*NAR Genomics and Bioinformatics*, 2024, **6**, 1–11
https://doi.org/10.1093/nargab/lqae005
Advance access publication date: 31 January 2024
**Standard Article**

# Functional domain annotation by structural similarity

**Poorya Mirzavand Borujeni[1] and Reza Salavati[1,2,*]**

[1]Institute of Parasitology, McGill University, Ste. Anne de Bellevue, Quebec H9X 3V9, Canada
[2]Department of Biochemistry, McGill University, Montreal, Quebec H3G 1Y6, Canada

[*]To whom correspondence should be addressed. Tel: +1 5143987721; Email: reza.salavati@mcgill.ca

## Abstract

Traditional automated *in silico* functional annotation uses tools like Pfam that rely on sequence similarities for domain annotation. However, structural conservation often exceeds sequence conservation, suggesting an untapped potential for improved annotation through structural similarity. This approach was previously overlooked before the AlphaFold2 introduction due to the need for more high-quality protein structures. Leveraging structural information especially holds significant promise to enhance accurate annotation in diverse proteins across phylogenetic distances. In our study, we evaluated the feasibility of annotating Pfam domains based on structural similarity. To this end, we created a database from segmented full-length protein structures at their domain boundaries, representing the structure of Pfam seeds. We used *Trypanosoma brucei,* a phylogenetically distant protozoan parasite as our model organism. Its structome was aligned with our database using Foldseek, the ultra-fast structural alignment tool, and the top non-overlapping hits were annotated as domains. Our method identified over 400 new domains in the *T. brucei* proteome, surpassing the benchmark set by sequence-based tools, Pfam and Pfam-N, with some predictions validated manually. We have also addressed limitations and suggested avenues for further enhancing structure-based domain annotation.

## Introduction

Functional annotation is a critical process that seeks to delineate the identity of a protein in three distinct dimensions: its final location, its function, and the biological processes in which it participates. Experimental functional annotation can be time-consuming and expensive, underscoring the significant importance of automated *in-silico* functional annotations. These tools offer the convenience of high-throughput analyses without the need for manual inspection. Currently, most automated *in silico* annotation tools rely on sequence similarity to manually annotated proteins. Among them, annotations based on 1:1 orthology and InterPro2GO are particularly prevalent (Gene Ontology Consortium website). In 1:1 orthology-based annotation, sequence search tools like BLASTP are primarily used to infer orthology relationships. InterPro2GO, the most widely adopted automated annotation tool (Gene Ontology Consortium website), identifies different sequence signatures in the query protein using InterProScan (1). The identified sequence signatures will be translated into GO terms using the InterPro2GO table, a curated table showing the associations between sequence signatures and the functional annotation of the proteins.

On the other hand, protein structure is shown to be three to ten times more conserved than its sequence (2). Consequently, it is anticipated that a larger number of proteins could potentially be annotated by integrating protein structure into the annotation process. Some studies have indeed succeeded in using protein structures to annotate domains of proteins that could not be annotated by sequence alone (3–5). However, these structure-based approaches have not been broadly adopted. One reason is the experimental resolution of protein structures has predominantly focused on well-studied proteins. Additionally, previous structure prediction tools were not sufficiently accurate, which made the prediction of unannotated protein structures challenging. Despite these limitations, these prediction tools have demonstrated potential for contributing to functional annotation (6,7).

Until recently, the structure-based functional annotation has not been the focal point for large-scale annotations mainly since only a limited number of protein structures had been resolved experimentally, and computational predictions were not sufficiently accurate. The introduction of AlphaFold2 sparked a revolution in predicting protein structure. AlphaFold2 is capable of predicting protein structures with an accuracy similar to experimental methods (8,9), thus opening new avenues in the field of functional annotation.

With protein structures predicted, the next step is to identify proteins with similar structures. Although advanced tools like DALI (10) and TM-align (11) are available for this purpose, they can be computationally intensive when used to search comprehensive structure databases. RUPEE (12) is a faster tool, but it only looks for structural similarity and not sequence similarity. Foldseek, a newly developed structure search tool, can find structurally similar proteins significantly faster than other tools (13).

The combination of AlphaFold2 and Foldseek may facilitate the annotation of a greater number of proteins. A recent study demonstrated that in the proteomic comparisons of species that are evolutionarily distant, there can be instances where the reciprocal best matches cannot be identified using sequence similarity methods. However, these matches can be detected when reciprocal top structural correspondences are considered (14). In a separate study, Ruperti et al. employed Foldseek to identify the closest structural match in model organisms to the query proteins and transferred the annotations from the top hits (15). Their findings were remarkable, as the

**Received:** September 12, 2023. **Revised:** December 3, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 15, 2024

use of Foldseek instead of sequence alignment tools could lead to the annotation of up to 50% more genes. Bordin et. al also used Foldseek and SSAP to annotate the CATHe domains of 21 model organisms and identify new structural domains in them (16).

As previously mentioned, InterPro2GO uses sequence-based domain annotation by utilizing multiple member databases within InterPro. Each database holds a distinct type of sequence signature. For example, Pfam, one of the most widely used members, holds the sequence signature of protein functional domains (17). These sequence signatures are extracted from multiple sequence alignments of well-recognized domain instances known as Pfam seeds. Pfam employs Hidden Markov Models (HMM) to model these signatures, which can then be searched against proteins of interest using the hmmscan module of HMMER (18). It is worth mentioning that Pfam also post-processes the hits and its post-processing might change the importance order of the hits reported by HMMER.

There are also tools for improving the sensitivity of Pfam annotation. For instance, by creating a profile of query proteins and aligning it with Pfam profiles, more hits can be identified, although the order of these hits might not necessarily reflect their relevance (19). A recent study found that by employing Convolutional Neural Networks (CNNs) to store sequence signatures, over 9.5% more new domains could be annotated (20). These new annotations are referred to as Pfam-N annotations.

In this study, we aim to explore a novel approach for Pfam domain annotation relying on Protein structure. As Pfam seeds represent a portion of a protein sequence, we created a database of Pfam seeds by dividing the corresponding full-length structures at their domain boundaries. From now on, we will refer to the source protein as the Full-Length Pfam Seed Source (FLPSS). Next, we evaluated the reliability of the structural alignment by aligning the FLPSS with the constructed domain structure database to determine the frequency of structural alignment between different Pfam seeds.

We focused on *Trypanosoma brucei*, an early-diverged eukaryotic parasite, and aligned its structome with our domain database as a case organism. We benchmarked the predicted domains by comparing them to the Pfam v35.0 and Pfam-N predictions as the gold standard. Additionally, we conducted a manual review of some of the domains that were predicted for the proteins involved in *T. brucei's* mitochondrial RNA editing and served as a case study.

## Materials and methods

### Construction of a Pfam domains structure database (PfamSDB)

The FLPSSs of Pfam v35.0 were retrieved from the AlphaFold2 database (AlphaFold2 DB, version 4). AlphaFold2 and Pfam use different versions of UniProt. As a result, for fewer than 0.2% of Pfam seeds, the sequence obtained by trimming the FLPSS based on the Pfam seed coordinates did not match the Pfam seed sequence. However, by applying the Needleman-Wunsch alignment method, we successfully mapped the domain locations from the Pfam sequence version to the corresponding sequences in the AlphaFold2 version for approximately two-thirds of these affected seeds. These instances were subsequently removed from the database. In the end, 6.3% of Pfam seeds were not present in the database, either because AlphaFold2 had not predicted their structure (6.2% of instances) or due to version discrepancies in the source sequence between Pfam and AlphaFold2 DB (0.1% of instances).

To develop the PfamSDB, we utilized two approaches. The first, which we will refer to as 'PDB_cut', involved dividing the Protein Data Bank (PDB) files corresponding to the Full-Length Pfam Seed Source (FLPSSs) at the domain borders. The second approach, henceforth known as 'FS_cut', entailed segmenting the Foldseek database files of the FLPSSs at the domain boundaries.

For the process of truncating PDB files, the source codes of the PDB-tools package (21) were modified to do cutting and format conversion to PDB simultaneously. GNU parallel was extensively used for parallelizing the computations. PfamSDB contained over 1.1 million structures.

### Alignment and labeling the alignments

The FLPSSs were aligned with the PfamSDB using Foldseek v8-ef4e960. The number of sequences passing the pre-filtering step was set to a high number by the '–max-seqs 1e9' option to get all possible alignments. A match with the same Pfam domain as the target seed instance was deemed a True Positive (TP), while a match with a different Pfam domain was considered a False Positive (FP). In our scoring system, for a result to be reported as a TP, the seed region on the query must be covered by more than 25% in alignment with an instance of the same Pfam domain. Conversely, if an instance of a different Pfam domain covers more than 25% of the seed region on the query in the alignment, it is considered an FP.

### Annotating *T. brucei* structome and benchmarking against Pfam and Pfam-N

The structome of *T. brucei*, taxonomy id: 185431, UniProt proteome ID: UP000008524, was downloaded from the AlphaFold2 website and it was aligned with PfamSDB using the same parameters as the former step and highest-scoring, non-overlapping hits were selected. By default, Foldseek hits are sorted based on the bitscore, which will henceforth be referred to as 'bits'. To select the hits with the highest 'bits/alnlen' (bitscore/alignment length), the hits of each query were sorted by bits/alnlen, and non-overlapping highest-scoring hits were selected. Our labeling approach was similar to the one employed previously, with one key difference: in this phase, we used regions annotated with either Pfam or Pfam-N as our gold standard for comparison, instead of using the seed regions that served as the gold standard in the previous step. If the domain predicted by the gold standard was not retrieved, it was labeled as False Negative (FN). We used Seaborn (22) and matplotlib (23) for visualization.

### Pfam domain annotation by MMseqs2

To investigate the added value of structural similarity for domain annotation, the same procedure as explained above was done by searching the sequences of *T. brucei* proteome against sequences of PfamSDB by MMseqs2 v14-7e284 (24), the same program used by Foldseek under the hood. The '-s 8.5' parameter was specified to run it with high sensitivity as Foldseek uses MMseqs2 in high sensitivity mode.

### Pfam domain annotation by HMMER

As mentioned earlier, Pfam annotation relies on the post-processing of HMMER hits. Post-processing relies on some curated data such as the score threshold for each Pfam domain. As we did not have such data for Foldseek alignment scores, we also evaluated Pfam domain annotation by simply selecting the best-ranking domains reported by HMMER. In this regard the 'hmmscan' was used for aligning the *T. brucei* proteome with the Pfam database without gathering threshold option. Hits with e-values up to 0.001 were considered for the rest of the analysis.

## Results and discussions

### PfamSDB contains high-confidence short structures

AlphaFold2 reports the estimated confidence level for each residue as the predicted Local Distance Difference Test (pLDDT). The pLDDT value ranges between 0 and 100, and higher values indicate more confident predictions. Early observations have shown that there is a high overlap between residues with low pLDDT and regions known as Intrinsically Disordered Regions (IDRs) that do not fold into specific structures (8). According to the AlphaFold2 website, residues with a pLDDT above 90 are expected to be highly accurate, while regions with a pLDDT between 70 and 90 are considered to have good backbone prediction. If low pLDDT regions correspond to IDRs, we do not expect accurate structural matches for those regions. Furthermore, the significance level of structural alignment hits has been shown to depend on protein length. Monzon et al. have demonstrated that Foldseek has difficulty establishing relationships for some short proteins (those with fewer than 200 amino acids) when identifying the reciprocal best hits (14). In contrast, sequence-based aligners do not exhibit this problem.

As mentioned, pLDDT is reported per residue, however, for the purposes of this discussion, we will refer to the average pLDDT of a region as avg_pLDDT. Figure 1A depicts the average of the avg_pLDDT values of instances of each Pfam domain, while Figure 1B illustrates the distribution of the average size of instances of each Pfam domain. Overall, the instances exhibit a high avg_pLDDT, and the average length of instances across different Pfam domains typically falls below 200. To be more specific, the third quartile (75th percentile) for the average size of instances for each Pfam is 209.

### TPs and FPs can be separated based on bits

We aimed to assess the frequency with which instances of one Pfam domain align to instances of a different Pfam domain, identified as FPs. We then sought to determine if Foldseek probability could distinguish these FPs from TPs. Using Pfam seeds as our gold standard, we aligned the FLPSSs to PfamSDB.

Our data indicates that 75% of the hits were true positives (TPs) in both databases, regardless of whether they originated from alignment with PDB_cut or FS_cut databases. The precision-recall curve of Foldseek probability showed that the recall (or sensitivity) changes less than 0.001 as we adjust the probability thresholds, while the precision changes from 0.75 to 0.80 when we change the Foldseek probability cutoff from its minimum (0.024) to its maximum (1). *F*-measure, the harmonic mean of Precision and Recall, provides a balanced assessment of a classification model's performance. The Foldseek probability set at 1 yielded the highest *F*-measure, which was 0.89 for PDB_cut and 0.88 for FS_cut, respectively.

Foldseek probability is a mapping between the alignment bits and the probability that the query and target belong to the same SCOPe superfamily (Personal communication with Milot Mirdita). We noticed that a Foldseek probability of 1 has been attributed to any alignment with a bits above 100. As the highest *F*-measure for Foldseek probability's performance was achieved when Foldseek probability of 1 was considered, there is a chance that selecting a higher bits as the threshold would lead to a higher *F*-measure. We plotted the precision-recall curve by considering different bits thresholds rather than Foldseek probabilities, shown in Figure 2. The highest *F*-measure achieved was 0.94 for alignments using the PDB_cut database and 0.93 for those utilizing the FS_cut database. Notably, both of these top scores were attained by setting the bits cutoff threshold at 152.

It is worth mentioning that here, we evaluated the precision and recall for all hits and found the optimum bits threshold based on FLPSS-against-PfamSDB alignment. There is also a single mapping between bits and Foldseek probability. However, the Pfam database considers different bits thresholds known as gathering thresholds for different Pfam domains, and the thresholds are manually curated. Although manual curation of the Foldseek bits threshold for each Pfam domain could enhance annotation performance, pursuing such an approach is beyond the purview of our current work.

To address the potential over-restrictiveness of a 152 bits threshold for shorter domains, we evaluated the impact of using a ratio of bits to alignment length on the *F*-measure. As illustrated in Figure 2, employing this ratio resulted in a significantly higher *F*-measure. In all the FLPSS against PfamSDB benchmark tests presented in this manuscript, the performance for classifying TPs from FPs have been comparable for alignments with either PDB_cut or FS_cut database. Our analysis revealed that the optimal *F*-measure for alignments with the FS_cut database was obtained at a bits/alnlen of 0.84, whereas for the PDB_cut database, the optimal ratio was 0.88. Based on these findings, we adopted an average threshold of 0.86, calculated from the optimal ratios for both the PDB_cut and FS_cut databases, as the cutoff for categorizing a hit as positive in subsequent analyses.

As mentioned earlier, we also used MMseqs2, a sequence alignment tool, to see the added value of structural alignment. MMseqs2 does not report a probability score. Our benchmark showed that 98.9% of the labeled alignments were TPs. The precision-recall curve showed the maximum *F*-measure (0.994) is achieved when the minimum bits (36) is used. So, we did not select any threshold on MMseqs2 hits.

### Short or low-confidence instances less likely to identify same Pfam domain matches

For each instance, the maximum number of TPs is equal to the number of instances in the same Pfam domain, a value we denote as $N$. Therefore, the maximum number of TPs for all instances of a particular Pfam domain amounts to $N^2$. We then define the 'proportion of retrieved instances' as the ratio of the 'total number of times instances of each Pfam domain were labeled as TP' to $N^2$. Figure 3 elucidates the relationship between this proportion of retrieved instances and two key parameters: the Average Length and the Average of 'avg_pLDDT of instances', per Pfam domain. These figures

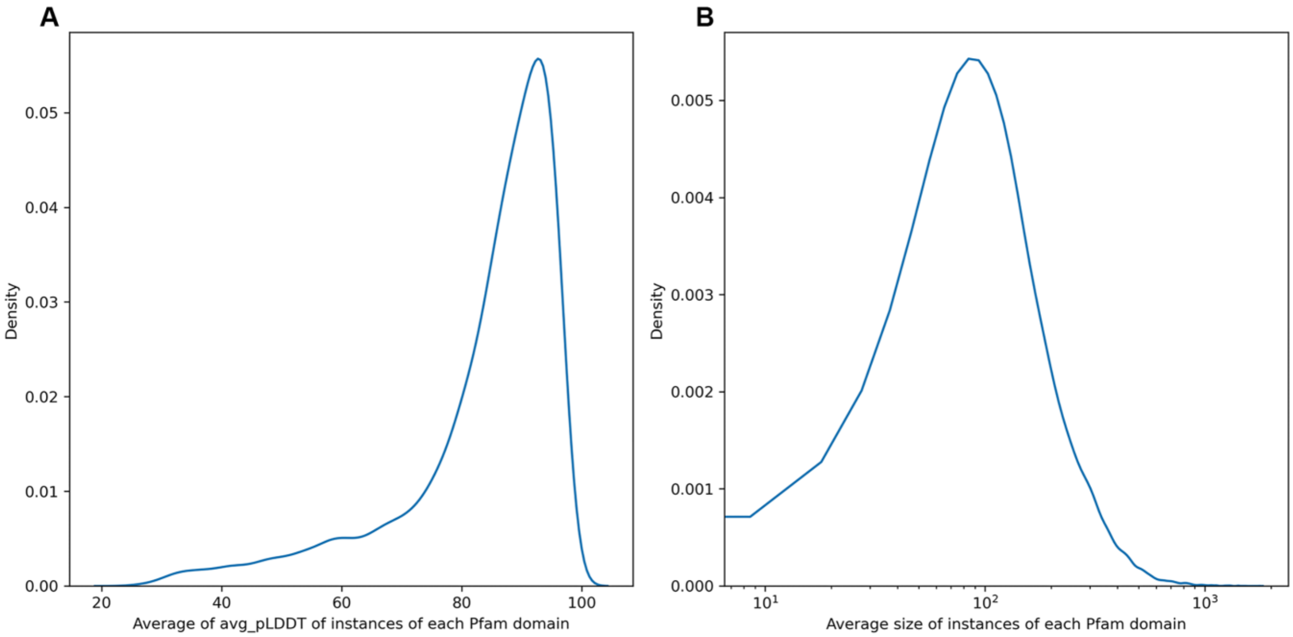

**Figure 1.** The kernel density estimation of the average of avg_pLDDT (**A**) and average of size of instances (**B**) of each Pfam domain.

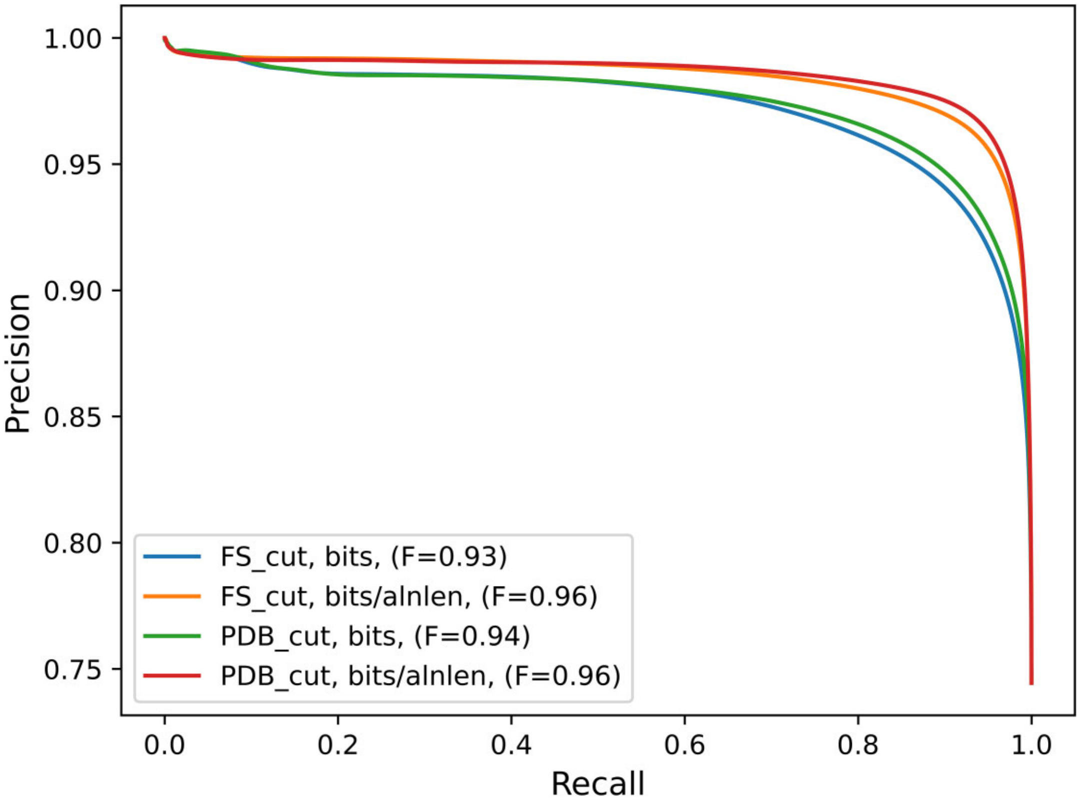

**Figure 2.** Precision–recall curves for different alignment criteria. The graph shows the performance of alignment with PDB_cut or FS_cut databases using either bits or bits/alnlen as the scoring condition. *F*-measure has been indicated for each.

were produced using the FS_cut database, and it is noteworthy that alignments with the PDB_cut database exhibited a similar trend. According to Figure 3, Pfam domains with longer average lengths and higher avg_pLDDT generally display a higher propensity to retrieve all instances of the same Pfam domain from the query. This observation is in line with the findings of (14), where short proteins and proteins rich in residues with low pLDDTs exhibited a reduced likelihood of identifying their reciprocal best hits when examining the reciprocal best structural hits across two organisms' proteomes.

In certain Pfam domains, instances, despite having a substantial length (>100) and a high avg_pLDDT (>80), failed to

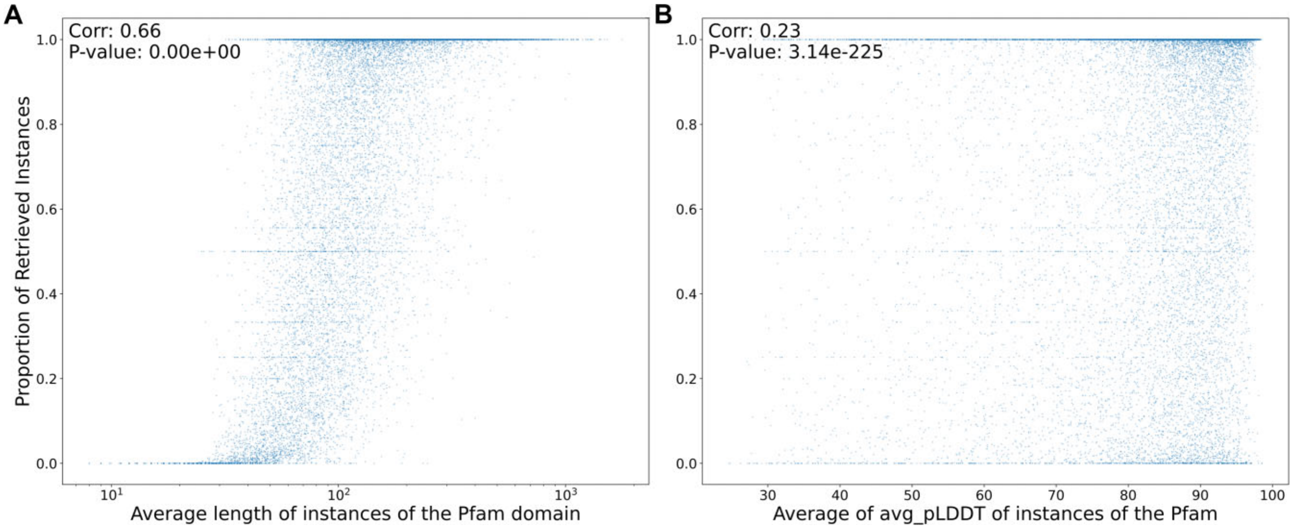

**Figure 3.** (**A**) Relationship between the proportion of retrieved instances and the average length of instances for each Pfam domain. (**B**) Relationship between the proportion of retrieved instances and the average avg_pLDDT of instances for each Pfam domain. The Spearman correlation coefficient, depicted in the upper left corner of each subfigure, provides a measure of the strength and direction of the relationship between the axes of each plot. These plots were created from the output of alignment with the FS_cut database.

retrieve any matches within the same Pfam domain - not even themselves. These particular instances predominantly exhibited low-complexity structures characterized by a profusion of helices. Further analysis revealed that the 3Di transformations of these structures also manifested this low complexity; most were sequences representing a single 3Di state. The 3Di structural alphabet, utilized by Foldseek, provides a representation of tertiary interactions between residues in a protein's spatial configuration. This offers enhanced information density and fewer false positives than traditional backbone structural alphabets (13). During prefiltering, Foldseek identifies matches with 3Di sequences similar to the query protein using sequence alignment. Notably, Foldseek's logs indicate that it automatically masks low-complexity regions during the 3Di state prefiltering. This likely explains the inability of low-complexity regions to align with related matches.

## Domain prediction: comparable in count to Pfam v35

We aligned the *T. brucei* proteins with the Pfam instances database, selecting non-overlapping hits as domain annotations. Figure 4 illustrates the number of domains annotated using various approaches according to which the number of domains predicted by Foldseek with the FS_cut database, selecting the highest passing bits threshold, is comparable to that of Pfam v35 and exceeds MMseqs2 predictions by 21%. In all cases, the number of domains predicted by using FS_cut database was higher than those predicted by using the PDB_cut database. Also, the use of bits as the ranking criteria ended up with more domain annotations.

Table 1 presents precision and recall at both Pfam and Clan levels, using either Pfam or Pfam-N as the gold standard. The Pfam database groups different Pfam domains with similar sequence signatures into the same Clan and Clan-level statistics are also depicted in Table 1. In all cases, the use of FS_cut database instead of PDB_cut database improves both precision and recall. Use of bits/alnlen as the ranking criteria improves the precision but will end up with a lower recall. In all scenarios, Clan level precision exceeds 90%. This suggests that even when a Pfam domain identical to the gold standard may not be predicted based on structure, a quite similar Pfam domain is often attributed to the same region.

Table 1 shows that when using Foldseek for domain annotation, recall is comparable to MMseqs2. Yet, previous studies have shown that Foldseek substantially outperforms BLASTP, a comparable sequence-based tool. For instance, Monzon *et al.* employed both Foldseek and BLASTP to determine the reciprocal best relationships among multiple model organisms (14). Notably, since Foldseek utilizes MMseqs2 in its high sensitivity mode, it is crucial for a balanced comparison of sequence and structure-based alignment to also run MMseqs2 in this mode. Endeavoring to provide such a comparison, we examined the reciprocal best relationships among the organisms featured in Monzon *et al.*'s study. Our results indicate that by using MMseqs2 in its high sensitivity setting (-s 8.5), the reciprocal best relationships for an additional 200 proteins can be identified, specifically when analyzing the proteomes of *Drosophila melanogaster* and *Homo Sapiens*. Despite this, the majority of statistics from the Monzon *et al.*'s study remained consistent with our experiment. Furthermore, while running MMseqs2 in high-sensitivity mode did yield more 'reciprocal best matches', Foldseek still significantly outstripped MMseqs2 in terms of established relationships. This implies that the modest edge Foldseek has over MMseqs2 in domain annotation might be influenced by the distinct biological questions under investigation.

To assess the impact of choosing a threshold on either bits or bits/alnlen on the domain annotation, we also evaluated the precision and recall without imposing any threshold. As detailed in Supplementary Table S1, the observed increase in precision when thresholds are applied is at least twice the magnitude of the decrease in recall. These findings suggest that employing the optimum thresholds, as calculated from FLPSS-against-PfamSDB alignments, can enhance *T. brucei* domain annotation.

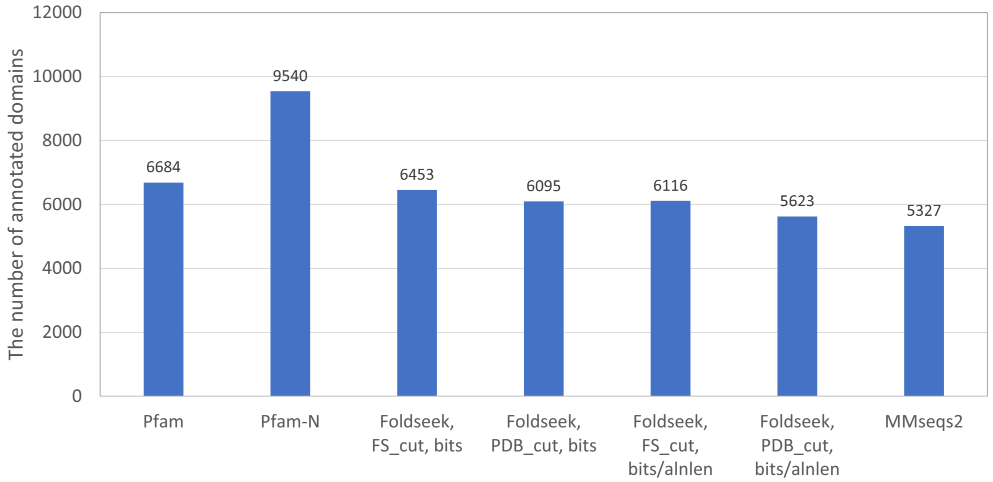

**Figure 4.** The number of domains annotated in different approaches. For the Foldseek method, both the database utilized and the ranking criteria are detailed.

**Table 1.** Precision and recall of domain annotations by MMseqs2 and Foldseek, using either Pfam or Pfam-N as the gold standard

| | | Gold standard: Pfam | | | | Gold standard: Pfam-N | | | |
|---|---|---|---|---|---|---|---|---|---|
| Alignment method | Ranking criteria | Precision Pfam-level | Recall Pfam-level | Precision Clan-level | Recall Clan-level | Precision Pfam-level | Recall Pfam-level | Precision Clan-level | Recall Clan-level |
| Foldseek, FS_cut | bits | 0.879 | 0.697 | 0.987 | 0.783 | 0.776 | 0.49 | 0.955 | 0.603 |
| Foldseek, PDB_cut | bits | 0.857 | 0.655 | 0.982 | 0.75 | 0.752 | 0.455 | 0.953 | 0.576 |
| Foldseek, FS_cut | bits/alnlen | 0.907 | 0.679 | 0.989 | 0.741 | 0.823 | 0.458 | 0.977 | 0.543 |
| Foldseek, PDB_cut | bits/alnlen | 0.896 | 0.637 | 0.987 | 0.702 | 0.815 | 0.423 | 0.976 | 0.507 |
| MMseqs2 | bits | 0.953 | 0.68 | 0.992 | 0.708 | 0.897 | 0.407 | 0.988 | 0.448 |

## Domain length impacts the recall significantly

Figure 5 illustrates the analysis of Pfam domains using Foldseek, MMseqs2 and HMMER. This figure particularly highlights the results from Foldseek alignment with FS_cut, plotting alignments under two distinct ranking criteria: bits and bits/alnlen. The observed trends align with those seen in the PDB_cut database (data not shown). Comparing Figure 5B and 5D directly, it becomes evident that for domains with lengths in the range of 40–50 amino acids, the false positive rate is reduced when ranking hits based on the bits/alnlen criterion.

In Figure 5A and 5C, domains located in areas with low pLDDT scores show reduced detection by Foldseek, indicated by a lower recall rate. Notably, MMseqs2 (Figure 5C) also displays low recall but high precision for domains in low pLDDT regions. This can be attributed to the fact that IDR regions exhibit low sequence similarity (resulting in low recall), but sequences resembling an IDR are indeed IDRs (resulting in high precision).

Figure 5B further demonstrates that the same observation holds true for the 'domain size,' with small domains having a notably low recall rate. A direct comparison between Figure 5B with Figure 5F and also Figure 5D with Figure 5F reveals that Foldseek does not enhance recall when retrieving small-sized domains, relative to MMseqs2; in fact, precision is even lower with Foldseek. This phenomenon can be attributed to the tendency of short domains not to fold into unique structures. Consequently, their structural resemblance to instances of different Pfam domains may increase the chance of alignment with a non-corresponding Pfam domain instance, thereby reducing precision. The same description can explain why the comparison of Foldseek and BLASTP for finding the reciprocal best hits between two organisms shows that many reciprocal best hits that are exclusively found by BLASTP, are <200 amino acids long (14).

Alignment bits depends on the length of the alignment and we expect that short domains would align with a lower bits even by HMMER, the program used by Pfam for domain annotations. However, since Pfam uses domain-specific bits thresholds, this can aid in annotating short domains. Indeed, Figure 5H shows that the HMMER top hit selection without considering the gathering threshold results in lower Precision for shorter domains. However, the reduction is less significant than Foldseek and MMseqs2.

Supplementary Figure S1 supports these observations, showing that similar patterns persist when Pfam-N is considered the gold standard. Additionally, Supplementary Figure S1 indicates that when Pfam-N is the gold standard, the precision of predictions for domains located in regions with low avg_pLDDT is lower than that of domains in high avg_pLDDT regions. The same figure also shows that the

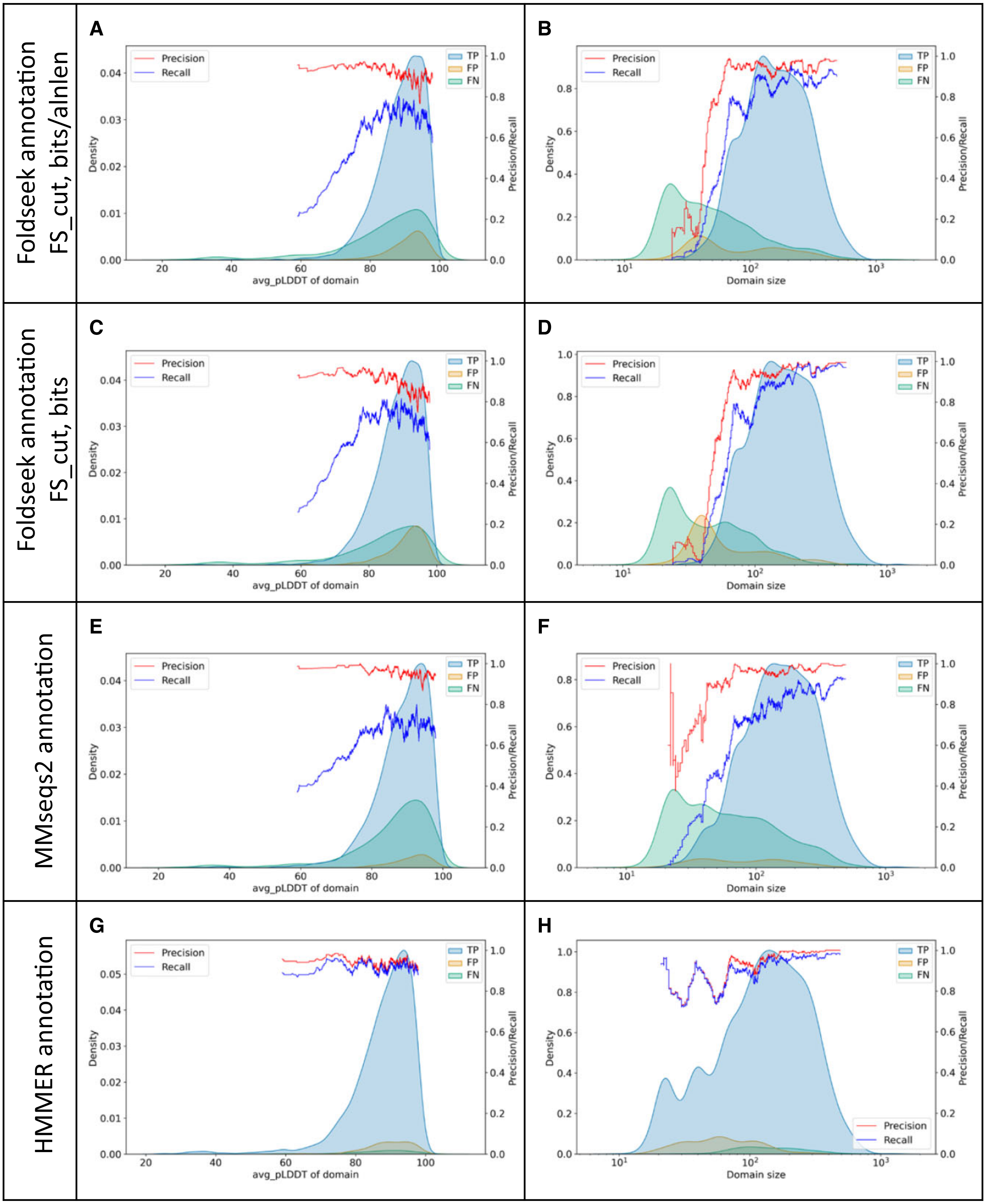

**Figure 5.** Analysis of Pfam domains: Foldseek (using FS_cut database), MMseqs2 and HMMER annotations. (**A**) avg_pLDDT distribution for Foldseek domains (hits ranked by bits/alnlen). (**B**) Size distribution of Foldseek domains (hits ranked by bits/alnlen). (**C**) avg_pLDDT distribution for Foldseek domains (hits ranked by bits). (**D**) Size distribution of Foldseek domains (hits ranked by bits). (**E**) avg_pLDDT distribution for MMseqs2 domains. (**F**) Size distribution for MMseqs2 domains. (**G**) avg_pLDDT distribution for HMMER domains. (**H**) Size distribution for HMMER domains. Distributions are colored according to their relationship with the gold standard (true positive (TP), false positive (FP), false negative (FN)). Precision and recall are calculated using a rolling method for every 200 consecutive hits.

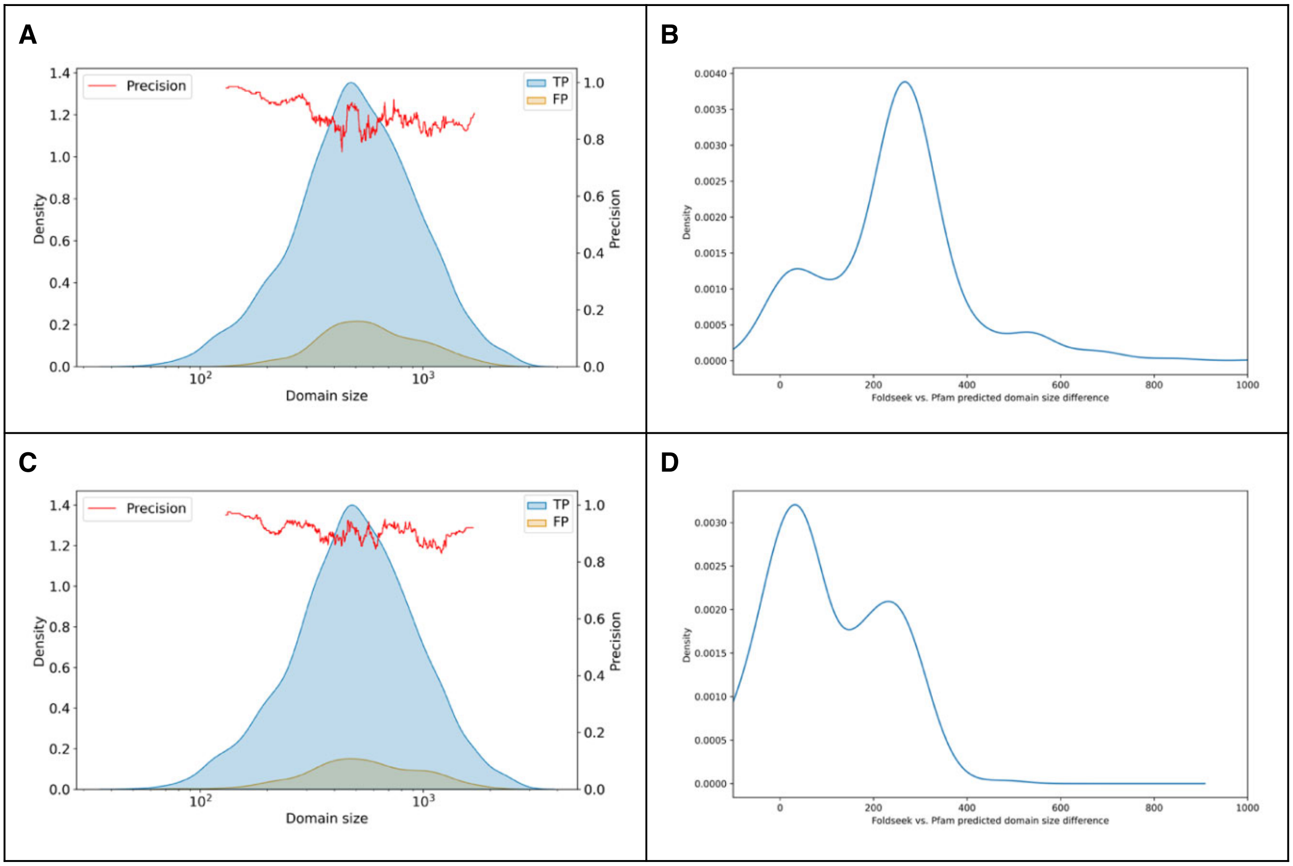

**Figure 6.** Analysis of Foldseek domains by Pfam. (**A**) The distribution of length of Foldseek domains (hits ranked by bits) categorized by their relationships to the Pfam annotations (true positives (TP), false positive (FP)). (**B**) The difference between the size of the false positive Foldseek domains (hits ranked by bits) and the Pfam domains predicted for the same region. (**C**) The distribution of length of Foldseek domains (hits ranked by bits/alnlen) categorized by their relationships to the Pfam annotations. (**D**) The difference between the size of the false positive Foldseek domains (hits ranked by bits/alnlen) and the Pfam domains predicted for the same region. Foldseek domains longer than 100 amino acids have been used for parts B and D.

HMMER top hit selection has low precision and recall for retrieving short Pfam-N domains.

The Spearman correlation coefficient between avg_pLDDT and domain size for Pfam and Pfam-N domains is -0.07 and 0.07 respectively, showing that there is no meaningful relationship between them.

While Figure 5D shows that when hits are ranked by bits, the precision and recall for annotating long domains is relatively high compared to short domains, Figure 6A shows that long domains attributed by Foldseek have a slightly lower precision than shorter ones. This discrepancy may arise because long Foldseek domains, labeled as FP, often correspond to regions where shorter Pfam domains are typically predicted, a relationship further illustrated by Figure 6B. Supplementary Figure S2 shows similar patterns by considering Pfam-N as the gold standard. For example, Pfam-N predicts PF01909, Nucleotidyltransferase domain, and PF03828, Cid1 family poly A polymerase, for the N- and C-terminal of Q38CM2, respectively. On the other hand, Foldseek predicts PF04928, Poly(A) polymerase central domain, for a region encompassing both regions. Moreover, Figure 6C demonstrates that using bits/alnlen as a ranking criterion can balance the precision across domains of varying lengths, a concept further exemplified by comparison of Figure 6D with Figure 6B.

## By increasing the e-value threshold, the majority of domains align with at least one related Pfam domain instance

As the next step, we explored if by using less stringent e-value thresholds, more domains could be annotated. Figure 7 illustrates how adopting less stringent e-value cutoffs in Foldseek leads to more domains aligning with at least one related seed. While this trend is also observed in MMseqs2 hits, the rate of increase relative to the e-value threshold is less pronounced than in Foldseek. However, using less stringent e-values may result in random structures aligning with query proteins, thereby raising the likelihood of inaccurate matches. Consistently, our FLPSS against PfamSDB benchmarking revealed that relying on hits with higher e-values leads to low-precision domain annotations, as anticipated (data not shown).

Our findings underscore a complex challenge in structure-based domain annotation. Specifically, while using more lenient e-values often results in the query structure aligning with instances of the same Pfam domain, the aligned seed does not necessarily rank higher than those of other structurally similar domains.

In another attempt to improve the ranking of the hits, after aligning the FLPSSs with the domain database domain, we tried training an Artificial Neural Network (ANN) to predict

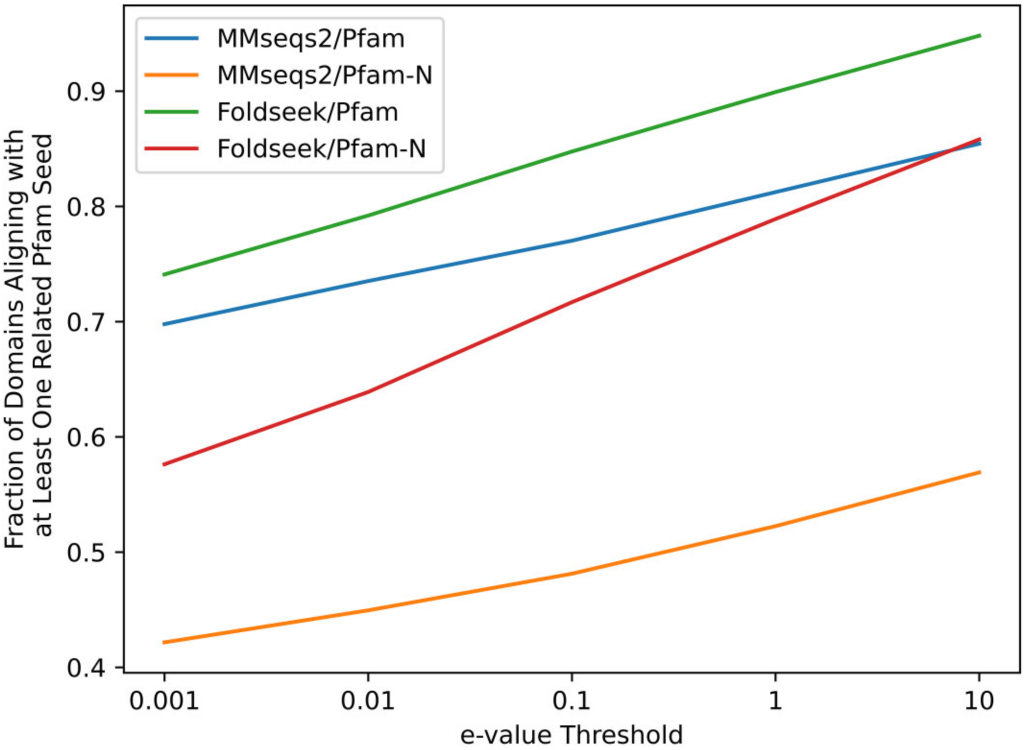

**Figure 7.** Comparison of the proportion of domains aligning with at least one related Pfam domain instance, using either Pfam or Pfam-N as the gold standard. The first part in the legend indicates the alignment tool used for domain annotation, while the second part specifies the gold standard.

**Table 2.** Comparison of various Foldseek alignment settings and their respective counts of predicted domains surpassing gold standard benchmarks

| Settings (databases, ranking criterion) | New predictions above Pfam | New predictions above Pfam-N | New predictions above Pfam and Pfam-N | Mean length | # Hypothetical proteins |
|---|---|---|---|---|---|
| FS_cut, bits | 1251 | 600 | 470 | 232.6 | 230 |
| PDB_cut, bits | 1108 | 514 | 408 | 254.3 | 197 |
| FS_cut, bits/alnlen | 998 | 526 | 390 | 127.3 | 185 |
| PDB_cut, bits/alnlen | 774 | 401 | 290 | 136.3 | 148 |

if the query and target have the same Pfam domain using alignment characteristics such as sequence identity, LDDT, and bits as input features. However, rescoring using the probability of trained ANN did not improve the precision significantly (data not shown). We speculate that re-ranking based on the conservation of critical residues could potentially enhance precision in future attempts.

## New domains predicted above domains predicted by Pfam and Pfam-N

In regions that lacked Pfam predictions, MMseqs2 pinpointed 336 domains. Similarly, it found 341 domains where Pfam-N had not identified any. Furthermore, in areas untouched by both Pfam and Pfam-N, MMseqs2 unveiled 156 new domains with an average length of 68. An examination of the MMseqs2 annotations reveals that the average sequence identity for the domains it annotated is 72.5%. In 51 of these cases, the query and target are 100% identical. This means that, in some instances, even when part of the query sequence has been utilized as a Pfam seed, the corresponding domain remains undetectable in the query protein by both Pfam and Pfam-N. For example, position 469–497 of the protein Q9N937 has been used as a Pfam seed for the domain PF00560 (Leucine Rich Repeat). However, Pfam has not annotated the same region probably because they had not satisfied the gathering thresholds.

Table 2 presents the number of domains identified by Foldseek under various settings, along with the average length of these predicted domains. Additionally, the table includes the count of hypothetical proteins in TriTrypDB (25) that were annotated with a domain. The mean length of the predicted domains is significantly larger than MMseqs2 predictions. Table 2 shows that using FS_cut database and bits as the ranking criteria, the highest number of domains are predicted. Analysis of Figures 4, 5, Tables 1, 2, indicates that while utilizing ‘bits/alnlen’ as the ranking criterion enhances performance in short domain annotation, overall, setting a threshold based on ‘bits/alnlen’ results in a lower recall compared to the use of ‘bits’ as the ranking criterion.

Significantly, PF13458 was attributed 35 times using alignments with FS_cut, where ‘bits’ served as the ranking criterion, and 31 times with PDB_cut, which also employed ‘bits’ for ranking purposes. According to the TriTrypDB website, this particular Pfam domain is primarily associated with genes described as ‘Adenylate cyclase’. Interestingly, in our workflow, the majority genes that were annotated with PF13458 also included ‘Adenylate cyclase’ in their description.

### Case study: domains predicted for proteins involved in *T. brucei* mtRNA editing

Trypanosomes possess a unique mitochondrial DNA structure composed of multi-kilobase-size fragments. These fragments are referred to as maxicircles, which are interconnected with numerous smaller minicircles (26). Within the maxicircles, 12 of the encoded genes are classified as cryptogenes, signifying that they cannot be directly translated after transcription. During the editing process, multiple uracil (U) bases must be either added or deleted to render these genes ready for translation (26). This intricate editing process is guided by guide-RNAs (gRNAs), which are primarily encoded by the minicircles. The gRNAs operate by marking the editing sites on the substrate RNA. They anneal to the substrate, creating bulges that signal the specific locations for catalytic enzymes to modify. For a comprehensive review of the editing process please refer to (26). As a case study, we look deeper into the domains predicted for some of the proteins involved in mtRNA editing. The predictions stem from alignments with either the FS_cut or PDB_cut databases, both utilizing bits as the ranking criterion.

Predictions with all settings suggest that KREPA4 (UniProt ID: Q38B91) may contain the PF00436 domain, known as 'Single-strand binding domains (SSB).' Pfam has also predicted PF00436 for KREPA6 (UniProt ID: Q38B90), with this prediction being consistent with that of Pfam-N. KREPA4 and KREPA6 are adjacent in the genome, and they appear as the closest non-self structural hits to each other when searched against the structure of all organism proteins. This suggests the possibility that they might be paralogs. The expectation that paralogous proteins contain similar domains supports this case. Analysis of the sequence signature of PF00436 reveals that positions 8, 69 and 76 are conserved and contain the amino acid Glycine. Correspondingly, these positions in KREPA4 also contain Glycine. The consistency in this specific amino acid placement between KREPA4 and the conserved signature of PF00436 provides evidence to support the prediction of this domain.

Both KREX1 (UniProt ID: Q57WU3) and KREX2 (UniProt ID: Q38BP2) are predicted to contain PF03159, which corresponds to the XRN 5′-3′ exonuclease N-terminus domain. This prediction is consistent with that made by Pfam-N for these proteins. An early study had hypothesized a similar function for these proteins (referred to as MP100 and MP99 in the publication), and this hypothesis was based on the conservation of certain amino acids. However, the confidence in attributing this domain to KREX1 and KREX2 had been limited at the time due to the low overall homology with the known XRN domains (27).

We predicted the presence of PF02940, identified as the 'mRNA capping enzyme, beta chain,' in RESC1 (UniProt ID: Q57XL7) and RESC2 (UniProt ID: Q586X1). The beta chain of the mRNA capping enzyme is known for its triphosphatase activity. A recent study by Dolce *et al.* elucidated the structure of the RESC1-2 complex using cryo-electron microscopy (28). The researchers reported a structural similarity between RESC1-2 and RNA capping enzymes, which typically have cationic cofactors and are involved in reactions that release phosphate. The study further examined the charged residues within the tunnels of the RNA capping enzyme, noting that one-half of the pattern interacts with the cationic cofactors, and the other half with the released phosphate. However, this pattern was not observed in RESC1-2, leading to the assessment that RESC1 and RESC2 are unlikely to be active enzymes (28).

For RESC5 (UniProt ID: Q389F5), the prediction includes PF19420, identified as '*N*,*N*-dimethylarginine dimethylhydrolase (DDAH) within eukaryotes', a domain related to arginine metabolism. This prediction was bolstered by a recent study that succeeded in crystallizing the structure of RESC5, revealing its structural similarity to the DDAH fold (29). However, this same study also uncovered key differences. Most notably, RESC5 was found to lack residue conservation in critical positions that are otherwise characteristic of the DDAH fold. Further investigation into RESC5's interaction with the DDAH substrate and product provided additional insights. The researchers conducted a Thermal shift assay, a technique used to assess protein-ligand interactions. The addition of the DDAH substrate and product to RESC5 had no discernible effect on the assay's results. This lack of effect serves as an indicator that there is no interaction between RESC5 and the DDAH substrate or product (29). The findings from this detailed analysis demonstrate that despite superficial similarities, RESC5 likely does not function in the same manner as DDAH.

## Conclusion and future work

In this study, we developed a database of domain structures by segmenting the structures predicted by AlphaFold2 at their domain boundaries. To annotate domains, we structurally aligned query proteins with this domain database using Foldseek. Subsequently, we selected the highest-scoring, non-overlapping hits. We either used the default ranking of Foldseek that is based on bits, or reranked the hits of each query based on the bits/alnlen. We then benchmarked these predictions against Pfam v35.0 and Pfam-N predictions.

Our data indicates that for short domains (those <100 amino acids in length), structure-based domain annotation is imprecise. Although ranking hits based on 'bits/alnlen' marginally enhanced precision for short domains, it generally resulted in a lower recall rate compared to the default 'bits'-based ranking. The lack of precision for short domains can be attributed to the fact that short sequences often lack distinctive folds. Consequently, there is a heightened likelihood of random structural similarities between different domains, resulting in reduced precision. Given that a significant fraction of domains are short, our results suggest that while structure-based Pfam annotation cannot supplant sequence-based domain annotation, it can complement it, particularly when annotating longer domains. We are keen to explore the synergy between sequence-based and structure-based domain annotations in future studies.

From an organism-specific standpoint, our study offers insights into the potential functions of genes in *T. brucei*. These insights are ripe for experimental validation. One standout prediction is the anticipated 5′-3′ exonuclease activity for KREX1 and KREX2; we are eager to corroborate this finding through wet lab experiments.

## Data availability

The scripts for creating the PfamSDB are available at https://github.com/Pooryamb/MakingPfamSDB. Scripts for benchmarking can be found at https://github.com/Pooryamb/BenchmarkingFS/. The PfamSDB and the FS_cut database can be accessed at https://zenodo.org/records/10246381.

## Supplementary data

Supplementary Data are available at NARGAB Online.

## Acknowledgements

We would like to express our sincere appreciation to Dr. Milot Mirdita for his expert guidance and detailed instructions on cutting a Foldseek database, which was crucial for the advancement of our work. Additionally, we extend our heartfelt gratitude to the reviewers for their insightful comments and constructive feedback, which significantly contributed to enhancing the quality of this manuscript. Special thanks go to Dr. Naghmeh Poorinmohammad for her invaluable comments. This project has been supported by a research grant 252733 from CIHR to R.S. We also acknowledge the Robert Harpur Foundation for providing financial support to P.M.B.

## Funding

CIHR [252733].

## Conflict of interest statement

None declared.

## References

1. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A. and Colwell,L. (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
2. Illergård,K., Ardell,D.H. and Elofsson,A. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins Struct. Funct. Bioinf.*, **77**, 499–508.
3. Bartas,M., Slychko,K., Brázda,V., Červeň,J., Beaudoin,C.A., Blundell,T.L. and Pečinka,P. (2022) Searching for new Z-DNA/Z-RNA binding proteins based on structural similarity to experimentally validated Z$\alpha$ domain. *Int. J. Mol. Sci.*, **23**, 768.
4. Li,G., Huang,J., Yang,J., He,D., Wang,C., Qi,X., Taylor,I.A., Liu,J. and Peng,Y.-L. (2018) Structure based function-annotation of hypothetical protein MGG_01005 from Magnaporthe oryzae reveals it is the dynein light chain orthologue of dynlt1/3. *Sci. Rep.*, **8**, 3952.
5. Zarembinski,T.I., Hung,L.-W., Mueller-Dieckmann,H.-J., Kim,K.-K., Yokota,H., Kim,R. and Kim,S.-H. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 15189–15193.
6. Gligorijević,V., Renfrew,P.D., Kosciolek,T., Leman,J.K., Berenberg,D., Vatanen,T., Chandler,C., Taylor,B.C., Fisk,I.M., Vlamakis,H., *et al.* (2021) Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.*, **12**, 3168.
7. Zhang,C., Freddolino,P.L. and Zhang,Y. (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.*, **45**, W291–W299.
8. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Ž'idek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold2. *Nature*, **596**, 583–589.
9. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold2 Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
10. Holm,L. (2022) Dali server: structural unification of protein families. *Nucleic Acids Res.*, **50**, W210–W215.
11. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
12. Ayoub,R. and Lee,Y. (2019) RUPEE: a fast and accurate purely geometric protein structure search. *PLoS One*, **14**, e0213712.
13. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2023) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-01773-0.
14. Monzon,V., Paysan-Lafosse,T., Wood,V. and Bateman,A. (2022) Reciprocal best structure hits: using AlphaFold2 models to discover distant homologues. *Bioinform. Adv.*, **2**, vbac072.
15. Ruperti,F., Papadopoulos,N., Musser,J.M., Mirdita,M., Steinegger,M. and Arendt,D. (2023) Cross-phyla protein annotation by structural prediction and alignment. *Genome Biol.*, **24**, 113.
16. Bordin,N., Sillitoe,I., Nallapareddy,V., Rauer,C., Lam,S.D., Waman,V.P., Sen,N., Heinzinger,M., Littmann,M., Kim,S., *et al.* (2023) AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun. Biol.*, **6**, 160.
17. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
18. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. and Finn,R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.
19. Söding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
20. Bileschi,M.L., Belanger,D., Bryant,D.H., Sanderson,T., Carter,B., Sculley,D., Bateman,A., DePristo,M.A. and Colwell,L.J. (2022) Using deep learning to annotate the protein universe. *Nat. Biotechnol.*, **40**, 932–937.
21. Rodrigues,J., Teixeira,J.M.C., Trellet,M. and Bonvin,A. (2018) pdb-tools: a swiss army knife for molecular structures. *F1000Research*, **7**, 1961–1961.
22. Waskom,M.L. (2021) seaborn: statistical data visualization. *J. Open Source Software*, **6**, 3021.
23. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
24. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
25. Amos,B., Aurrecoechea,C., Barba,M., Barreto,A., Basenko,E.Y., Belnap,R., Blevins,A.S., Böhme,U., Brestelli,J., Brunk,B.P., *et al.* (2022) VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.*, **50**, D898–D911.
26. Aphasizheva,I., Alfonzo,J., Carnes,J., Cestari,I., Cruz-Reyes,J., Göringer,H.U., Hajduk,S., Lukeš,J., Madison-Antenucci,S., Maslov,D.A., *et al.* (2020) Lexis and grammar of mitochondrial RNA processing in trypanosomes. *Trends Parasitol.*, **36**, 337–355.
27. Worthey,E.A., Schnaufer,A., Mian,I.S., Stuart,K. and Salavati,R. (2003) Comparative analysis of editosome proteins in trypanosomatids. *Nucleic Acids Res.*, **31**, 6392–6408.
28. Dolce,L.G., Nesterenko,Y., Walther,L., Weis,F. and Kowalinski,E. (2023) Structural basis for guide RNA selection by the RESC1–RESC2 complex. *Nucleic Acids Res.*, **51**, 4602–4612.
29. Salinas,R., Cannistraci,E. and Schumacher,M.A. (2023) Structure of the *T. brucei* kinetoplastid RNA editing substrate-binding complex core component, RESC5. *PLoS One*, **18**, e0282155.

**Received:** September 12, 2023. **Revised:** December 3, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 15, 2024