Tracing the function expansion for a primordial protein fold in the era of fold-based function prediction: beta-trefoil

Moushmi Goswami[1] and Subhashini Srinivasan[1, 2]

1. Institute of Bioinformatics and Applied Biotechnology, Biotech Park, Electronic City Phase I, Bangalore 560100, India
2. Corresponding author

Key words: DALI; Foldseek; Functional genomics; CASP; NR-database; UniProt database; PDB database; PFam database; Homology modeling; Protein structure prediction; Protein evolution;

Acronym: PDB- Protein Data Bank; BFD- Big Fantastic Database; DNA- Deoxyribonucleic Acid; EST- Expressed Sequence Tag; BLAST- Basic Local Alignment Search Tool; HMM- Hidden Markov Model; CASP- Critical Assessment of Structure Prediction; ML- Machine Learning; AI- Artificial Intelligence; r.m.s.d.- Root Mean Square Deviation; AFdb- AlphaFold database; CATH- Class, Architecture, Topolography and Homology; DALI- Distance-matrix ALIgnment; 3Di- Three-Dimensional Interactome; PFam- Protein Families; RefSeq- Reference Sequence

ABSTRACT

The incredibly narrow protein fold bottleneck, which separates the billions of unique proteins on one side to deliver diverse biological functions on the other, arises from folds tolerating mutations during evolution. One such fold, called beta-trefoil, is present in functionally diverse proteins including cytokines involved in the immune system. The unrecognizable sequence-level diversity even among paralogs of interleukin-1, one of the cytokines within the same chromosomal locus, suggests the resilience of this fold to mutational on-slaught. Furthermore, beta-trefoil containing-proteins are known to exist along with other domains allowing them to bind to diverse macromolecular types in achieving the observed functional diversity. In this study, we use the beta-trefoil, a fold prevalent across kingdoms, to assess the reach and limitations of function prediction using fold-fold comparison. We found thirty-two distinct functions for proteins containing at least one trefoil domain by mining both the PDB and AlphaFold databases. Among the novel domain architecture including trefoil domain, we find domains such as chitinase, lipase, beta-glucosidase, protein kinase, peptidoglycan-binding + peptidase matrixin, glycosyl hydrolases family 3 + PA14 + fibronectin type- III, alpha galactosidase A, PhoD-like phosphatase, insecticidal crystal toxin, trypsin, alginate lyase and two unknown domains. The limitations of structure-structure comparison tools may arise from converging evolution in folds. This was addressed by creating phylogenetic tree from structure-based multiple sequence alignment of trefoil domains from diverse kingdoms and functions. The clades by kingdoms as opposed to functions may support the hypothesis that beta-trefoil evolved independently, perhaps creating a twilight zone in function prediction using fold-fold comparison.

INTRODUCTION

Nobel Laureate Francois Jacob, in his essay on Evolution and Tinkering, stated that nature is a better tinkerer than an inventor. As evolution proceeds to create complex organisms from a

45 small repertoire of protein folds, one can expect accumulating diversity in the amino acid
46 sequences of proteins sharing the same fold. Considering that protein functions are dictated by
47 their unique three-dimensional structures, the accepted mutations during evolution are expected to
48 preserve the overall structural fold of the original protein. While, in some functional families, one
49 can recognize the remnant sequence motifs of the ancient protein, there are protein fold-families
50 with no recognizable sequence signatures, yet sharing folds and, perhaps, function.

51 In 1992, Cyrus Homi Chothia estimated the number of unique folds to be as low as one
52 thousand[3]. As of 2023, the protein data bank (PDB) has 3D structures of 227,933 proteins with
53 ~7000 clusters less than 30% identity representing only 1,257 unique folds[1]. The five-fold
54 reduction in the number of folds from sequence-based clustering reflects both i) the diversity in
55 sequences within a family below the 30% identity cutoff, and ii) the differing domain architecture
56 within a protein family limiting homology over good lengths of the proteins. More recently, the
57 number of sequence clusters in the BFD (big fantastic database), using a collation of 2 billion
58 proteins, is as large as 65,983,866 using a 30% identity cutoff and 90% coverage over the length
59 of proteins[2].

60 The exponential lag in solving 3D structures of proteins and, until recently, our inability
61 to predict structures of divergent and novel proteins from primary sequences, made sequence
62 homology based-function prediction a norm for the last several decades. We have come a long
63 way since the efforts by Dayhoff in collating thousands of published protein sequences[3]. The
64 development of tools for sequence homology-based function prediction kept pace with growing
65 sequence databases, bridging the lag in structure determination. For example, the early 1990's saw
66 a boom in DNA sequence databases from EST-sequencing efforts, triggering an unprecedented
67 race to discover novel paralogs of genes of therapeutic importance. Function prediction from

68 fragmented DNAs was the need of the time, which was met wholesomely by the timely
69 development of the most popular molecular biology tool in 1990 called BLAST[4]. To improve the
70 sensitivity in sequence-based structure prediction, among divergent paralogs/homologs, hidden
71 markov models (HMM) for individual fold families were developed[5].

72 Theoretically one should be able to predict functions of divergent proteins using the shared
73 folds. Parallel efforts continued to predict structures of proteins based on homology at the
74 sequence-level. However, the progression of homology-based structure prediction can be likened
75 to decades where nothing happened. Slowly but surely, novel strategies including the incorporation
76 of multiple templates, the refinement of secondary structure prediction, the optimization of *ab*
77 *initio* loop modeling, and the use of backbone-dependent rotamer libraries for side-chain
78 placement continued to improve structure prediction methods over decades[6]. That the native fold
79 is present among the thousands of simulated structures kept the lights-on for researchers through
80 community-based challenge[7]. Wodak et al. has provided a detailed review of progress in
81 methodologies achieved via collective intelligence through community experiments like CASP
82 (critical assessment of structure prediction) and other efforts[8]. Significant improvements in
83 methods were observed by CASP10 by introducing co-evolution to guide the folding process[9].
84 Parallel to this effort, the deluge in protein sequences from advances in next generation sequencing
85 revealed more and more coevolving pairs within folding domains adding additional distance
86 constraints between distal amino acids in the primary sequence that are likely to be spatially
87 proximal — advancing structure prediction by leaps and bounds. By 2020, the convergence of
88 collective and artificial intelligence along with advances in high-throughput sequencing
89 technologies, we entered an era when decades happen in weeks. By the time of CASP14,
90 AlphaFold2 predicted 3D structures of test proteins to unprecedented accuracy[10].

91 Currently, AlphaFold-predicted structure database (AFdb) contains hundreds of millions
92 of predicted structures providing the basis for functional genomics using fold-fold comparison.
93 Methods to predict functions using AFdb are on the rise. In one such effort, using DeepFRI, the
94 contribution of AFdb in enhancing the accuracy of protein function prediction achieved previously
95 by just using PDB was demonstrated[11]. Their results highlighted the importance of AFdb structures
96 among all three Gene Ontology (GO) categories.

97 The success of AlphaFold in predicting high-quality tertiary structures inspired
98 development of PANDA-3D[12], which is a novel computational tool designed to predict protein
99 functions by leveraging the high-quality 3D structural models in AFdb. Developed as a deep-
100 learning-based method, PANDA-3D focuses on combining the geometric vector perceptron graph
101 neural networks and transformer decoders for multi-label classification of GO over the traditional
102 function prediction tools using sequence-based annotation such as in PANDA[13].

103 The research by Durairaj et al. explores the diversity of protein structures and functions
104 through advanced computational tools, leveraging the AFdb[14]. By integrating structure, sequence
105 similarity and machine learning tools like DeepFRI, the study developed an annotated sequence
106 similarity network to explore protein functions and evolutionary relationships. The study analyzed
107 over 350 million protein sequences from UniProt, revealing that many proteins remain
108 unannotated/uncharacterized due to limitations in sequence-based annotation methods. Their
109 efforts significantly enhance understanding of these "dark" proteins by the identification of new
110 protein families, such as TumE-TumA, a superfamily of toxin-antitoxin systems and discovery of
111 the beta-flower fold, a unique structural motif.

112 In yet another effort the authors attempted to cluster the 214 million structures in AFdb
113 using a novel structure comparison tool[15]. The authors use Foldseek, a scalable structural-

114 alignment-based method, to identify ~18 million clusters with 2.3 million non-singletons. Out of
115 the non-singleton clusters, 31% was unannotated representing potentially novel structures. Their
116 key findings include species-specific clusters comprising 4% of the total clusters, hinting at *de*
117 *novo* gene birth. Evolutionary insights suggest most clusters are ancient, with some human
118 immune-related proteins showing remote homology to prokaryotic species, suggesting ancient
119 systems may have been co-opted into specialized functions in higher organisms. Also, they report
120 new domain families, expanding the understanding of protein evolution and function.

121 Building on these advances, we have used both PDB and AFdb to explore the functional
122 diversity achieved by a single fold such as a primordial beta-trefoil fold[16]. This approach could be
123 considered a bottom-up approach to the work just described above, which can be considered top-
124 down. There are 1,387 protein containing trefoil domain in the current version of the PDB.
125 According to SCOP classification, beta-trefoil is implicated in eight distinct functional families
126 with capacity to bind to DNA (3brf), to carbohydrate (1jly, 1jlx), to chlorophyll (2dre), to
127 immunoglobulin-like receptor domain (1ira), to actin in a pH dependent fashion (1hcd), proteases
128 (6dwu) and AbfB (1wd3 and 2wd4).

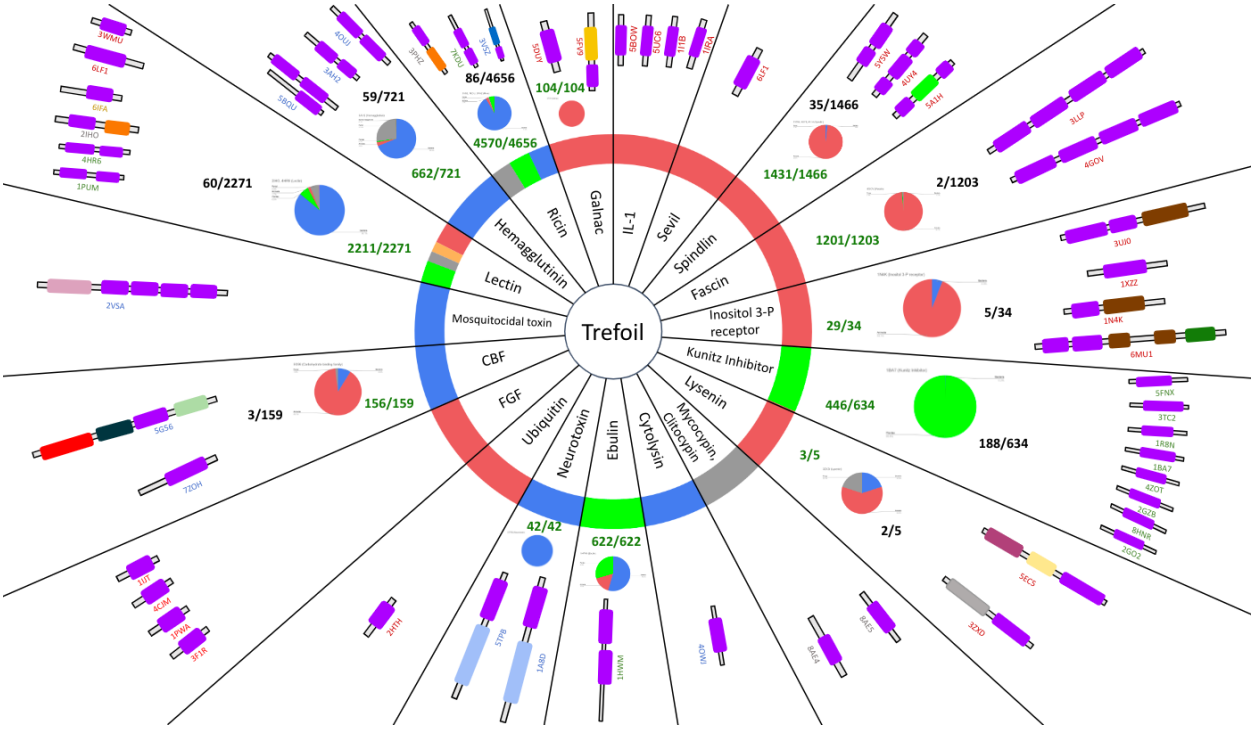
129 **RESULTS**

130 In this work, the beta-trefoil fold was chosen to demonstrate the reach and limitations of
131 functional classification of a primordial fold using fold-fold comparison. Trefoil fold not only
132 displays functional diversity by promiscuously binding to every class of macromolecules, but is
133 the most divergent fold family challenging the sequence-based function prediction even among
134 paralogs of the same functional class within the same locus in human[17].

135

136 **Functional diversity of trefoil fold in PDB**

As representative of the trefoil fold, we selected accessions for 3D structures of five known proteins including 1i1b, 5uc6, 4p0j, 5bow and 6ncu from PDB. During the first iteration, each of these structures were searched for potential homologs against the PDB using DALI[18], a structure comparison tool. The r.m.s.d. cutoff for homolog selection was set at 2 Å and the query coverage greater than 80%. This process revealed nineteen distinct functions among proteins containing the trefoil fold. This is however an underestimate, since we missed other functions present in PDB because of the stringent cutoff. For example, we missed proteins containing the trefoil fold in the PDB such as LAG-1 (3brf), AbfB (1wd3) and beta-L-arabinopyranosidase (3a21). As shown in Figure 1, we found 51 entries containing trefoil domain spanning diverse functions in the PDB including GalNAc, Hemagglutinin, Lectin, Mosquitocidal toxin, Carbohydrate-Binding Family, Fibroblast Growth Factor, Ubiquitin, Neurotoxin, Ebulin, Cytolysin, Mycocypin/Clitocypin, Ricin, Lysenin, Kunitz Inhibitor, Inositol-3 Phosphate receptor, Fascin, Spindlin, Sevil and IL-1.



151

152 Figure 1: The core of the sundial represents the trefoil fold, a primordial recurring structural motif
153 in various proteins. The second ring classifies the proteins into functional families. The third layer
154 shows the distribution of trefoil fold across biological kingdoms, color-coded as blue for bacteria,
155 gray for fungi, green for plants, red for Metazoa, and yellow for protozoa. Pie charts within
156 functional families depict kingdom-level distribution of multidomain trefoil proteins in AFdb. The
157 outermost layer illustrates domain architecture, with colored blocks indicating different protein
158 domains, where purple represents the trefoil domain and others denote additional domains that
159 contribute to the function within the spoke such as Agglutinin (orange), Glycosyl transferase
160 family (dark-yellow), Glycosyl hydrolases family (dark-blue), Clostridium neurotoxin receptor-
161 binding domain (light-blue), RIH domain (brown), Ryanodine receptor-associated domain (dark-
162 green), Unknown domain 1 (light-pink), Unknown domain 2 (dark-pink), Unknown domain 3
163 (lemon-yellow), Unknown domain 4 (grey), Unknown domain 5 (pastel-green), Spindlin/Ssty
164 family domain (light-green), Cellulase (red), Carbohydrate binding family domain (forest-green).

165

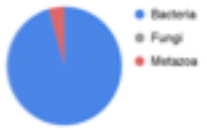
166 The domain architecture of every protein identified containing the trefoil fold from the
167 PDB was characterized using HMM and their domain architecture is shown in the outermost ring
168 within each of the respective functional spokes. Table 1 lists 18 representative protein structures
169 containing distinct domain architecture that were selected for our second iteration. These
170 representative protein structures were queried against the AFdb v2.0 using the 3Di option in
171 Foldseek[19]. During this iteration, the search results were filtered to include only those with an
172 r.m.s.d. below 3 Å and covering more than 30% of the query structure. The 3 Å increase in cutoff
173 is based on our experience with PDB search, where we missed some functions because of the 2 Å

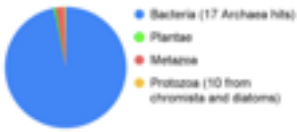
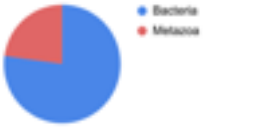
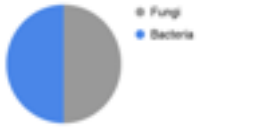
174 cutoff. The stringency of 30% query coverage for this iteration was essential to identify multi-
175 domain proteins containing trefoil from AFdb.

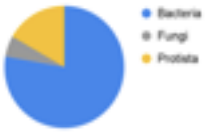
176 Details of the search results from mining AFdb are elaborated in Table 1. The second and
177 third column contain the query IDs for the representative query structures and their respective
178 kingdoms. The fourth column reports the number of hits with 80% versus 30% coverage from
179 AFdb and the last column highlights the kingdom diversity among the 30% hits. With a few
180 exceptions, hits with 80% and more coverage show less diversity at the kingdom level. We find
181 considerable kingdom-level diversity from AFdb with respect to the queries in column 3. For
182 example, with the query 3vsz belonging to bacteria, we find AFdb hits covering all four kingdoms.
183 On the other hand, 7kdu belonging to plants does not yield a single hit from plants. Similarly, 1ba7
184 from plants yields the structures only from bacteria.

185

186 **Table 1:** Validation of the approach taken here

Protein function	**Representative Query PDB ID**	**Query Kingdom**	**Total number of Hits (>80% /30% coverage)**	**Number of Hits from a different kingdom among 30% coverage and less**
Ricin	7KDU	Plantae	221/2250	

Ricin	3VSZ	Bacteria	2/368	
Ricin	3PHZ	Fungi	30/2038	
Hemagglutinin	3AH2	Bacteria	14/721	
Lectin	2IHO	Fungi	37/2271	
Lectin	4HR6	Plantae	152/553	
Carbohydrate binding family (CBF)	5G56	Bacteria	5/159	
Neurotoxin	5TPB	Bacteria	39/42	-
Ebulin	1HWM	Plantae	175/622	
Lysenin	3ZXD	Metazoa	5-Mar	

Lysenin	5EC5	Metazoa	12-Jan	
Kunitz inhibitor	1BA7	Plantae	631/634	
Inositol 3-P receptor	1N4K	Metazoa	Jun-34	
Fascin	4GOV	Metazoa	1049/1203	
Spindlin	5Y5W	Metazoa	16/55	
Spindlin	4UY4	Metazoa	76/188	
Spindlin	5A1H	Metazoa	1069/1223	
Galnac	5FV9	Metazoa	102/104	-

In column 4 of Table 1, the number of hits with 80% and 30% query coverage is shown. For some of the queries, the number of hits are large for 80% cutoff allowing us to interrogate the accuracy of predicted structures with respect to sequence identity. For the eleven query structures with sufficient number of hits (purple, Table 1: column 2), the r.m.s.d. versus sequence identity

are plotted in Figure 2. As expected, there is a clear inverse correlation between r.m.s.d. and sequence identity among these hits, validating that AlphaFold is a mega homology modeling tool.

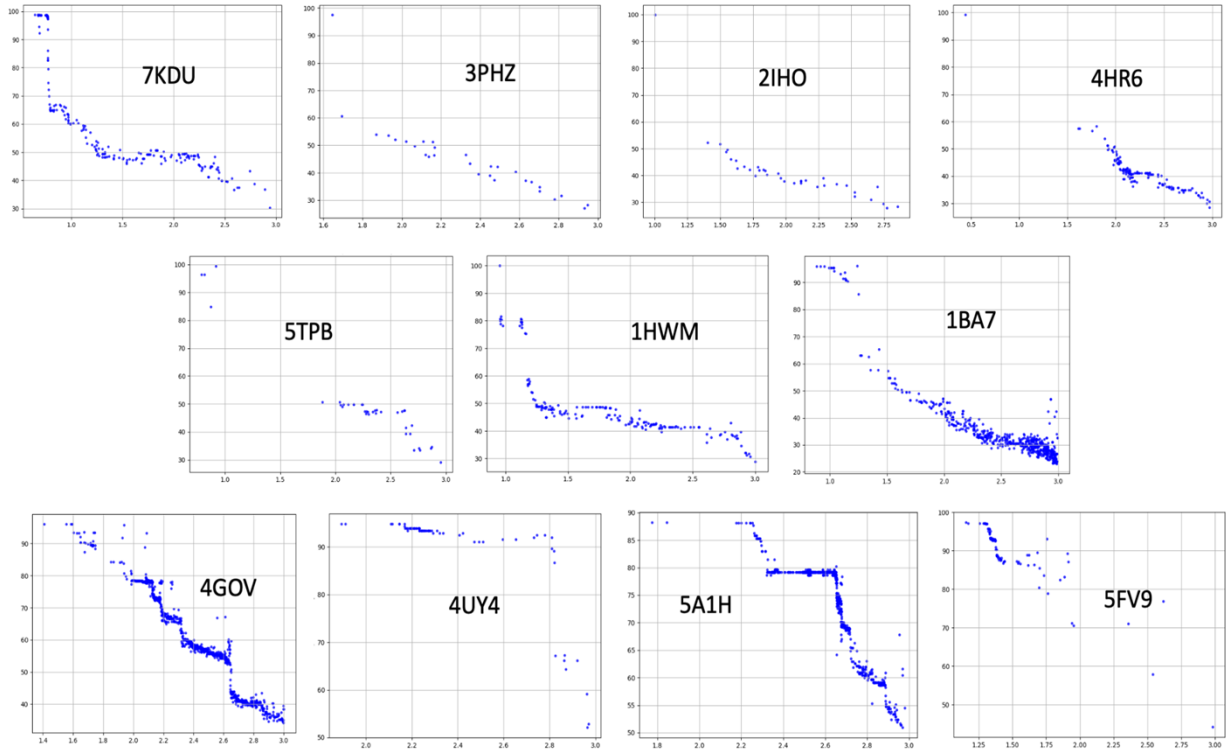


Figure 2: For the eleven query PDB structures with accession mentioned inside, each plot shows r.m.s.d. (x-axis: 0-3 Å) versus percent identities (y-axis: 30-100) for hits against AFdb with more than 80% query coverage.

Figure 2 shows the plot of r.m.s.d on the x-axis verses sequence identity on the y-axis for hits covering 80% of the query sequence against AFdb, for the eleven query structures inset within the respective plots. The hits from AFdb for the eleven selected query PDB structures shown in Figure 2, vary in pattern widely. A few follow a step pattern with each step representing distinct classes or orders. For example, each step in the hits from 4gov, from metazoa, represent homologs

206 from birds at 78%, to fish at 66%, whale at 55% and Arthropoda at 42% identity. Similarly, hits
207 from 5a1h, from metazoa, have a step at 78% representing birds and bats. Other predominant
208 patterns in Figure 2 include those with missing structures above ~60% identity cutoff and below
209 ~1.5 r.m.s.d. cutoffs, such as hits from 4hr6, 2iho and 5tpb. For example, the query 4hr6, is a lectin
210 from snake gourd- *Trichosanthes anguina*, a plant with hits between 60% - 50% belong to the
211 order Cucurbitales, hits between 50% - 40% mostly belong to the order Dipsacales and
212 Malpighiales, and a few from Rosales and Fabales, hits between 40% - 30% mostly belong to the
213 order Malvales and very few from Malpighiales, Cucurbitales and Caryophyllales. The query 2iho
214 is a lectin, which is from mushroom, and finds hits that are also from mushrooms (Basidiomycetes)
215 with a few exceptions from Ascomycetes, which is a sac fungus. The query 5tpb is a toxin from
216 bacterium *Clostridium botulinum* with most hits belonging to the same species and a few from the
217 same genus. In order to explain the gap in the region spanning low r.m.s.d. (<1.5 Å) and high
218 sequence identity (> 60%) in hits using query structure 4hr6, 2iho and 5tpb in Figure 2, we blasted
219 the sequence for the query protein against NR (Supplementary Figure 2 1C, 2B and 3C) and
220 UniProt databases (Supplementary Figure 2 1B and 3B) and find that this is resulting from
221 incompleteness in AFdb resulting from missing UniProt entries.
222 The third pattern is a sudden drop to 60% identity still showing low r.m.s.d. such as in
223 7kdu, a Rosid and in 1hwm, an Asterid. The 3-4 layers of hits against 7kdu are all plants from
224 different orders. For example, 98% identity is for another Rosid, 66% for plant species under
225 magnolids, and species under monocots show up at 43% identity. Similarly, hits against 1hwm at
226 81% come under Asterids and those at 43% belong to Rosids.
227

228 **Functional annotation**

229 Functional annotations of UniProt proteins available online are by homology at the
230 sequence level. These will include partial domain-level homology and may miss domain
231 architecture-based functional annotation. Hereafter, even for characterized entries in UniProt,
232 independent annotation was attempted using fold-fold comparison. In order to discover novel
233 domain architecture containing the trefoil fold, we interrogated hits with query coverage greater
234 than 30% for multi-domain query structures with the assumption that trefoil fold covers at least
235 30% of the multi-domain proteins. Kingdom-level diversity in functions for the query functions in
236 Table 1 is shown in pie-chart within the respective functional spokes in Figure 1 and Table 1. This
237 makes it evident that the fold-fold comparison strategy has the reach to identify homologs across
238 kingdoms. For example, the CBF (carbohydrate binding family) extends to metazoans and is not
239 limited to microbes as observed from PDB.

240 In Figure 3, we report thirteen novel UniProt proteins from AFdb containing a trefoil
241 domain along with an additional domain, which are not represented together in the PDB. These
242 proteins were categorized based on the presence or absence of these additional domains in the PDB
243 with or without trefoil domain. i) Category-1: Proteins that share their trefoil fold, but the structure
244 of the other domain is not present in the PDB database, suggesting novel domain architecture. ii)
245 Category-2: Proteins whose additional domain is found in the PDB without the trefoil fold. In these
246 cases, it was verified that the trefoil domain is absent even in the full length protein of the
247 corresponding PDB structures, indicating that while the additional domain is well represented, its
248 association with the trefoil fold is unique to the proteins identified. iii) Category-3: Proteins that
249 are represented in the PDB with both the trefoil fold and the extra domain as the hits, perhaps
250 missed during our first iteration because of stringent r.m.s.d. cutoff.

251

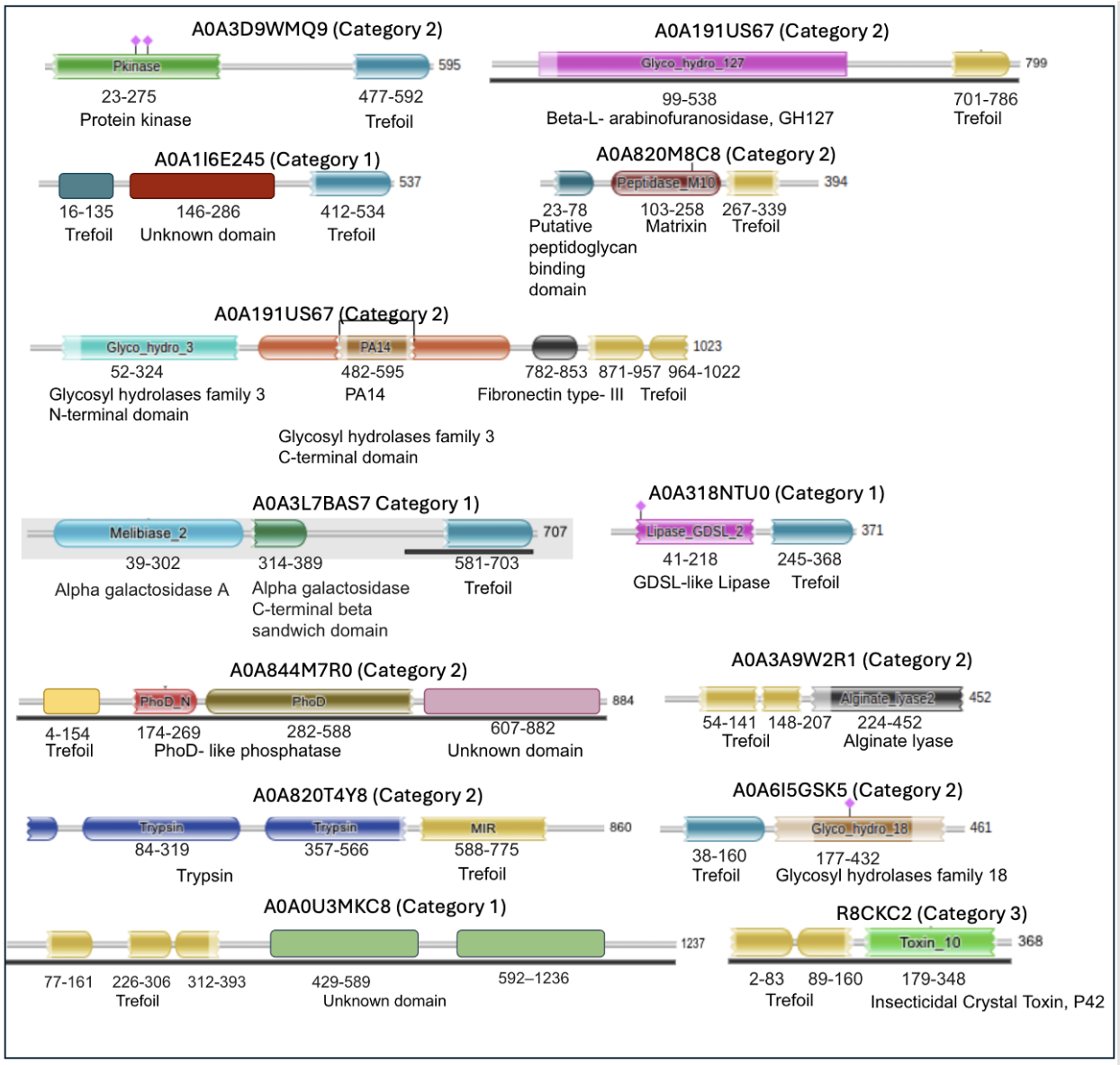


Figure 3: Novel domain architecture of UniProt proteins containing trefoil domains predicted by AlphaFold that are not yet represented in the PDB. A tabular format is included in Supplementary Table 1.

Among the 13 hits from AFdb, we found four category-1 and eight category-2 hits as shown in Figure 3 and listed in Supplementary Table 1. Figure 4a, 4c, 4e, 4f, 5c, 5d, 5e and 5f shows hits belonging to category-2. A0A6I5GSK5, which is annotated as chitinase sharing its chitinase

260 domain with 2dsk (r.m.s.d 1.342 Å), lacks trefoil domain even in the corresponding full-length
261 proteins (WP_317259669.1) with 742 amino acids. A0A4Q2KX23, sharing its beta-glucosidase
262 domain with 4i3g (r.m.s.d 1.016 Å) lacks trefoil domain in the corresponding full-length protein
263 (WP_230883105.1) with 832 amino acids. A0A3D9WMQ9 shares its protein kinase domain with
264 3f61 (1.95 Å), which lacks trefoil domain in the corresponding full-length protein
265 (WP_057333683.1) of length 626 amino acids. A0A820M8C8 shares its metallopeptidase domain
266 with 1slm (0.824 Å) lacking trefoil domain in the corresponding full-length protein (NP_002413.1)
267 with 477 amino acids. A0A191US67 shares its Beta-L- arabinofuranosidase domain with 4qjy
268 (2.83 Å) lacking trefoil domain in the corresponding full-length protein (WP_371878443.1) with
269 642 amino acids. A0A844M7R0 shares its phoD domain with 2yeq (0.741 Å) lacking trefoil
270 domain in the corresponding full-length protein (WP_009969242.1) with 583 amino acids.
271 A0A820T4Y8 shares the trypsin domain with 7n7x (0.628 Å) lacking trefoil domain in the
272 corresponding full-length protein (XP_054205941.1) with 436 amino acids. A0A3A9W2R1
273 shares the alginate lyase domain with 3zpy (0.976 Å) lacking trefoil domain in the corresponding
274 full-length protein (WP_303396801.1) with 446 amino acids.

275 Figures 4b, 4d, 5a and 5b show category-1 hits. The trefoil domains of A0A318NTU0,
276 A0A1I6E245, A0A3L7BAS7 and A0A0U3MKC8 gives an r.m.s.d. of 0.795 Å, 1.867 Å, 0.686
277 Å and 0.888 Å against a representative trefoil domain to confirm the presence of this domain in
278 these novel hits. The non-trefoil domains of these proteins there is no representation in the PDB.
279 Perhaps the annotation for these domains in UniProt is based on the presence of the annotated
280 domain in PFam.
281

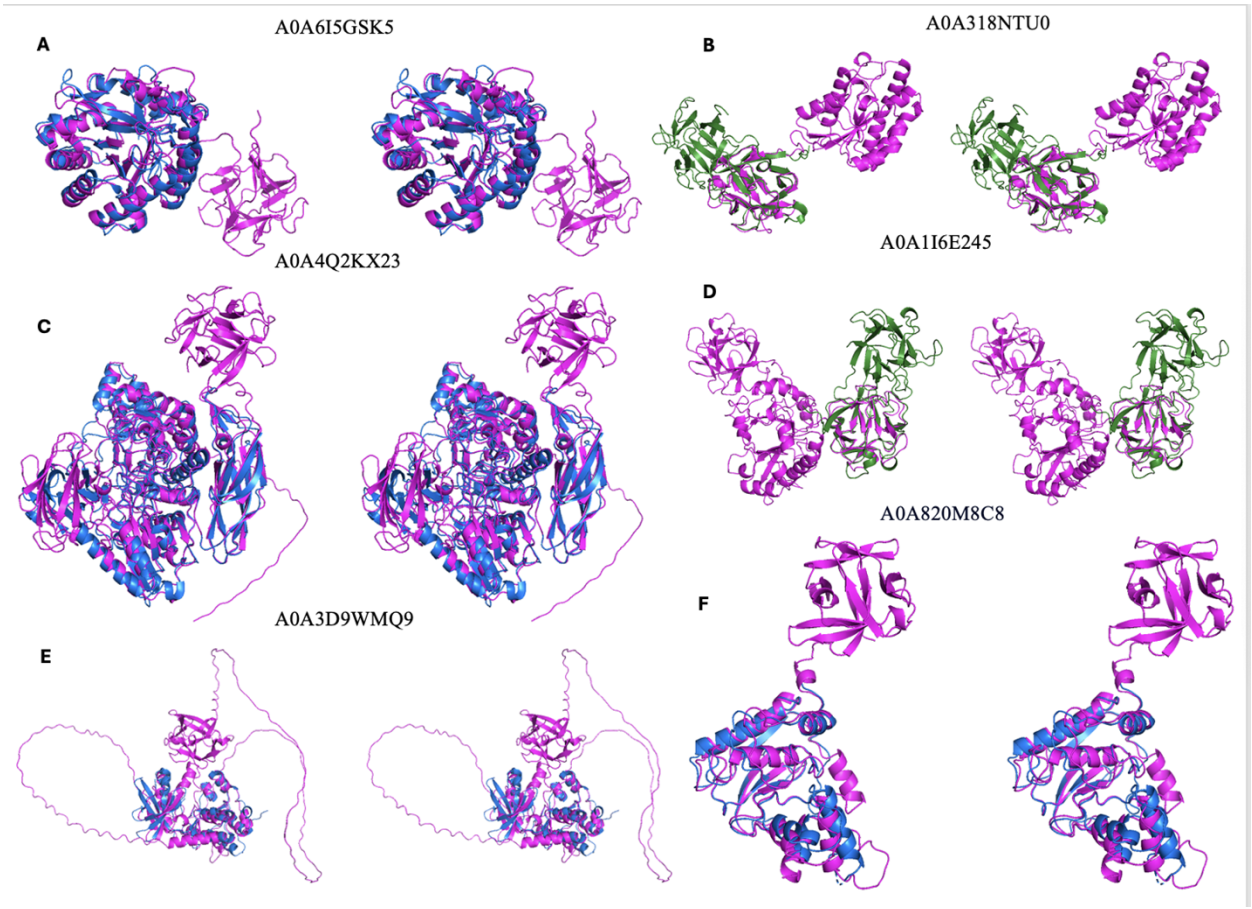


Figure 4: Stereo diagram of the six characterized UniProt predicted structures not represented in PDB. PDB queries in B and D representing category-1 are in green, AFdb structures in pink, PDB hits for categories-2 and 3 in A, C, E and F are in blue.

Figure 5g shows the structure of R8CKC2, category-3, superimposed on the full structure of 8bad annotated as a toxin from *Bacillus thuringiensis* with an r.m.s.d. of 0.976 Å. An enlarged view of all proteins in Figure 4 and Figure 4 are shown individually in stereo figures in Supplementary Figure 3a-f and Supplementary Figure 4a-g.

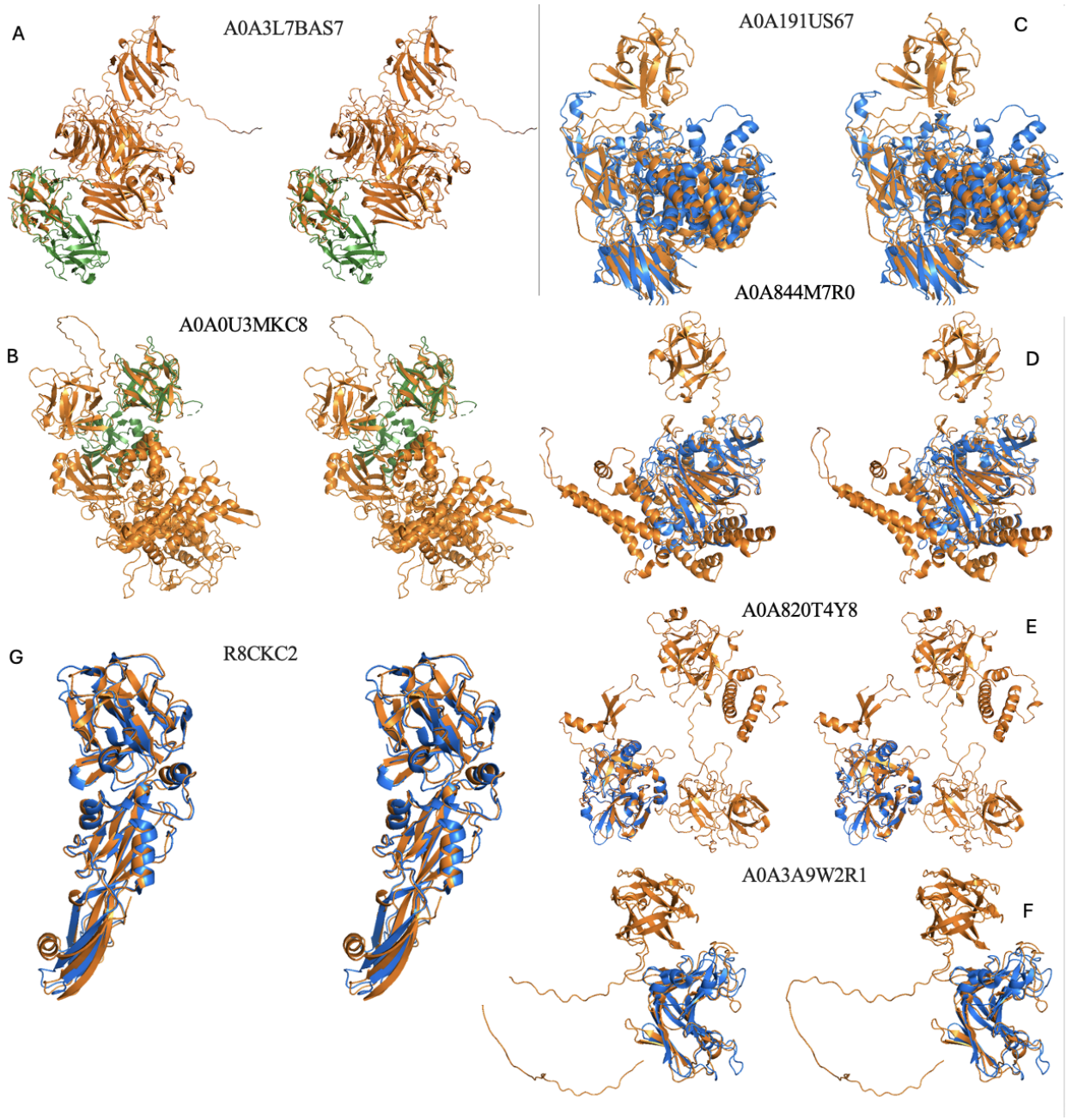


Figure 5: Stereo diagram of the seven uncharacterized proteins from AFdb containing trefoil domain not part of the functions and domain architecture presented in Figure 1. PDB queries in A and B representing category-1 are in green, AFdb structures in orange, PDB hits for categories-2 and 3 in C through G in blue.

DISCUSSION

298 The incomprehensible convergence in fold for billions of proteins, divergent at the
299 sequence-level, can now be used in functional characterization of thousands of proteins under each
300 organism using AFdb. Here, by identifying novel domain architecture in the AFdb structures
301 containing the trefoil fold, one of the primordial protein folds, we demonstrate the usefulness of
302 fold prediction in exploring functional diversity.

303 While the SCOP database represents eight functional diversities among proteins containing
304 the trefoil domain, our search in the PDB revealed 19 distinct functions for proteins containing the
305 trefoil fold shown in Figure 1. Except for a few, these functions span all kingdoms including plants,
306 bacteria, fungi and Metazoans. This could be an underestimate because of the 2 Å r.m.s.d. cutoff
307 used to limit false positives.

308 The discovery of the functional diversity from PDB is likely underestimated, as the PDB
309 is highly redundant and/or biased towards proteins of therapeutic or economic importance to
310 humans. Therefore, extending our analysis further, we searched the AFdb of over 200 million
311 predicted protein structures using representative PDB structures for each functional class
312 containing trefoil fold (inner ring, Figure 1). Not surprisingly, as shown in Figure 2, among the
313 AFdb structures from a given PDB query with 80% coverage, there is a clear inverse correlation
314 between sequence identity and r.m.s.d. for the predicted structures. Interestingly, the steps in the
315 plots of r.m.s.d. against sequence identity in Figure 2, represent a distinct class or order or
316 kingdom. In that, AlphaFold is a mega homology modeling tool on the one end and a novel fold
317 prediction tool on the other.

318 In order to look for novel functional architecture among the trefoil containing proteins in
319 AFdb, we interrogated hits with greater than 30% query coverage. Note that this cutoff leads to
320 false negatives by eliminating large proteins in which trefoil domains may span less than 30%. For

321 example, proteins with domain structures like 6mu1 (Supplementary Figure 6) where the trefoil
322 region spans only 15% of the entire structure. As mentioned in the results section, the hits are
323 binned into various categories depending on the presence of trefoil and/or other domains in PDB.

324 Among the proteins from AFdb containing trefoil fold, we discovered novel domain
325 architectures where the trefoil fold is found along with an additional domain that is either not yet
326 represented in the PDB or the complete domain architecture represents a novel domain
327 arrangement that is yet to be structurally characterized. With the exception of A0A820T4Y8, all
328 the novel proteins containing a trefoil domain identified here are from culturable bacteria and are
329 amenable to functional characterization using experimental techniques. For example,
330 A0A3L7BAS7 is a protein from *Micromonospora sp.* CV4, A0A191US67 is a protein from
331 *Streptomyces parvulus*, A0A844M7R0 is a protein from *Scytonema sp.* UIC 10036
332 (cyanobacteria), R8CKC2 from *Bacillus cereus*, A0A820T4Y8 from *Rotaria sp. Silwood2*
333 (Metazoa) and A0A0U3MKC8 from *Roseateles depolymerans*.

334 It should be mentioned that FoldSeek is able to find homologs of trefoil across diverse
335 kingdoms and functions, which is not possible using sequence-based annotation especially of folds
336 like trefoil (work not published). On the other hand, the extra reliance on Foldseek algorithm for
337 this project may also have limited the discovery of novel proteins containing trefoil fold, because
338 Foldseek reduces the structural elements into alphabets, which may still be biased toward sequence
339 similarity for short range interaction. For example, the performance of Foldseek to find structural
340 homologs of helical cytokines (results not shared) was poor. On the other hand, Foldseek performs
341 better with folds containing beta sheets, such as the TNF fold family (manuscript in preparation).
342 We believe that this is because helical cytokines are overrepresented on short-range interactions
343 compared to long-range interactions, which is the opposite for beta-sheet proteins.

344 Trefoil folds appear to be highly tolerant of amino acid changing mutations even within
345 the core. For example, paralogs within the same genome loci binding to the same receptor show
346 no significant sequence homology. For example, the function of IL1RA, an interleukin-1 receptor
347 antagonist, could only be predicted using 1D-3D threading approaches[17]. The work reported here
348 shows that the trefoil fold is present in all kingdoms with diverse domain architecture.
349 Furthermore, the diversity in functions of trefoil domain containing proteins is by binding to
350 diverse macromolecules including, DNA, carbohydrate, ig-like receptor domains, proteases,
351 chlorophyll, actin and by large oligomerization forming pores. In order to find out if beta-trefoil is
352 a converging or diverging evolution, we generated a phylogenetic tree using structure-based
353 sequence alignment for all proteins shown in Supplementary Figure 7. In divergent evolution one
354 would expect beta-trefoil containing proteins within the same functional family to cluster together
355 irrespective of kingdom. As shown in the phylogenetic tree (Supplementary Figure 7) these
356 proteins actually cluster by kingdom and not by function, suggesting convergent evolution. The
357 clustering by kingdom along with the observed diversity in delivering function via binding to
358 diverse macromolecules, it is tempting to hypothesize that this fold may have evolved
359 independently several times. If true, fold-based function prediction will hit a twilight zone just like
360 sequence-based function prediction. However, based on the amino acids at the equivalent positions
361 forming the core, shown in Supplementary Figure 8, the jury is still out on this hypothesis

362 # METHOD

363 In Figure 6 we show the flowchart of the approach taken here. The r.m.s.d. cutoff of 2 Å
364 was used during the first round of searching PDB database for trefoil containing proteins. The

365 cutoff was relaxed to 3 Å while searching the AFdb with a query containing multiple domains.

366 Nevertheless, these cutoffs are stringent and may have led to false negatives.

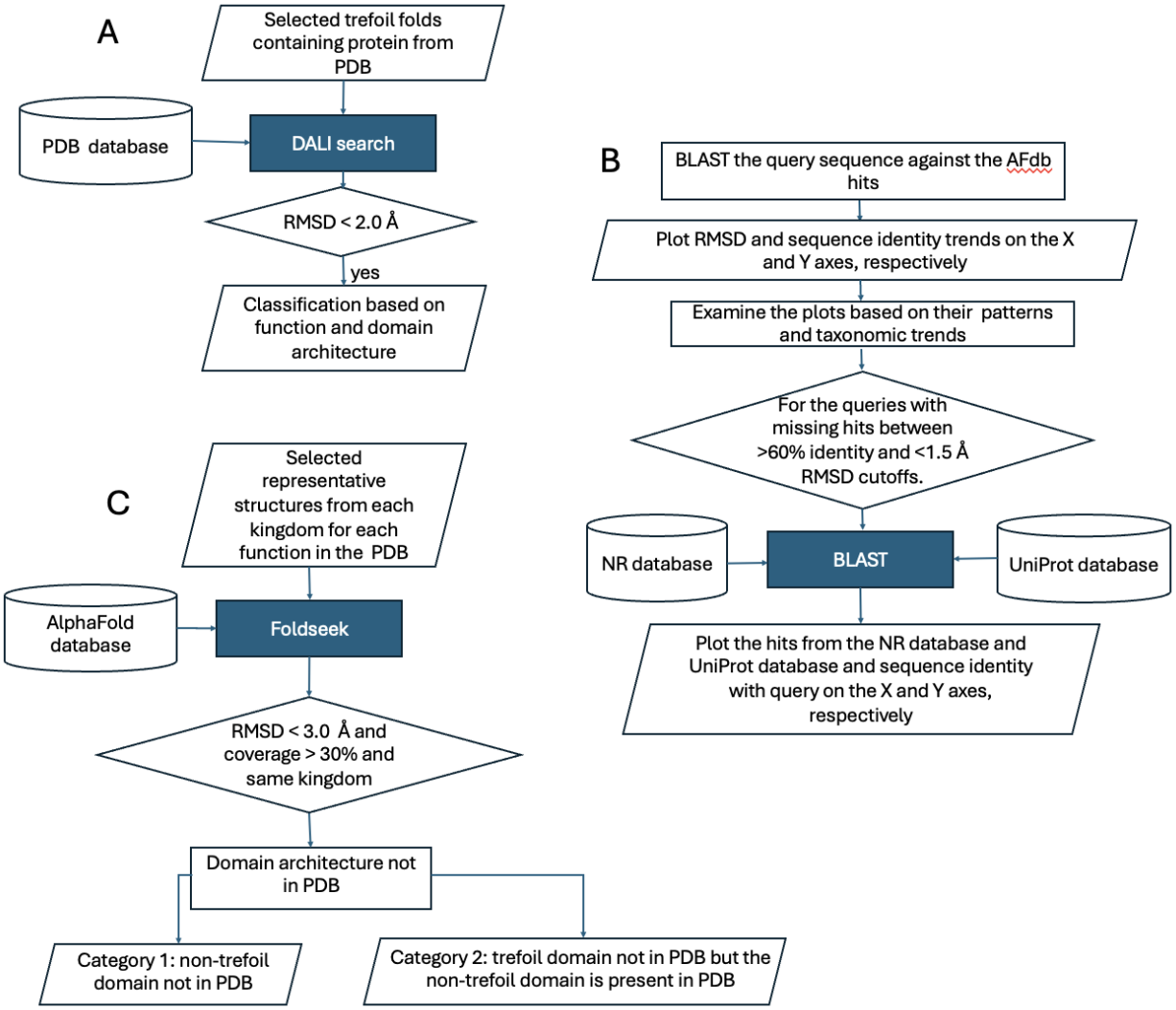
367

368 **Structure-based Functional Genomics**

369 As shown in Figure 6a a few sample structures containing trefoil domain from the PDB database

370 were searched against PDB using DALI for all known structures containing trefoil fold. A filter of

371 2 Å was used for functional classification.



372

373 Figure 6: Flowchart of various steps along with filtering criteria

374

375 **Validation of our approach**

376 As shown in Figure 6b, the accepted hits for a given query against AFdb with a query coverage of
377 80% and more is plotted using r.m.s.d. against sequence identity ranging from 30% to 100%.

378

379 **Identification of novel protein containing trefoil fold from PDB**

380 As shown in Figure 6c, an expanded list of representative structures were selected from each
381 kingdom representing a given function including those containing multiple domains. These were
382 used as query sequences to search AFdb using Foldseek with 3Di option and a cut of 3 Å was used
383 to select hits for downstream analysis. These hits included characterized and uncharacterized
384 proteins based on existing annotation and were classified as various categories based on the
385 existence of 3D structure of individual domains in PDB.

386 **ACKNOWLEDGEMENT**

387 The authors thank GoK for funding for data analysis personnel via BioIT grant and computing
388 infrastructure via the Department of IT, BT and ST.

389 **CONFLICT OF INTEREST**

390 Authors have no conflict of interest.

391 **AUTHOR CONTRIBUTION**

392 MG: The entire bioinformatics work, SS: for conceptualization, guiding bioinformatics work and
393 writing of the manuscript

DATA AVAILABILITY

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

BIBLIOGRAPHY

1. Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Dayhoff, M., Schwartz, R. & Orcutt, B. *Atlas of Protein Sequence Nd Structure*. vol. 5 (1978).
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
5. Eddy, S. R. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365 (1996).
6. Hameduh, T., Haddad, Y., Adam, V. & Heger, Z. Homology modeling in the time of collective and artificial intelligence. *Comput. Struct. Biotechnol. J.* **18**, 3494–3506 (2020).
7. Simons, K. T., Strauss, C. & Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199 (2001).
8. Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D. & Bates, P. A. Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes. *Annu. Rev.*

416 *Biophys.* **52**, 183–206 (2023).

417 9. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based
418 residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad.*
419 *Sci. U. S. A.* **110**, 15674–15679 (2013).

420 10. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of
421 methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).

422 11. Ma, W. *et al.* Enhancing Protein Function Prediction Performance by Utilizing AlphaFold-
423 Predicted Protein Structures. *J. Chem. Inf. Model.* **62**, 4008–4017 (2022).

424 12. Zhao, C., Liu, T. & Wang, Z. PANDA-3D: protein function prediction based on AlphaFold
425 models. *NAR Genomics Bioinforma.* **6**, lqae094 (2024).

426 13. Wang, Z., Zhao, C., Wang, Y., Sun, Z. & Wang, N. PANDA: Protein function prediction
427 using domain architecture and affinity propagation. *Sci. Rep.* **8**, 3484 (2018).

428 14. Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature*
429 **622**, 646–653 (2023).

430 15. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein
431 universe. *Nature* **622**, 637–645 (2023).

432 16. Tenorio, C. A., Longo, L. M., Parker, J. B., Lee, J. & Blaber, M. Ab initio folding of a
433 trefoil-fold motif reveals structural similarity with a β-propeller blade motif. *Protein Sci.*
434 *Publ. Protein Soc.* **29**, 1172–1185 (2020).

435 17. Bazan, J. F., Timans, J. C. & Kastelein, R. A. A newly defined interleukin-1? *Nature* **379**,
436 591 (1996).

437 18. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends*
438 *Biochem. Sci.* **20**, 478–480 (1995).

19. Fast and accurate protein structure search with Foldseek - PubMed. https://pubmed.ncbi.nlm.nih.gov/37156916/.