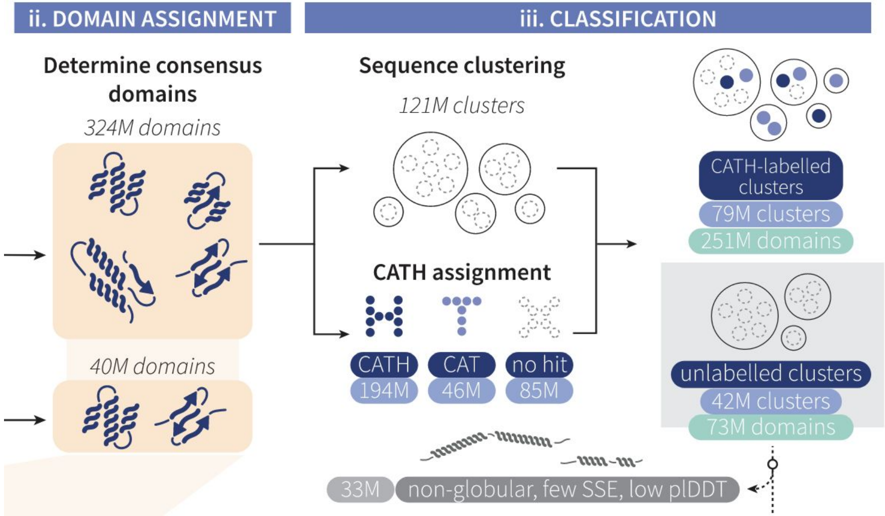
Domain classification

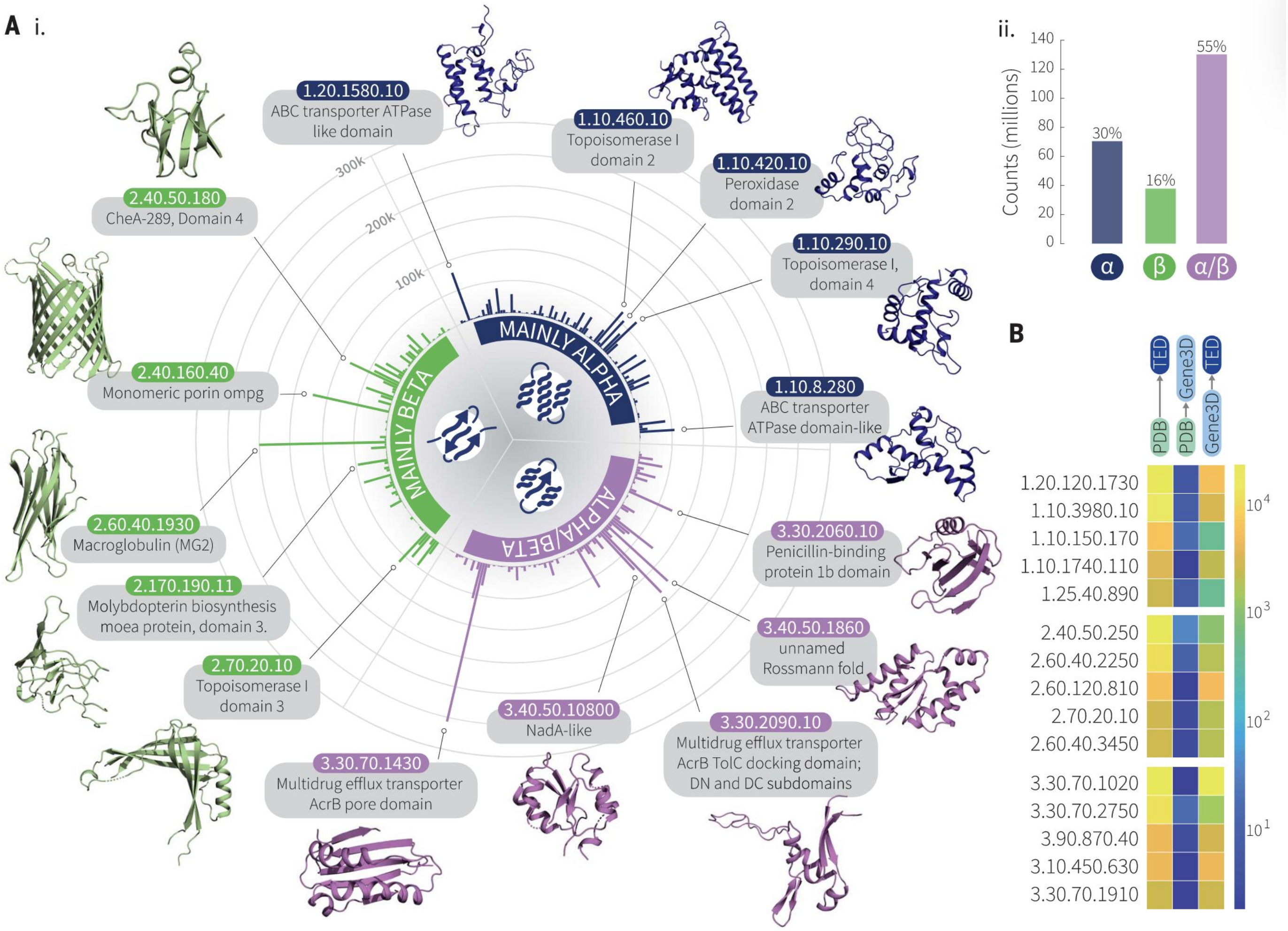
1. TED-100 domains are processed by MMseqs2, creating >121 million clusters at 50% identity.

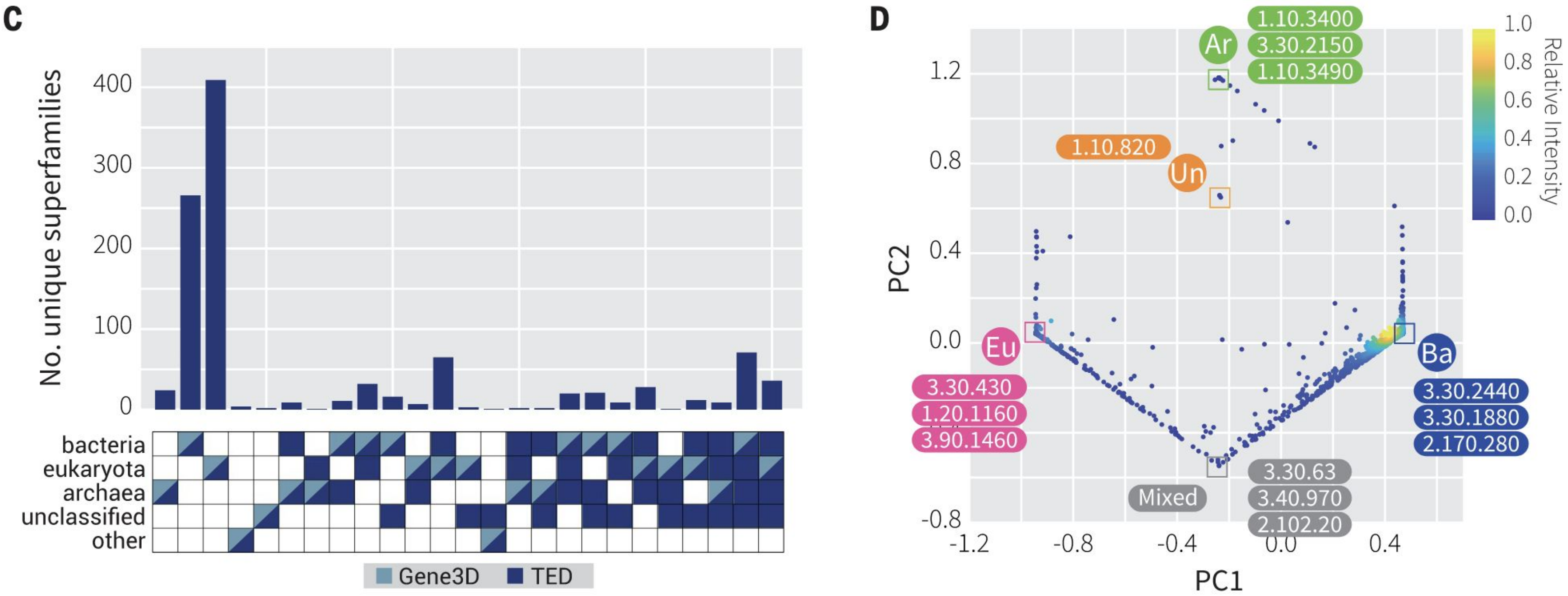
2. Foldseek was optimized using a curated CATH dataset (3186 domains) to establish thresholds for identifying superfamily (H-level) and fold (T-level) matches with 98% precision.

3. Foldseek scanned 324,389,697 domains against a CATH SSG5 library, identifying 193,939,494 H-level and 16,026,530 T-level matches, leaving 114,423,673 domains unmatched.

4. Remaining unmatched domains were analyzed using Merizo-search, which leverages Foldclass (an equivariant graph neural network) and TM-align to assign domains based on structure-embedding similarity (matches are considered as topology-level matches).




A
i.
300k
200k
100k
MAINLY ALPHA
MAINLY BETA
ALPHA/BETA
1.20.1580.10
ABC transporter ATPase like domain
1.10.460.10
Topoisomerase I domain 2
1.10.420.10
Peroxidase domain 2
1.10.290.10
Topoisomerase I, domain 4
1.10.8.280
ABC transporter ATPase domain-like
3.30.2060.10
Penicillin-binding protein 1b domain
3.40.50.1860
unnamed Rossmann fold
3.30.2090.10
Multidrug efflux transporter AcrB TolC docking domain; DN and DC subdomains
3.40.50.10800
NadA-like
3.30.70.1430
Multidrug efflux transporter AcrB pore domain
2.70.20.10
Topoisomerase I domain 3
2.170.190.11
Molybdopterin biosynthesis moea protein, domain 3.
2.60.40.1930
Macroglobulin (MG2)
2.40.160.40
Monomeric porin ompg
2.40.50.180
CheA-289, Domain 4
ii.
Counts (millions)
140
120
100
80
60
40
20
0
30%
16%
55%
α
β
α/β
B
PDB → TED
PDB → Gene3D
Gene3D → TED
1.20.120.1730
1.10.3980.10
1.10.150.170
1.10.1740.110
1.25.40.890
2.40.50.250
2.60.40.2250
2.60.120.810
2.70.20.10
2.60.40.3450
3.30.70.1020
3.30.70.2750
3.90.870.40
3.10.450.630
3.30.70.1910
10^4
10^3
10^2
10^1


C
No. unique superfamilies
400
300
200
100
0
bacteria
eukaryota
archaea
unclassified
other
Gene3D
TED
D
Ar
1.10.3400
3.30.2150
1.10.3490
1.10.820
Un
Eu
3.30.430
1.20.1160
3.90.1460
Ba
3.30.2440
3.30.1880
2.170.280
Mixed
3.30.63
3.40.970
2.102.20
PC2
1.2
0.8
0.4
0.0
-0.8
PC1
-1.2
-0.8
-0.4
0.0
0.4
Relative Intensity
1.0
0.8
0.6
0.4
0.2
0.0

Novel Domain Identification Workflow

Initial Filtering:

- 41.8M domain clusters were assessed using normalized radius of gyration and packing density metrics.
- Secondary structure element (SSE) filtering excluded domains with fewer than six SSEs.
- The plDDT80 metric ensured retention of well-folded domains while excluding clusters with unfolded tail regions.

Library Comparisons: Iterative searches were performed against PDB, CATH, ECOD, and SCOPe libraries using Foldseek to identify novel domains.

Core Subset Selection: High-symmetry domains were separated to identify potential novel folds, accounting for repetitive units like WD40 repeats.

Chopping Quality Assessment: Retrained Foldclass networks evaluated domain segmentation, distinguishing between well-chopped domains and over/under-chopped segments.

Novelty Ranking: Final domains were ranked using Foldclass embeddings, calculating Euclidean distances to known domain neighbors in established libraries.

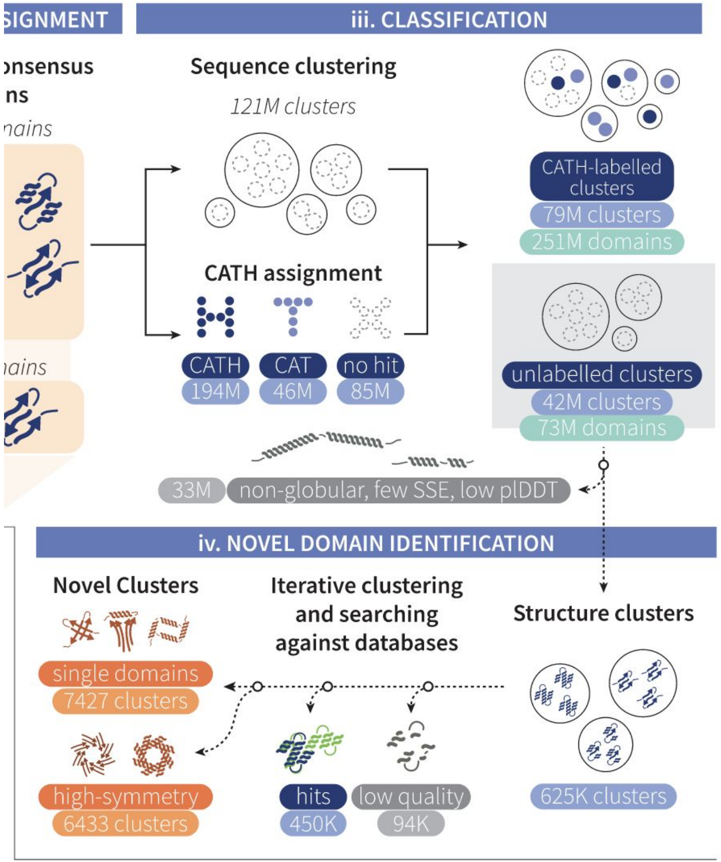
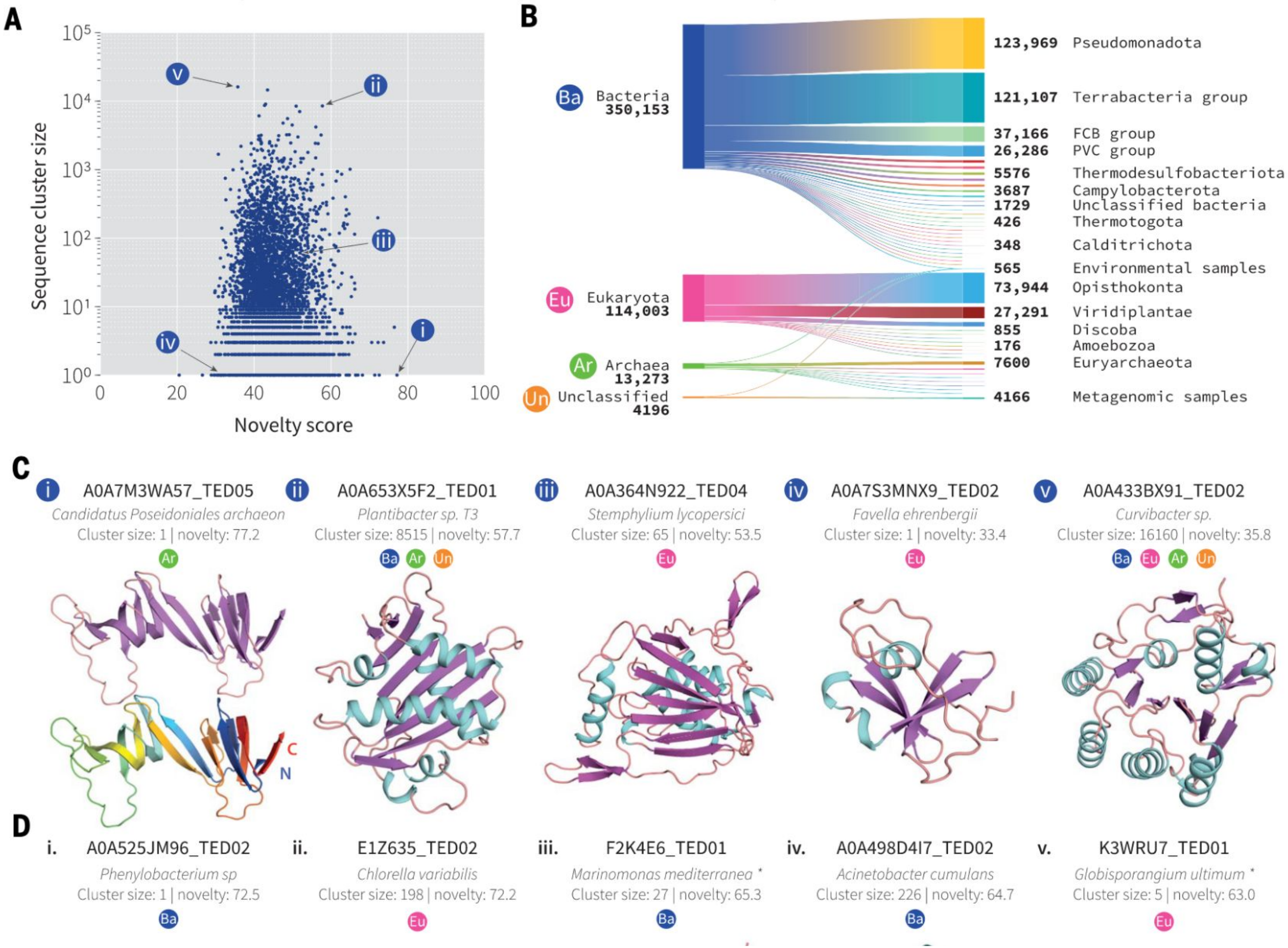


Fig. 4. Examples of novel domain clusters identified in TED. (**A**) Comparison of domain novelty score versus sequence cluster size (*n* = 7427). Novelty scores are predicted by the Foldclass algorithm where novel domains are ranked with a score close to 100. (**B**) Taxonomic distribution of novel 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**) (i) to (v) correspond to labels shown in (A). In (i), the bottom subpanel shows the arrangement of strands that form the coiled hairpin loop from the N terminus (blue)

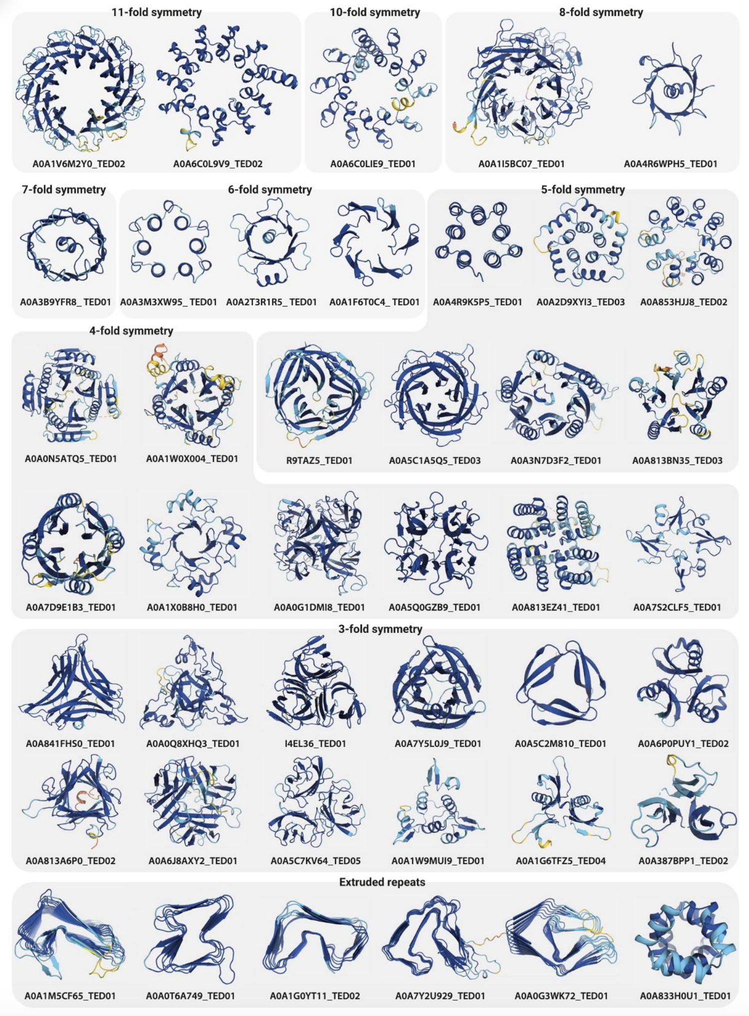


High symmetry domains and extruded repeats:

Symmetric Domain Clusters: Domains with high internal symmetry, such as WD40 beta-propellers, require special categorization. Using the SymD program, clusters with a symmetry Z score >9 were grouped into 6433 highly symmetric novel fold clusters. These include unique architectures like an 11-bladed beta-propeller and various other propeller arrangements.

Extruded Repeat Architectures: A category of cyclic repeats, termed "extruded repeats," extruded along an axis to form highly repetitive and symmetric structures. Examples include alpha- and beta-solenoids and horseshoes with diverse, unstructured loops.

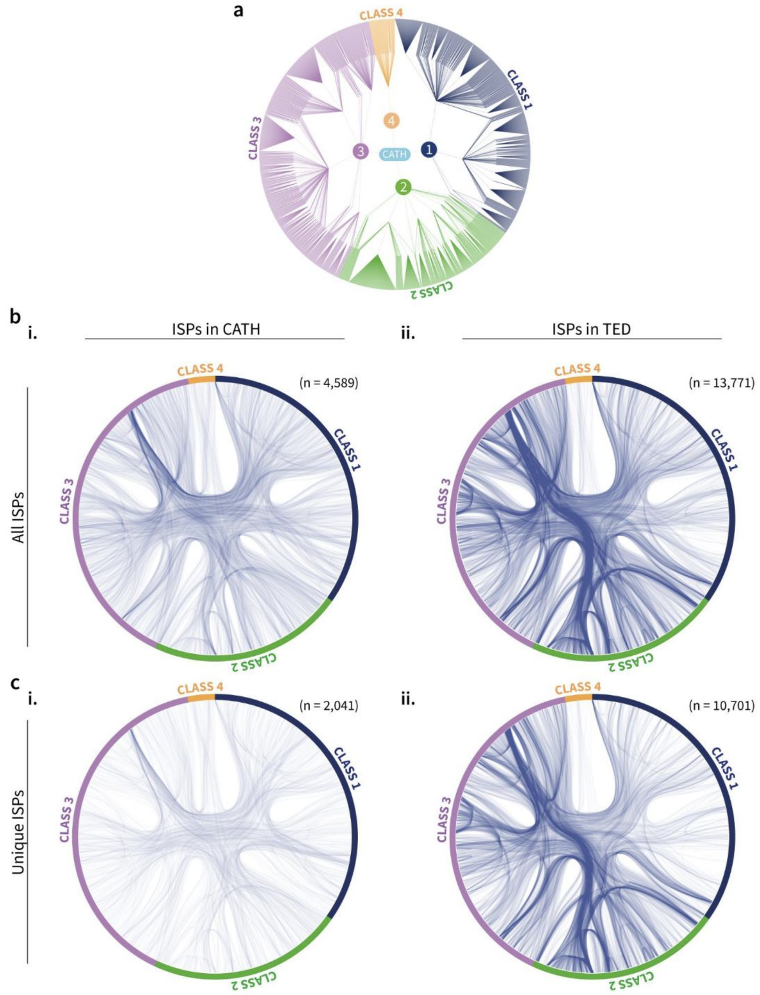
Systematic Studies and Databases: Repeat architectures have been curated by studies and databases (e.g., DbStRiPs), revealing connections to known solenoid folds, including HEAT, ankyrin, and armadillo repeats.



Interactions between domain pairs

Domain Interaction Identification: Domains are defined as interacting if at least three Cβ (or Cα for Gly) atoms are within 8 Å. An interacting Superfamily Pairs (ISP) is the collection of all instances of such contacting pairs of domains belonging to specific superfamilies.

Comparison of TED and CATH ISPs: TED includes 27,280,057 domain interaction instances across 13,771 ISPs, significantly outnumbering CATH’s 196,234 instances across 5111 ISPs. Most ISPs in TED are unique, while 2041 are exclusive to CATH.

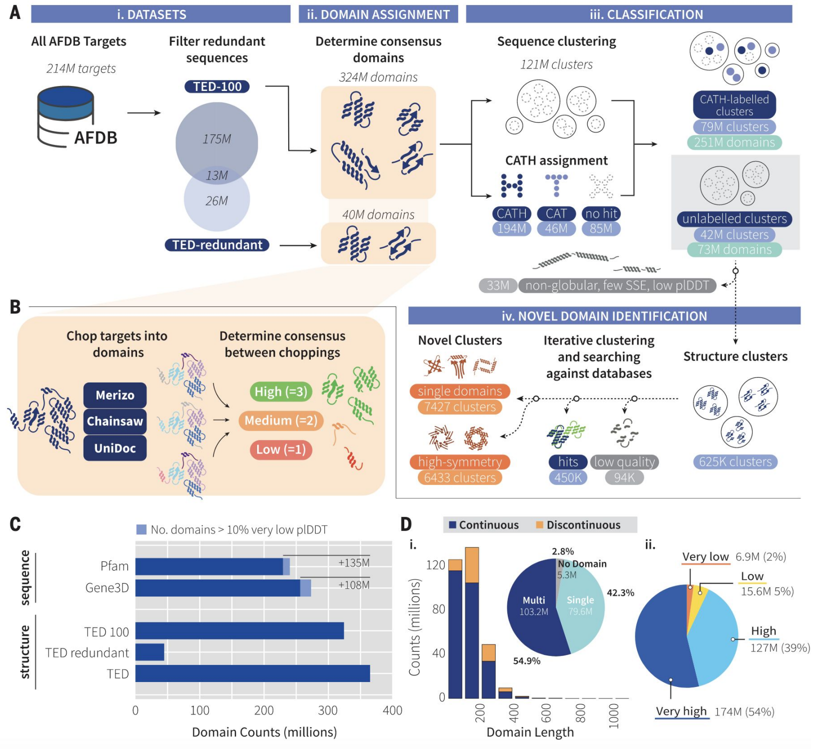


Conclusion

- TED fills critical gaps left by sequence-based methods, offering a comprehensive resource for identifying and classifying protein domains within the AFDB.
- By uncovering over 100 million previously undetectable domains and revealing thousands of new folds, TED substantially expands our understanding of protein structure diversity.
- TED highlights evolutionary trends, including conserved and lineage-specific folds, providing insights into protein evolution across the Tree of Life.
- The resource uncovers more than 10,000 new structural interactions and enriches knowledge of domain-packing geometries, tripling known superfamily interactions.
- With ongoing updates aligned with AFDB releases, TED facilitates functional analyses, from drug discovery to evolutionary studies, setting the stage for future breakthroughs in structural biology.

Thanks for listening!

Questions-Comments?


A
i. DATASETS
ii. DOMAIN ASSIGNMENT
iii. CLASSIFICATION
All AFDB Targets
214M targets
AFDB
Filter redundant sequences
TED-100
175M
13M
26M
TED-redundant
Determine consensus domains
324M domains
40M domains
Sequence clustering
121M clusters
CATH assignment
CATH
194M
CAT
46M
no hit
85M
CATH-labelled clusters
79M clusters
251M domains
unlabelled clusters
42M clusters
73M domains
33M
non-globular, few SSE, low pIDDT
B
Chop targets into domains
Merizo
Chainsaw
UniDoc
Determine consensus between choppings
High (=3)
Medium (=2)
Low (=1)
iv. NOVEL DOMAIN IDENTIFICATION
Novel Clusters
single domains
7427 clusters
high-symmetry
6433 clusters
Iterative clustering and searching against databases
hits
450K
low quality
94K
Structure clusters
625K clusters
C
No. domains > 10% very low pIDDT
sequence
structure
Pfam
Gene3D
TED 100
TED redundant
TED
+135M
+108M
0
100
200
300
400
Domain Counts (millions)
D
Continuous
Discontinuous
i.
120
80
40
0
Counts (millions)
200
400
600
800
1000
Domain Length
2.8%
No Domain
5.3M
Multi
103.2M
Single
79.6M
42.3%
54.9%
ii.
Very low
6.9M (2%)
Low
15.6M 5%)
High
127M (39%)
Very high
174M (54%)