School of Biological Sciences

ASSESSMENT COVER SHEET AND TEMPLATE

Section A – to be completed by the student

Student Number	**2456386**		
Programme	**Msc Bioinformatics**		
Unit Name	**Bioinformatics Research Project**	Unit Code:	**BIOLM0034**
Assessment name	**Research Project Report**		
Word Count	**5921**		
Do you give permission for you work to be used anonymously in examples given to students in the future? Yes			

By submitting this assignment cover sheet, I confirm that I understand and agree with the following statements:

- 'I have not committed plagiarism, cheated or otherwise committed academic misconduct as defined in the University's Assessment Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- 'I have not submitted this piece, in part or in its entirety, for assessment in another unit assignment (including at other institutions) as outlined in section 4 of the University's Assessment regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- I understand that this piece will be scrutinised by anti-plagiarism software and that I may incur penalties if I am found to have committed plagiarism, as outlined in sections 3 of the University's Examination Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)

CHALLENGES AND STRATEGIES IN FINDING DIELS-ALDER ENZYMES: A COMPARATIVE STUDY OF STRUCTURAL AND SEQUENCE-BASED METHODS

Nicole Alejandra Morveli Flores

Dr Marc Van der Kamp

Bioinformatics 2023

Abstract

This study evaluates structural similarity-based search methods for identifying Diels-Alder (DA) enzyme homologs, focusing on spirotetronate cyclases. Traditional searches use sequence similarity, but structural methods like Foldseek and AlphaFold clusters offer potential advantages. We compared two approaches: Method A, which applies Foldseek with TM-align and 3Di alignment methods, and Method B, using AlphaFold clusters.

Our analysis found that Foldseek's performance is influenced by the choice of alignment method, with TM-align and 3Di yielding different results. Although unique hits were often irrelevant, both methods successfully identified relevant DA homologs. Method B, using AlphaFold clusters, proved more reliable, consistently retrieving accurate hits due to its structured clustering approach.

The study also revealed limitations in relying solely on structural similarity. Incorporating sequence similarity improved search comprehensiveness, as sequence-based methods identified additional relevant hits not captured by structural searches. Predicting closed conformations for DA enzymes using AlphaFold was challenging, highlighting biases in training data and difficulties with prediction parameters.

In summary, combining structural and sequence-based methods enhances the discovery of DA enzyme homologs. Effective protocols should integrate Foldseek with a well-chosen alignment method, incorporate sequence similarity for broader coverage, and use AlphaFold clusters to refine results. Future research should further explore these methods, including additional parameters and experimental validations to confirm potential DA enzyme candidates.

Acknowledgements

I would like to express my sincere gratitude to Dr. Marc Van der Kamp for his exceptional guidance and support throughout this project. His patience, insightful advice, and encouragement have been invaluable in helping me learn and discover my passion for this research. To my amazing friends: Isabella, David, Camila, Danyra and all my housemates. To Calum for being a great listener and being patience. Thank you for being my home and supporting me in this adventure. To my parents and my sister for being my main inspiration and support. To my mom, I do not have words to express my gratitude to you. Finally, to Remigio my main emotional supporter.

Author's Declaration

I declare that this thesis has been composed solely by myself and that is has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Contents of Tables and Figures

Tables

Figures

1. Introduction

In the pharmaceutical industry, carbon-carbon bond formation is essential for synthesizing complex molecules with stereospecificity (1). The Diels-Alder (DA) reaction, a [4+2] cycloaddition between a diene and a dienophile, facilitates this by forming two carbon-carbon bonds and generating cyclohexene structures. Despite its utility, the DA reaction faces challenges such as low efficiency at room temperature and difficulties in controlling stereoselectivity, often requiring high temperatures or pressures that are energetically costly (2,3)

To enhance the efficiency and specificity of the DA reaction, researchers are exploring biocatalysts like enzymes. Enzymatic reactions typically exhibit higher reaction rates and specificity while operating under milder conditions. Among these, the spirotetronate cyclase family has shown promise in catalyzing true [4+2] cycloadditions (4)

Applying enzymes to perform the DA reaction in industry can greatly improve its performance, both in efficiency and specificity. The search for naturally occurring DA enzymes has a long history. However, only a few types of enzymes have been proven capable of catalyzing true [4+2] cycloadditions. One of the proven enzymes is the spirotetronate cyclase family. They are a group of DA enzymes involved in the synthesis of polyketide natural products with a spirotetronate moiety. Well-characterized members of this enzyme family include PyrI4, AbyU, AbmU and AbnU (5–8).

For instance, AbyU is part of the abyssomicin C biosynthetic pathway. It consists of two eight-stranded antiparallel β-barrels with (+1)8 topology sealed on both ends. A salt bridge formed by Glu (78) and Arg (122) seals one end. At the opposite end, there is a hydrophobic loop. It is formed by a β1-β2 linker capping loop. The central part of the barrel forms a large hydrophobic cavity that can be accessed through the capping loop. The capping loop is flexible, allowing substrate entry into the cavity. It becomes more ordered once the substrate is bound. Therefore, AbyU can adopt two conformations: open and closed. The conformation is determined by the position of the capping loop (6,9).

Contrary to other enzymes, there is no evidence of highly conserved amino acid features in the active site. The reaction likely proceeds through hydrophobic and steric complementarity, driven primarily by the accommodation of the substrate inside the cavity (2). To date, over 400 compounds have been identified as products of the DA reaction, but only a few DA

enzymes have been fully characterized. This suggests that many DA enzymes remain to be discovered.

Given the industrial importance of Diels–Ader (DA) enzymes, there is a growing interest in discovering more of these enzymes. Protein databases contain many protein sequences, but few structures have been characterized. To address this, sequence similarity searches are commonly used. They identify homologous sequences to a desired query. However, inferring functional similarity from sequence similarity is not straightforward (10). This process becomes even more complex when dealing with remote homologs. These are proteins that share structure and function but exhibit low sequence identity (11).

This complexity is evident in the case of the spirotetronate DA enzymes, which often share less than 25% sequence identity (6). Despite this, the spirotetronate cyclase Cyc15 was successfully identified using sequence-based searching. The strategy involved creating a library of sequences encoding known or predicted DA enzymes. Then, the search was expanded using BLASTp to find homologous sequences in public databases. To visualize the results, a sequence similarity network (SSN) was constructed. A SSN groups sequences into nodes connected by edges based on a selected sequence similarity threshold. A representative sequence from each node was then expressed and tested for enzymatic activity, using the natural substrate of AbyU. Notably, Cyc15 exhibited considerable DA activity and resulted in a product with the opposite stereochemistry to the AbyU product (6). This case demonstrates the effectiveness of using sequence similarity to find homologous proteins. However, it may not be comprehensive: some hits can be missed. In the case of DA enzymes, there is no evidence of specific sequence features retained in the active site. Consequently, local and global sequence identity may be too small to be statistically significant (12).

Structural alignment tools (structure-based methods) offer an alternative approach for detecting homologous proteins with similar functions. Such methods compare the overall fold of the query protein to identify hits that share the same structure. Another alternative is protein language models. They represent proteins as embeddings. They can capture complex patterns in sequences, such as PLM-Blast (13).

Foldseek is an efficient structural search algorithm that converts the query to the dictionary which describes the tertiary interactions of amino acids (3Di dictionary). This allows for the identification of proteins with a common structure (14). In the Foldseek web server, a 3Di-based search in structural databases can be performed, with the hits ranked based on the e-

value obtained. As an alternative to 3Di, TM-align can be used. It performs the alignment based on the position of the backbone Cα atoms (15). Foldseek has been used to cluster the AlphaFold database (AFDB), creating a database called AlphaFold Clusters (16). The COFACTOR platform offers another approach to searching protein structures based on the TM-align algorithm, using the BioLip database (17)

These methods provide alternatives to sequence-based searching for finding distant homologs. Thus, they can be applied to expand the search for spirotetronate DA enzymes. Foldseek has demonstrated high efficiency, sensitivity, and alignment quality. A benchmark study compared its performance with other structural aligners such as Dali, CE, and TM-align (14). It concluded that Foldseek sensitivity is comparable to Dali and significantly higher than CLE-SW and MMseqs2. These findings support the use of Foldseek for finding spirotetronate DA homologs.

Many proteins share the β-barrel fold found in the characterized spirotetronate DA enzymes. However, they do not all display DA (18). Therefore, additional features must be evaluated to determine whether a hit could be a DA homolog. One key feature is the presence of a salt bridge. Such an evaluation should include the presence of a suitable cavity for the substrate to be confined in.

To accurately compare cavity metrics, DA enzymes must be in their closed conformations. This requires the capping loop to adopt its ordered structure and seal one end of the β-barrel. AlphaFold2 (19), e.g. as applied through ColabFold (20), is a tool capable of accurate protein conformation predictions and could thus be used to predict the closed conformation of DA enzymes.

Structure-based approaches offer a promising strategy for detecting remote homologs. However, their implementation could have some challenges. First, there is no previous structure-based search protocol successfully tested for finding DA enzymes. Therefore, the main aim of this study is to develop an effective and reliable search protocol to identify putative spirotetronate DA enzymes. The different parameters (query, alignment method) will be assessed. Previously characterized members of the spirotetronate DAse family will be used as queries. The results obtained by the different methods (with their specific parameters) will be compared. A Sequence Similarity Network (SSN) will be used to differentiate hits that could be relevant. Another important aim of this study is to find DA enzymes with alternative properties (e.g. different cavity size or stereoselectivity). Therefore, the relevant hits obtained will be structurally evaluated to compare them with current known

DA enzymes. Finally, predictions of their closed conformation will be performed to evaluate their cavity volume.

2. Methods

Note: all methods were run in default parameters unless otherwise indicated.

2.1 Data collection

2.1.1 Searching based on structure

Different structural-based tools were used to find DA enzyme homologs. They were chosen according to previous benchmark studies that proved their effectiveness and sensitivity for finding remote homologs (14,16). Foldseek (search.foldseek) was used for method A using TM-align and 3Di search. AlphaFold clusters (search.AFDBcluster) were used for method B. The clusters obtained by each query were collected. The web server also provides a list of 10 similar clusters every search. The ten similar clusters to AbyU were included to complement the results. Structural alignment based on Cα positions has also proven high sensitivity, although at the expense of speed. To have a robust comparison between current structural-based search tools, COFACTOR (search.COFACTOR) was used for Method C. Finally, protein language models are a new alternative to structural alignments. PLM-blast has been reported to excel at distant homolog proteins search (21). Therefore, Method D employed PLM-blast (search.PLM-blast). All search methodologies are detailed in Table 1. Hits sharing >95% identity (based on Blastp) were discarded (Table S1). The presence of previously characterized spirotetronate cyclases with confirmed or predicted DA activity(6) was evaluated (positive control) (Table S2). PDB files and fasta sequences of each query used can be found in Figure S1 and Table S3.

2.1.2 Searching based on sequence

The fasta sequences obtained by (6)were collected and searched in UniProt (search.UniProt). The variants (sequences with >95% identity) were discarded using Blastp. The sequences that did not yield matches in UniProt were discarded. These hits are named ‘sequence hits’.

2.2 Visualizing and selecting relevant hits

Sequence similarity networks (SSNs) were built using EFI-EST (22) with UniProt IDs. from each search method (sequences without start or stop codons were excluded). Table S4 lists the parameters used for each network. The SSNs were visualized with Cytoscape (23), using the organic layout. Only members of the cluster(s) containing the positive controls were selected for further analysis (name as relevant hits).

2.3 Evaluating the relevant hits 3D structure

Structure visualization was performed in PyMOL (24), using structures from PDB if present, or else from the Alphafold database (search.AlphaFold). Alignment to AbyU was performed in PyMOL. For sequence hits do not present in UniProt, structures were predicted using ESMfold (searchESMMetagenomicAtlas) (25).

2.4 Evaluating the active cavity site

2.4.1 Assessing the presence of the salt bridge

Hits were aligned based on sequence using Clustal Omega (26)(search.ClustalOmega). The MSA was visualized using a guide phylogenetic tree made by Clustal Omega. They were aligned based on structure using TExpresso (27) (searchTCoffee.com).

2.4.2 Predicting the closed conformations

Predicted structures were considered to have a closed conformation if the capping loop acquired a similar position to that of AbyU (see Figure S1). AF2 prediction in ColabFold (20)was used in an attempt to predict the closed conformation of remaining hits. Different strategies were tested:

- Closed structures of AbyU (PDB: 5DYQ) or Cyc15 (AFDB prediction of UniProt A0A401MXE6) were used as a template (template-based). The highest-scoring models were visualized.
- Adjustments were made to dropout rates, number of seeds (num_seeds), and MSA depth (max_msa) (parameter-based). The first four highest-scoring models were visualized.

2.4.3 Calculating the cavity volume

CastP (searchCastp) (28) was used to calculate the cavity volume using a sphere radius 185 size of 1.1Å.

3. Results

3.1 Structure-based search results

3.1.1 Method A

3.1.2.1 Effect of the alignment

Method A employed Foldseek with four different queries and two alignment methods (3di or TM) (see Table 1). The hits obtained are summarized in Table 2. In general, both alignment methods retrieved a similar number of hits independent of the query. However, each alignment method retrieved some unique hits that were not found by the other. If these unique hits are irrelevant (e.g., not complete β-barrels structures) then the search can be performed either using TM-align or 3Di. To confirm this, the unique hits were searched in the list of relevant hits (sub-cluster members obtained after SSN visualization). None of them were found, thus the relevant hits were found by both methods equally. The highest-scoring hits retrieved by TM-align and 3di were evaluated (Figure 1A). They differ significantly in structure. B6TGG4 (retrieved by TM-align) contains a partial β-barrels with a loop, while Q75KD7 (retrieved by 3Di) contains a β-barrels with an additional non-folded structure.

3.1.2.2 Effect of the query

The number of (unique) results also differs per query, with Cyc15 recovering the most unique hits. A Blastp alignment was conducted to assess the similarity between the queries (all vs. all). The goal was to determine if this similarity affects the number of unique hits retrieved. The logic was that the more similar the queries are to each other, the fewer unique results they should retrieve. AbnU shares 49% similarity with AbyU (see Table 3) when compared directly, yet each still retrieved unique results. To further assess this, a similar sequence to AbyU (identified in the phylogenetic tree in Figure S3) was used as a query in Foldseek. The query chosen, A0A246RFQ9, shares 83% similarity with AbyU. The hits obtained were compared to those from AbyU. A0A246RFQ9 retrieved some unique hits, but to a much lesser extent compared to AbnU (Table 4). This suggests that considering sequence similarity between queries is relevant to expanding the search. However, the unique hits retrieved by each query were also searched in the list of relevant hits and they were ultimately irrelevant for this study. Then, using different queries expands the search but does not increase its specificity as all the relevant hits were found by all the queries. The unique hits still display structures β-barrel like structures, but they are incomplete, or they are part of a larger complex. The highest-scoring unique hits retrieved by each query were visualized in PyMOL (Figure 1B). Their respective score is also shown (either tm or 3di). For reference, a hit with a TM score below 0.5 or and e-value higher than 0.01 is often considered not significant (14).

Notably, method A did not retrieve AbyU as a hit when AlphaFold/Uniprot50 v4, AlphaFold/Swiss-prot v4, AlphaFold/Proteome v4 were used as databases. The search was expanded by adding PDB100. After performing these changes, AbyU was found. A possible explanation could be that the pre-filter stage of Foldseek was too restrictive. The pre-filter performs an alignment that reduces false positives and non-homologous sequences. The criterion is based on double k-mer matches which non-homologous sequences are less likely to have (14). However, when turning off the pre-filter AbyU was not retrieved. Other reasons not explored here influenced Foldseek's inability to retrieve AbyU when the PDB100 is not used as a database.

3.1.2 Method B

` 3.1.2.1 Effect of the query

The hits obtained by method B are summarized in Table 5. AbnU, AbyU and Cyc15 were clustered together, the total members of the cluster were 78 proteins. PyrI4 was clustered separately with 32 proteins. The total results obtained are higher compared to the hits retrieved by method A. However, all the hits are concentrated in the A0A022Q700 cluster with 4000 members. Notably, the effect of the query is very different for both methods. In method A each query retrieved a significant amount of unique hits. Conversely, in method B AbnU, AbyU and Cyc15 retrieved the same hits. Cluster A0A0K9XAS6 and A0A3A9ZPZ1 were found by the queries. Therefore, the members of these clusters could be relevant hits. They were searched in the list of Method B's relevant hits (121 hits). Clusters A0A0K9XAS6 and A0A3A9ZPZ1 retrieved 75 hits of the relevant hits. The other 34 hits were found by the added clusters. In that sense, including the additional clusters expands the search and still retrieves relevant hits. The search is enriched when queries with lower sequence similarity (PyrI4) are added. The clusters' representative structure was evaluated in Pymol (Figure 1C). They all display a β-barrels structure that resembles the queries used. Therefore, expanding the search in method B proved to be more effective in this study than in method A.

3.1.3 Method C & D

The hits obtained method C and D were also evaluated (Table S6 and Table S7). Currently these webservers only employ PDB100 as the database. Therefore, the number of hits obtained are significantly lower compared to the other methods. The hits were evaluated manually. Method C was originally run as a "state-of-the-art" approach, where all the positive controls were expected to be found. As expected, method C hits included well-known DA enzymes. However, it retrieved some other hits that are not characterized by DA enzymes,

such as pterocarpan synthases (29). Also, an allene oxide cyclase is involved in oxylipin biosynthesis (30). The hits obtained by method D included some well-known spirotetronate DA enzymes. However, other hits displayed non-beta barrel structures and beta barrels without DA activity. For instance, JAC1 (JAC1.PDB) has been reported to have chaperone activity (31). Therefore, the results obtained by methods C and D were not considered for further analysis, except for those already included in other methods.

3.2 Relevant hits selection and visualization

After the results retrieval for each method, the SSNs were visualized. The SSN of methods A and B were compared. The connection threshold used was 25% sequence similarity. For both methods, all the positive controls were clustered together and in a separated cluster. Therefore, members of these sub-clusters were considered relevant hits. The relevant hits unique to each method were counted (Table 6). The sub-clusters of each method can be found in Figure 2. Both methods retrieved members of the positive control. Method B, retrieved the most compared to method A. However, is important to consider various search methodologies as both methods found unique hits.

3.3 Structure evaluation

Given that they were clustered together, all the relevant hits were expected to display an 8-stranded β-barrels. However, some of the structures found were not β-barrels, they displayed different folds. K4FDH3 displayed a structure containing two beta barrels connected by a loop with different sequence content (Figure S4A)., Each beta barrel was thus evaluated as an individual hit. A0A522WLH4 and A0A838IMS4 displayed a structure indicated as ‘complex’. It contained a β-barrels similar to other structures, but also significant additional structural elements. Thus, they were not included in further analysis (Figure S4B and Figure SBC).

The structures of the hits previously obtained using sequence-based search (6)were also evaluated. Some of the hits displayed a β-barrels structure but other hits displayed other folds. The combined results from the structure-based search (methods A and B) were counted and compared with the hits obtained by Zorn et al (see Table 7). The structure-based search retrieved more unique hits, although both methods obtained hits that were not found by the other. Therefore, searching homologs based on sequence similarity also retrieves relevant hits. However, searching by structure provided more results that displayed the β-barrels fold. Only hits obtained by structure-based search (structural hits) were considered for further analysis.

3.4 Active Cavity size Evaluation

3.4.1 Salt bridge evaluation

Most of the reactions in the pharmaceutical industry are inter-molecular. The ideal DA enzyme will have enough cavity space to perform a reaction with two substrates. AbyU typically displays a salt bridge formed by Arg-His in positions 78 and 122 (see Figure 3A) The presence of the salt bridge limits the cavity volume, thus hits without the Arg -His salt bridge will be of interest. Structure-based sequence alignment (using T-COFFEE) was performed. The presence of a salt bridge that caps the active site (positions 78 and 122 in AbyU numbering) was found in many, but not all hits (Table 8). Some hits do not contain the same Glu/Asp-Arg amino acid pair (as in AbyU and others) but have similar amino acids. They can still cap the cavity through hydrogen bonding (e.g. Arg replaced by Tyr) Further, bulky amino acids like Arg would still affect the cavity volume even if they do not form a salt bridge with the amino acid on the opposite side of the barrel. Therefore, to reveal hits with possible alternative cavity shapes/sizes, particular attention was given to hits that do not contain Arg. However, despite not having Arg in the same position, some hits still have bulky amino acids such as leucine (L), methionine (M), tyrosine (Y), tryptophan (W), and leucine (L). Their presence could still limit the cavity volume even without forming a salt bridge. In position 122, the amino acids tend to be bulky. Conversely, in position 78 the amino acids vary significantly in composition and do not follow a particular trend.

3.4.2 Closed conformation prediction

As described in the introduction for AbyU, the cavity is often delimited by a salt bridge at one end and the capping loop at the other. Then, the cavity is lined by the side chains of several residues that point into the cavity (Figure 3). When the loop acquires an ordered structure, the proteins are considered in its closed conformation. The conformations present in the AFDB either have an open or a closed capping loop or domain (Table 9).

In an attempt to obtain closed conformations for those hits with an open conformation, custom predictions were tested. Tests were initially performed with hit A0A1Y0RPR1 (AFDBentry), using closed structures of AbyU (from PDB ID 5DYV) or Cyc15 (AlphaFold prediction of A0A401MXE6) as templates (Figure 4 and Table 10). When a custom template is used, this structure is used as the basis for the initial ‘pair representation’ prediction in AlphaFold. It was assumed that if the template and the query share a higher sequence identity in the capping loop region, it will help get alternative conformations. However, with both AbyU and Cyc15 as templates (with Cyc15 being more similar to A0A1Y0RPR1 in the loop region), an open conformation was predicted (Figure S6). Notably, only one temple-based prediction was successful, A0A345XXX2 (AFDB.entry) using AbyU as a template (Figure 5).

Therefore, further prediction attempts used a different approach. They were based on previous reports that successfully predicted alternative protein conformations. The assumption was that conformations can be decoded from sequence (32)Using the Da Silva approach, the MSA and the number of seeds were changed. Alternative capping conformations were predicted (Figure 6, Table 11), but this does not resemble a fully closed conformation.

Notably, the alternative conformations were obtained when the number of seeds was changed to 16 in all the MSA depths. Then, it was hypothesized that the number of seeds was controlling the change in the conformations rather than the MSA depth.
Additional predictions were made while keeping the MSA depth parameter in its default setting and varying the number of seeds. All the number of seeds (from 2-8) obtained some movement of the capping loop (except for number of seeds of 16). When compared to the previous predictions where the MSA depth was also changed (Figure 11C), the movement of the capping loop is less evident. Therefore, changing the MSA depth and the number of seeds had a stronger effect on the capping loop position.

3.4.3 Cavity volume evaluation

Although closed models were not obtained for all hits, the cavity of those models that do have a closed conformation can be further evaluated. Hits without the salt bridge are expected to have a higher cavity volume, as the cavity may extend further into the barrel (as observed for PyrI4). After separating hits that either have an Arg at position 122 (‘salt bridge’ or similar capping of the cavity) or not, mean cavity volumes of the closed hits were calculated (Table 12). The results indicate that in general, hits lacking the salt bridge tend to have a higher cavity volume. However, to attribute this to the presence of the salt bridge is difficult to conclude. Only 4 cavity volumes without the salt bridge were measured (compared to 17 with the salt bridge). Therefore, there is not enough data to make a confident comparison.

4. Discussion

This study investigated different search protocols based on structural similarity to find DA homologs. While earlier studies have successfully found DA enzymes using sequence similarity searches, they have not employed structural similarity as a search parameter. Given the low sequence similarity they share, using structural-based search may help uncover new DA enzymes. Foldseek (method A) and Alpha Fold clusters (method B) were employed as search algorithms. As there are no previous reports using these methods to find DA enzymes, the parameters of each search were compared. The goal is to develop an effective search method that expands the search but also is specific enough to find putative DA enzymes. Additionally, the relevant hits obtained were evaluated structurally (cavity measurement) to present them as a possible DA enzyme which could later be tested experimentally.

We found that the Foldseek search is more complex than expected. Initially, Foldseek was thought to retrieve the same hits independent of the query or alignment method. In Method A each query and alignment method retrieved unique hits (see Table 2). However, the unique hits were ultimately irrelevant. In a benchmark study (14), different alignment methods were compared including Foldseek (3Di) and Foldseek-TM. Their ability to classify members of the same family and superfamily was tested (Scope dataset). Both methods had the highest and third highest performance. Although very similar, they still had small variations when finding true positives. Therefore, when used in the webserver they are expected to retrieve different results. One of the key differences between the alignment methods relies on the local and global differences detection. TM-align specializes in detecting global similarities. In contrast, 3Di focuses on local structures. However, both methods utilize the 3Di dictionary to convert the query. TM-align, then, is not applied in its 'pure' form as in other algorithms (eg. COFACTOR). The 3Di conversion considers tertiary interaction between aminoacidic pairs, while TM-align only considers Cα positions. The different hits retrieved by each method, then, could be due to the different scoring systems that each alignment applies. However, the effect of the scoring system on the structures of the hits retrieved is not evident. Overall, each alignment method retrieved unique hits that do not have an evident structural difference between them. The alignment method, then, does not capture a specific structure. For this study, both alignment methods effectively retrieved the relevant hits. Therefore, when using method A the search can be performed with either TM-align or 3Di indistinctly.

The queries used in Foldseek also impact the hits obtained. Given that they share a similar fold, they were expected to retrieve similar hits. However, each query retrieved different

results. This divergence could be explained by how Foldseek converts aminoacidic composition into the 3di dictionary. For each residue i, it identifies its nearest neighbour j which is then categorized into one of 20 possible states. Although the descriptors consider spatial conformation, the amino acid identities also influence how these conformations are described. Therefore, even though the queries AbyU, AbnU, Cyc15, and PyrL4 share a similar fold, their sequence content influences the retrieval of different results. This is shown in Table S4, A0A246RFQ9 still manages to retrieve unique hits but at a much lower capacity than AbnU. However, the unique hits were irrelevant. When comparing their structure, they displayed a variety of folds. Notably, their respective scores (TM or e-value) are much lower when compared to the relevant hits.

All the queries retrieved the relevant hits, irrespective of their sequence content. Conversely, non-relevant hits display different folds. This suggests that the highest information density is encoded in conserved protein cores, while non-conserved coil/loop regions carry lower information density (14). The 3Di effectively captures this, as all the highest-scoring hits were found by all queries and most of them displayed the β-barrels. Conversely, the unique hits retrieved by each query were retrieved when aligning partial β-barrels structures or additional motifs which were found by the specific sequence content of each query.

Moreover, we found that relying uniquely on structure is not effective. The importance of sequence similarity was first shown in the different hits obtained by methods A and B. Both methods retrieved members of the positive control, but method B proved to be more effective. It retrieved more unique hits and positive control than method A (see Table 4). Another key difference was the impact of the query in both methods which was not expected. Although Foldseek and AlphaFold clusters are different algorithms, they both employ the 3Di dictionary. Then, their implementation should be similar. In the AFDB AbyU, AbnU, and Cyc15 are clustered together. When used as queries in foldseek, then, they should retrieve similar hits. However, each query retrieved a large amount of unique hits. This is partly explained by the difference in sequence content. However, it does not fully explain why in method A queries (AbyU, AbnU, and Cyc15) could expand the search while in method B they retrieved the same results. This divergence can be explained by the difference between the methods. Unlike the Foldseek algorithm, the initial stage of clustering for the AF clusters is based on a multiple sequence alignment (MSA). It uses MMseqs with 50% identity and 90% overlap to select a cluster representative. Subsequently, the 3Di dictionary from Foldseek is used for structure-based clustering (16). AF clusters filter hits by a sequence identity threshold, whereas Foldseek skips this step. As a result, in method A the query is compared with a larger pool of target sequences. Then, the hits obtained are much more diverse than method B. However, this diversity does not mean that the hits are

relevant. Adding more queries in method B increases the relevant hits obtained. Conversely, in method A, as there is no previous sequence identity filter adding more queries did not increase the relevant hits retrieved.

Sequence similarity also proved to be useful when filtering relevant results. In the SSN of all methods, the relevant hits formed a separated cluster. Furthermore, sequence-based search also retrieved results that were not found by structural-based search. Therefore, sequence similarity needs to be considered to perform a reliable search. Pereira et al. (33)showed the importance of a combined approach. They successfully classified the 'dark matter' proteins in UniProt by using structure features and sequence similarity. Therefore, a combined approach with sequence similarity and structural alignment proves to be effective for finding remote DA homologs.

Predicting the closed conformation proved more challenging than anticipated. Notably, most of the hits did not need a prediction as they were already in a closed conformation. Del Alamo et al. ((34), concluded that AlphaFold is biased towards known structures. This can be explained by the data that was used to train AlphaFold neural networks. The data available in PDB before April 30 2018 was used. By 2018, only AbyU and PyrI4 from the positive control had a crystal structure entry in PDB. AbyU has 5DYQ (open conformation) and 5DYV (closed conformation) and PyrI4 has 5BTU (open conformation) and 7DVK (closed conformation). Thereby, there is a bias toward previously known proteins (34). AlphaFold already has the information of two characterized DA conformational states (open or closed). Then, the structure prediction of putative DA enzymes could acquire the open or closed conformation. However, using ColabFold to predict the closed conformation of the open-structured hits was challenging. The use of templates in their closed conformation (AbyU or Cyc15) was irrelevant to predicting an alternative conformation. In contrast, changing the MSA depth (max_msa parameter) predicted intermediate conformations. This agrees with the results obtained by Del Alamo *et al.* where a template in the outward-facing (OF) conformation of MCT1 was provided to ColabFold, along with MSA of various sizes. The desired OF conformation was obtained only with MSAs of 16-32 sequences. The other conformations obtained were either inward-facing (IF) or intermediate conformations. The same results were observed when using LAT1 in its OF or IF template or PTH1R in its active or inactive state. Their alternative conformations were obtained only when the MSA was shallow (34). This suggests that templates are useful when the MSA depth is changed. As the MSA depth increases, the template becomes less informative. Moreover, as suggested by Wayment-Steele *et al*. different proteins need an appropriate MSA depth to predict a desired conformation (35). In this case, the accurate MSA depth was not found. The number

of seeds was also a relevant parameter to obtain different conformations. Then, the appropriate parameters combination (num_seeds and max_seq) would predict a closed conformation.

This study, however, has some limitations. First, when accessing the difference between the methods (alignment and queries) only the highest scoring hits were evaluated. To have a robust comparison, all the unique hits should be evaluated. Their respective scores and structures should be compared. However, as they were ultimately irrelevant, they did not affect the results obtained.

The closed conformations prediction also has some limitations. First, not many DA enzymes have been characterized. Then, using tools that rely on neural networks may limit the results obtained as they are trained on characterized proteins available in databases (e.g. PDB). Alternatively, another prediction tool could be used to predict the closed conformations. This was performed by Mbatha *et al*. (9), they used Modeller (search. modeller) to predict the closed conformation of AbmU. However, using ColabFold helped notice the existing AF bias in DA conformation prediction. Moreover, it also showed the importance of configuring the MSA depth and number of seeds.

Avenues for future research include evaluating a set of remote homolog enzyme functions. For instance, the complete search protocol (including search, SSN visualization, and structure evaluation) had relevant results for finding putative DA enzymes. However, other enzymes should be tested to conclude if combining structure-based search and sequence similarity is an optimal strategy. Cavity volume and the presence of the salt bridge are parameters that help classify DA enzymes. However, the evaluation of more parameters is important to evaluate alternative spirotetronate DA enzymes. For instance, cavity size and cavity lining residues can be included as classification parameters. Moreover, docking experiments and reaction simulations should be implemented to state that a hit could be an alternative spirotetronate DA enzymes.

While new structural-based tools to search remote homologs are available, an effective protocol for implementing these tools in the search for unknown DA enzymes has not been done yet. The parameters, the algorithm, and the sequence similarity used have an impact on the results obtained. To do an effective search, sequence similarity, and structural approaches must be combined. These are key recommendations when searching remote enzyme homologs:

- Foldseek: the search parameters can include one well-known query and any alignment method (TM-align or 3Di). The use of multiple queries is not necessary.
- Sequence similarity: Alternatively, an MSA can be done with sequences retrieved from multiple databases. Then, only sequences sharing a % of sequence similarity will be kept. Next, they can be used as queries in Foldseek.
- AlphaFold clusters: The first step should be identifying the clusters containing the well-known enzymes. Then, uncharacterized members of this cluster can be used as queries to search again. This step can be repeated until the structure does not resemble the initial query's fold.

These recommendations would help refine the search and increase the possibility of finding a putative remote homolog of the desired enzyme.

5. Bibliography

1. Gao L, Yang J, Lei X. Enzymatic intermolecular Diels-Alder reactions in synthesis: From nature to design. Vol. 2, Tetrahedron Chem. 2022.
2. Byrne MJ, Lees NR, Han LC, Van Der Kamp MW, Mulholland AJ, Stach JEM, et al. The Catalytic Mechanism of a Natural Diels-Alderase Revealed in Molecular Detail. J Am Chem Soc. 2016;138(19).
3. Gregoritza M, Brandl FP. The Diels–Alder reaction: A powerful tool for the design of drug delivery systems and biomaterials. European Journal of Pharmaceutics and Biopharmaceutics. 2015 Nov 1;97:438–53.
4. Sara AA, Um-E-Farwa UEF, Saeed A, Kalesse M. Recent Applications of the Diels-Alder Reaction in the Synthesis of Natural Products (2017-2020). Synthesis (Germany). 2022;54(4).
5. Tian Z, Sun P, Yan Y, Wu Z, Zheng Q, Zhou S, et al. An enzymatic [4+2] cyclization cascade creates the pentacyclic core of pyrroindomycins. Nat Chem Biol. 2015;11(4).
6. Zorn K, Back CR, Barringer R, Chadimová V, Manzo-Ruiz M, Mbatha SZ, et al. Interrogation of an Enzyme Library Reveals the Catalytic Plasticity of Naturally Evolved [4+2] Cyclases. ChemBioChem. 2023;24(14).
7. Li B, Guan X, Yang S, Zou Y, Liu W, Houk KN. Mechanism of the Stereoselective Catalysis of Diels-Alderase PyrE3 Involved in Pyrroindomycin Biosynthesis. J Am Chem Soc. 2022;144(11).
8. Kashyap R, Yerra NV, Oja J, Bala S, Potuganti GR, Thota JR, et al. Exo-selective intermolecular Diels–Alder reaction by PyrI4 and AbnU on non-natural substrates. Commun Chem. 2021;4(1).
9. Mbatha SZ, Back CR, Devine AJ, Mulliner HM, Johns ST, Lewin H, et al. Antibiotic origami: selective formation of spirotetronates in abyssomicin biosynthesis. Chem Sci. 2024;
10. Pearson WR. An introduction to sequence similarity (“homology”) searching. Curr Protoc Bioinformatics. 2013;(SUPPL.42).
11. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform. 2018;19(2).
12. Pertsemlidis A, Fondon JW. Having a BLAST with bioinformatics (and avoiding BLAST phemy). Vol. 2, Genome Biology. 2001.
13. Kaminski K, Ludwiczak J, Pawlicki K, Alva V, Dunin-Horkawicz S. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. Bioinformatics. 2023;39(10).

14. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2).
15. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7).
16. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, et al. Clustering predicted structures at the scale of the known protein universe. Nature. 2023;622(7983).
17. Zhang C, Freddolino PL, Zhang Y. COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Res. 2017;45(W1).
18. Fairman JW, Noinaj N, Buchanan SK. The structural biology of β-barrel membrane proteins: A summary of recent reports. Vol. 21, Current Opinion in Structural Biology. 2011.
19. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873).
20. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6).
21. Pantolini L, Studer G, Pereira J, Durairaj J, Tauriello G, Schwede T. Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. Bioinformatics. 2024;40(1).
22. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58(41).
23. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: Empowering workflow-based network analysis. Genome Biol. 2019;20(1).
24. DeLano WL. Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography. 2002;40.
25. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model [Internet]. Vol. 379, Science. 2023. Available from: https://esmatlas.com
26. Sievers F, Higgins DG. Clustal Omega. Curr Protoc Bioinformatics. 2014;2014.
27. Poirot O, O'Toole E, Notredame C. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res. 2003;31(13).
28. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: Computed atlas of surface topography of proteins. Nucleic Acids Res. 2018;46(W1).

29. Meng Q, Moinuddin SGA, Kim SJ, Bedgar DL, Costa MA, Thomas DG, et al. Pterocarpan synthase (PTS) structures suggest a common quinone methide-stabilizing function in dirigent proteins and proteins with dirigent-like domains. Journal of Biological Chemistry. 2020;295(33).
30. Neumann P, Brodhun F, Sauer K, Herrfurth C, Hamberg M, Brinkmann J, et al. Crystal Structures of physcomitrella patens AOC1 and AOC2: Insights into the enzyme mechanism and differences in substrate specificity. Plant Physiol. 2012;160(3).
31. Huwa N, Weiergräber OH, Fejzagić A V., Kirsch C, Schaffrath U, Classen T. The Crystal Structure of the Defense Conferring Rice Protein OsJAC1 Reveals a Carbohydrate Binding Site on the Dirigent-like Domain. Biomolecules. 2022;12(8).
32. da Silva GM, Cui JY, Dalgarno DC, Lisi GP, Rubenstein BM. Predicting relative populations of protein conformations without a physics engine using AlphaFold2. Biophys J. 2024;123(3).
33. Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, et al. Uncovering new families and folds in the natural protein universe. Nature. 2023;622(7983).
34. Del Alamo D, Sala D, McHaourab HS, Meiler J. TITLE: Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife. 2022;11.
35. Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature. 2024;625(7996).

7. Tables & Figures

7.1 Tables

Method	Webserver	Queries	Databases
A	clusters.foldseek.com	AbnU, AbyU, Cyc15,PyrI4	AlphaFold/Uniprot50 v4 AlphaFold/Swiss-prot v4 AlphaFold/Proteome v4 PDB100
B	search.foldseek.com	AbnU, AbyU, Cyc15,PyrI4	AlphaFold/Proteome v4
C	search.COFACTOR	AbyU	PDB100
D	search.plmBLAST	AbnU, AbyU, Cyc15,PyrI4	PDB100

TABLE 1 - Detailed Search Protocol: Queries, Methods, and Databases

Detailed search methodologies, including the specific search queries, methods employed, and databases utilized for data collection. The table provides a comprehensive overview of the approach used in each search instance, outlining the search parameters, the databases accessed, and the strategies applied to gather relevant data for the study. This structured format ensures transparency and reproducibility of the search process.

	AbnU	AbyU	Cyc15	Pyrl4
AbnU	100%	49%	31%	26%
AbyU	49%	100%	39%	31%
Cyc15	31%	39%	100%	31%
Pyrl4	26%	31%	26%	100%
A0A246RFQ9	83%			

TABLE 2- Analysis of Sequence Similarity Among Queries: Insights and Comparisons. Sequence similarity analysis of queries, including A0A246RFQ9, to assess the relationship between sequence similarity and the retrieval process. A0A246RFQ9 was selected based on a multiple sequence alignment (MSA) performed with ClustalW, where it showed the highest sequence similarity to AbyU. Sequence similarity was calculated using BLASTp to evaluate if higher similarity between queries influences their results. The table displays the similarity scores between the sequences, providing insight into the comparison of query sequences used in the study.

	Subject		
Query	AbyU	AbnU	A0A246RFQ9
AbyU	0	1123	408
AbnU	1968	0	
A0A246RFQ9	1274	0	0
Total hits	2345	1500	1479

TABLE 3 - Relationship Between Sequence Similarity and Unique Hit Retrieval.
Unique hit analysis for query A0A246RFQ9 using Foldseek. A0A246RFQ9 was included due to its high sequence similarity with AbyU. The table compares the number of unique hits retrieved by A0A246RFQ9 and the subject, allowing for an assessment of how the query's high sequence similarity influences the retrieval of unique hits. This analysis provides insight into the relationship between sequence similarity and the distinctiveness of the hits retrieved.

Query	Per query	Only 3Di	Only TM-align	Total
AbnU	694	720	780	1500
AbyU	1357	1126	1219	2345
Cyc15	2054	1463	1585	3048
Pyrl4	1538	1130	1224	2354

TABLE 4 – Unique Hit Distribution Across Queries: Foldseek Method A Analysis. Unique hits retrieved by each alignment or query using Foldseek (Method A). Unique hits are defined as those retrieved exclusively by individual alignments or queries. These hits were obtained by merging all the results from Foldseek and comparing the hits retrieved by each alignment and query. The table highlights the distinct hits retrieved by each query, showcasing the unique retrieval profiles based on the Foldseek analysis.

Query	Cluster representative	Total members
AbnU	A0A0K9XAS6	78
AbyU	A0A0K9XAS6	78
Additional	A0A0B6VM83	37
Additional	A0A022Q700	4092
Additional	A0A2P1BT29	2
Additional	A0A4Y8Y1N5	3
Additional	A0A4P7ZYP8	2
Additional	A0A833W1H9	49
Additional	L8H095	126
Cyc15	A0A0K9XAS6	78
Pyrl4	A0A3A9ZPZ1	38
	Total hits	4583

TABLE 5 – Query Distribution Across Clusters: Method B and Search-Derived Additions. Cluster analysis of queries using Method B. In this method, each query is assigned to a cluster, with each cluster represented by a designated representative sequence. The table presents the number of members in each cluster, reflecting the distribution of queries within their respective clusters. Additional clusters identified during the query searches were included in the analysis, providing a comprehensive overview of the cluster memberships and the additional clusters discovered during the process.

Method	Total hits (relevant)	Unique hits	Positive control
Method A	120	76	AbmU, AbsU, AbnU, Cyc15, AbyU, ChlL, TcaU4, LonU2, VstJ, LobD1, PyrI4, Tmn8, Tsn15
Method B	45	5	AbmU, AbnU, AbyU, Cyc15, QmnH, ChlL, TcaU4, Tmn8, Tsn15

TABLE 6- Unique Hit Retrieval and Positive Control Detection: Method A vs. Method B. Comparison of unique hits obtained by both Method A (Foldseek) and Method B. The table shows the count of unique hits retrieved by each method. Additionally, it includes an assessment of the presence of positive controls within the results, providing insight into the effectiveness and accuracy of each method in identifying unique hits and detecting positive controls.

Search type	Unique hits	Total hits
Sequence-based search	36	76
Structure-based search	58	98

TABLE 7- Comparative Effectiveness: Structure vs. Sequence in Hit Retrieval. Analysis of hits obtained from Zorn *et al*., (sequence hits) with variant hits excluded. The table presents the count of unique hits retrieved by each method, after discarding any variant hits reported by Zorn et al. A comparative analysis of the unique hits retrieved by each method is provided, offering insights into the consistency and effectiveness of the methods in identifying distinct hits. This table effectively compared the two different approached of using sequence or structure for doing a homolog search.

Uniprot ID	Salt bridge (78 + 122 AbyU)
A0A0B6VRF8	A-R
A0A0F7N0P6	Y-L
A0A0G3A932	E-R
A0A0J1H8G7	E-R
A0A0K9XAS6	E-R
A0A0L8NIA4	E-R
A0A0M2RZN8	N-R
A0A0M8XYC3	E-R
A0A0Y0AWW3	E-R
A0A101JBC1	E-R
A0A109B5R0	E-R
A0A132CHM2	E-R
A0A168IKI6	E-R
A0A1B4FKD2	E-R
A0A1B4FUY1	E-R
A0A1B4SIL4	E-R
A0A1C3HDP6	E-R
A0A1C4PF23	E-E
A0A1C6V0W8	T-M
A0A1J4Q2U2	A-R
A0A1M7FPM3	Y-R
A0A1Q4P580	Y-R
A0A1Q5TWR6	Y-R
A0A1Q7WCG8	D-Y
A0A1S1QVM7	E-R
A0A1S1R073	Y-R
A0A1S1R326	E-R
A0A1S2PAK5	L-M
A0A1Y0RPR1	L-I
A0A1Y2SP38	Y-R
A0A239CWW2	E-R
A0A246RFQ9	E-R
A0A248YPZ6	I-M
A0A2P1BT29	Y-R
A0A2P4UBX2	I-W
A0A2S6GLP7	Y-R
A0A2S6GLP9	R-R
A0A2S6H1S2	F-R
A0A318JTK0	D-T
A0A345XXX2	Y-R
A0A367F712	S-M
A0A3A9YQ06	I-R
A0A3A9ZPZ1	S-E
A0A3D1PI25	E-R

A0A3E0H6P9	F-R
A0A401MXE6	V-R
A0A401MXH6	V-R
A0A438LZ58	T-E
A0A495Q9P3	Y-R
A0A4R2JL81	Y-R
A0A4R4MHI3	
A0A4R5A4Q9	Y-R
A0A4R5E5X7	T-L
A0A4U3F911	Y-R
A0A4U3LVR5	Y-R
A0A4Y8Y1N5	Y-R
A0A518WFP6	E-R
A0A544YAL3	E-R
A0A5D0NLL8	A-R
A0A5P8WEY5	E-I
A0A5P9PSM5	L-M
A0A5Q0GZD2	L-M
A0A5R8YLG6	F-R
A0A5S4GQR6	F-R
A0A5S4H4D6	V-R
A0A5S8WF63	V-L
A0A640SQ86	E-R
A0A640SVV6	E-R
A0A6G3UFY1	Y-R
A0A6G9F6I2	L-M
A0A6I4MT57	S-L
A0A6N7ZAT6	F-R
A0A6N7ZB01	S-R
A0A743Z1Y4	E-R
A0A7D3ZDF8	V-L
A0A7J5DHM5	E-R
A0A7K2LH80	E-R
A0A7K2N5N2	E-M
A0A7U4PB97	E-R
A0A7W4FA88	E-R
A0A7X0IJX5	E-R
A0A7Y0J7D7	E-R
A0A7Y0WQ23	E-R
A0A8B4GMM4	E-R
A7BEY9	S-L
F4F7G1	E-R
F8AWP7	E-R
F8K320	N-M
K4FDH3_2	
K4FDH3_1	E-R

K7QVW7	S-
L7RSB5	A-R
Q0R4M0	S-M

TABLE 8- Comparative Analysis of Salt Bridges at AbyU-Equivalent 78-122 Positions. Evaluation of the salt bridge composition between Glu (78) and Arg (122) AbyU positions. The presence of this salt bridge was assessed through structural alignment of the hits. The table details the alignment results and the subsequent visualization of the structure in PyMOL to confirm the position of the salt bridge. This analysis provides insight into the structural features of the hits and their relevance to the salt bridge configuration compared to that of AbyU.

UniProt ID	Conformation in AFDB	Conformation in PDB
A0A0B6VRF8	Open	Not available
A0A0F7N0P6	Open	Not available
A0A0G3A932	Closed	Not available
A0A0J1H8G7	Closed	Not available
A0A0K9XAS6	Closed	Not available
A0A0L8NIA4	Closed	Not available
A0A0M2RZN8	Open	Not available
A0A0M8XYC3	Closed	Not available
A0A0Y0AWW3	Closed	Not available
A0A101JBC1	Closed	Not available
A0A109B5R0	Closed	Not available
A0A132CHM2	Closed	Not available
A0A168IKI6	Closed	Not available
A0A1B4FKD2	Closed	Not available
A0A1B4FUY1	Closed	Not available
A0A1B4SIL4	Closed	Not available
A0A1C3HDP6	Closed	Not available
A0A1C4PF23	Open	Not available
A0A1C6V0W8	Open	Not available
A0A1J4Q2U2	Open	Not available
A0A1M7FPM3	Open	Not available
A0A1Q4P580	Closed	Not available
A0A1Q5TWR6	Closed	Not available
A0A1Q7WCG8	Open	Not available
A0A1S1QVM7	Closed	Not available
A0A1S1R073	Closed	Closed
A0A1S1R326	Closed	Not available
A0A1S2PAK5	Closed	Not available
A0A1Y0RPR1	Open	Not available
A0A1Y2SP38	Closed	Not available
A0A239CWW2	Closed	Not available
A0A246RFQ9	Closed	Not available
A0A248YPZ6	Open	Not available
A0A2P1BT29	Open	Open
A0A2P4UBX2	Open	Not available

A0A2S6GLP7	Closed	Not available
A0A2S6GLP9	Closed	Not available
A0A2S6H1S2	Closed	Not available
A0A318JTK0	Loop not defined	Not available
A0A345XXX2	Open	Not available
A0A367F712	Open	Not available
A0A3A9YQ06	Loop not defined	Not available
A0A3A9ZPZ1	Open	Not available
A0A3D1PI25	Loop not defined	Not available
A0A3E0H6P9	Closed	Not available
A0A401MXE6	Closed	Open
A0A401MXH6	Closed	Not available
A0A438LZ58	Open	Not available
A0A495Q9P3	Closed	Not available
A0A4R2JL81	Closed	Not available
A0A4R4MHI3	Closed	Not available
A0A4R5A4Q9	Open	Not available
A0A4R5E5X7	Open	Not available
A0A4U3F911	Closed	Not available
A0A4U3LVR5	Closed	Not available
A0A4Y8Y1N5	Open	Not available
A0A518WFP6	Closed	Not available
A0A522WLH4	Non-beta barrel fold	Not available
A0A544YAL3	Closed	Not available
A0A5C8PZY7	Open	Not available
A0A5D0NLL8	Closed	Not available
A0A5P8WEY5	Loop not defined	Not available
A0A5P9PSM5	Closed	Not available
A0A5Q0GZD2	Closed	Not available
A0A5R8YLG6	Closed	Not available
A0A5S4GQR6	Closed	Not available
A0A5S4H4D6	Open	Not available
A0A5S8WF63	Open	Not available
A0A640SQ86	Open	Not available
A0A640SVV6	Closed	Not available
A0A6G3UFY1	Open	Not available

A0A6G9F6I2	Open	Not available
A0A6I4MT57	Open	Not available
A0A6N7ZAT6	Open	Not available
A0A6N7ZB01	Closed	Not available
A0A743Z1Y4	Closed	Not available
A0A7D3ZDF8	Open	Not available
A0A7J5DHM5	Open	Not available
A0A7K2LH80	Closed	Not available
A0A7K2N5N2	Closed	Not available
A0A7U4PB97	Closed	Not available
A0A7W4FA88	Closed	Not available
A0A7X0IJX5	Closed	Not available
A0A7Y0J7D7	Closed	Not available
A0A7Y0WQ23	Closed	Not available
A0A838IMS4	Non-beta barrel fold	Not available
A0A8B4GMM4	Closed	Not available
A7BEY9	Open	Not available
F4F7G1	Closed	closed and open
F8AWP7	Closed	Not available
F8K320	Closed	Not available
K4FDH3_1	Open	Not available
K4FDH3_2	Open	Not available
K7QVW7	Open	closed and open
L7RSB5	Open	Not available
Q0R4M0	Open	Not available

TABLE 9- Evaluation of Hit Conformations and Their Representation in the PDB. Assessment of structural conformations and availability in the PDB. The table categorizes hits based on their structural conformations: some exhibited a closed conformation, while others lacked a defined capping loop, making their conformation indeterminate. Additionally, the presence of these structures in the Protein Data Bank (PDB) was evaluated, providing insights into the structural characteristics and reported conformations.

UniProt ID	Template	Result
A0A0F7N0P6	AbyU	Open
A0A0F7N0P6	Cyc15	Open
A0A0M2RZN8	AbyU	Open
A0A0M2RZN8	AbyU	Open
A0A1C6V0W8	AbyU	Open
A0A1M7FPM3	AbyU	Open
A0A1M7FPM3	AbyU	Open
A0A1Q7WCG8	AbyU	Open
A0A1S2PAK5	AbyU	Open
A0A1S2PAK5	AbyU	Open
A0A1Y0RPR1	AbyU	Open
A0A1Y0RPR1	Cyc15	Open
A0A248YPZ6	AbyU	Open
A0A2P4UBX2	AbyU	Open
A0A345XXX2	AbyU	Closed
A0A367F712	AbyU	Open
A0A3A9ZPZ1	AbyU	Open
A0A438LZ58	AbyU	Open
A0A4R5E5X7	AbyU	Open
A0A4Y8Y1N5	AbyU	Open
A0A5P8WEY5	Cyc15	Open
A0A5Q0GZD2	AbyU	Open
A0A5S8WF63	AbyU	Open
A0A6G9F6I2	AbyU	Open
A0A6I4MT57	AbyU	Open
A0A6I4MT57	AbyU	Open
A0A7D3ZDF8	AbyU	Open
A0A7J5DHM5	AbyU	Open
A0A7K2N5N2	AbyU	Open
A7BEY9	AbyU	Open
F8K320	AbyU	Open

TABLE 10- Template-Based Closed Conformation Analysis: AbyU and Cyc15 in ColabFold. Closed conformation prediction of structural hits performed with ColabFold. AbyU and Cyc15 were used as templates for the predictions, selected based on their sequence similarity to the subject (refer to Figure SX for details). This table summarizes the predictions made using these templates to assess the closed conformations of the structural hits.

Test	Model	Subject (UniProt ID)	Template	Num_seeds	dropout	max_msa	Result
Test_1_1	Model 1	A0A1Y0RPR1	AbyU	2	checked	256:512	no change
Test_1_2	**Model 2**	A0A1Y0RPR1	AbyU	2	checked	512:1024	no change
Test_1_3	Model 1	A0A1Y0RPR1	AbyU	2	checked	64:128	no change
Test_1_4	Model 2	A0A1Y0RPR1	AbyU	2	checked	32:64	no change
Test_1_5	Model 2	A0A1Y0RPR1	AbyU	2	checked	16:32	no change
Test_2_1	Model 1	A0A1Y0RPR1	AbyU	16	checked	256:512	alternative conformation
Test_2_2	Model 1	A0A1Y0RPR1	AbyU	16	checked	16:32	alternative conformation
Test_2_3	Model 1	A0A1Y0RPR1	AbyU	16	checked	512:1024	alternative conformation
Test_2_4	Model 2	A0A1Y0RPR1	AbyU	16	checked	64:128	alternative conformation
Test_2_5	Model 3	A0A1Y0RPR1	AbyU	16	checked	32:64	alternative conformation
Test_3_1	Model 1	A0A1Y0RPR1	AbyU	2	checked	default	no change
Test_3_1	Model 2	A0A1Y0RPR1	AbyU	2	checked	default	alternative conformation
Test_3_1	Model 3	A0A1Y0RPR1	AbyU	2	checked	default	alternative conformation
Test_3_1	Model 4	A0A1Y0RPR1	AbyU	2	checked	default	no change
Test_4_1	Model 1	A0A1Y0RPR1	AbyU	4	checked	default	no change
Test_4_1	Model 2	A0A1Y0RPR1	AbyU	4	checked	default	alternative conformation
Test_4_1	Model 3	A0A1Y0RPR1	AbyU	4	checked	default	alternative conformation
Test_4_1	Model 4	A0A1Y0RPR1	AbyU	4	checked	default	alternative conformation
Test_5_1	Model 1	A0A1Y0RPR1	AbyU	8	checked	default	alternative conformation
Test_5_1	Model 2	A0A1Y0RPR1	AbyU	8	checked	default	no change
Test_5_1	Model 3	A0A1Y0RPR1	AbyU	8	checked	default	alternative conformation
Test_5_1	Model 4	A0A1Y0RPR1	AbyU	8	checked	default	alternative conformation
Test_6_1	Model 1	A0A1Y0RPR1	AbyU	16	checked	default	no change
Test_6_1	Model 2	A0A1Y0RPR1	AbyU	16	checked	default	no change
Test_6_1	Model 3	A0A1Y0RPR1	AbyU	16	checked	default	no change
Test_6_1	Model 4	A0A1Y0RPR1	AbyU	16	checked	default	no change

TABLE 11- Optimized ColabFold Predictions: Assessing Structural Hit Conformations. Closed conformation prediction of structural hits using ColabFold. The predictions were made with adjustments to the multiple sequence alignment (MSA) depth and number of seeds, based on prior recommendations. For Test 1 and Test 2, the highest-scoring models were selected for visualization. In Tests 4, 5, and 6, the four highest-scoring models were visualized. A conformation was classified as alternative if the capping loop was positioned differently compared to the original subject.

	Total	Mean volume
Hits in a closed conformation	54	
Hits in a closed conformation with the salt bridge	51	124
Hits in a closed conformation without the salt bridge	3	229
Hits in an open conformation	34	
Total hits	96	

TABLE 12- Comparative Cavity Volume Assessment of Closed Conformation Hits Using CastP.

Calculation of cavity volumes for hits in a closed conformation using CastP. The table includes the volume measurements for hits with a closed conformation, with separate calculations provided for hits lacking the salt bridge (or without Arg at position 122). This distinction allows for the evaluation of how the presence or absence of the salt bridge affects the cavity volumes of the hits.

6.2 Figures

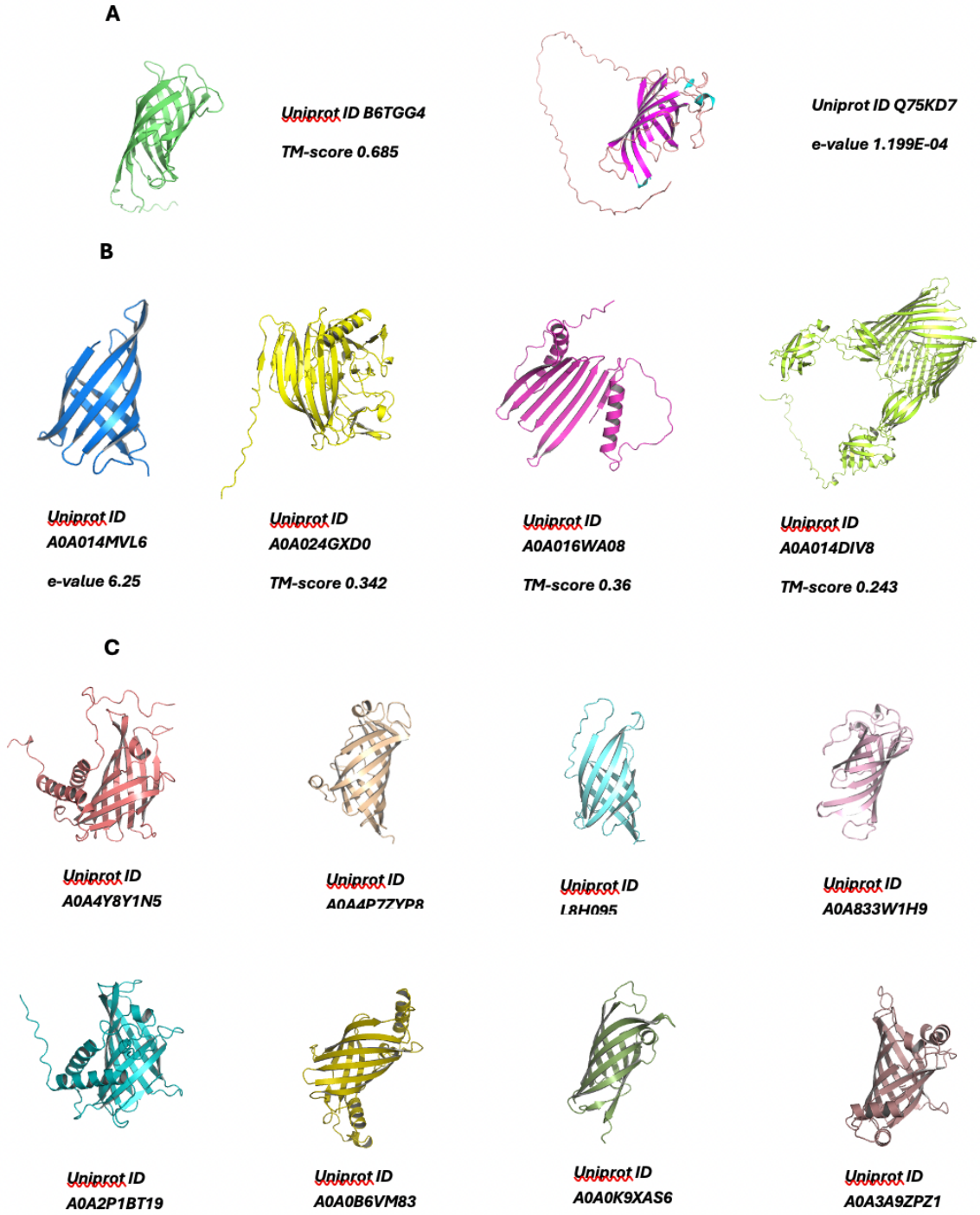


FIGURE 1- Overview of unique hits and cluster representatives retrieved by alignment methods.

(A): Unique hits retrieved by the alignment method. These hits display similar structures, indicating no particular retrieval advantage for each alignment method. Additionally, these hits are considered irrelevant to the study.

(B): Unique hits retrieved by individual queries. These hits also exhibit non-beta barrel folds or incomplete beta barrels, rendering them irrelevant. This finding supports the accuracy of information density capture by Foldseek, as these hits have lower e-values and TM scores, demonstrating that Foldseek correctly identifies relevant structures.
(C): Cluster representatives of each cluster retrieved by Method B. The representatives display similar structures to the queries, even though some are uncharacterized proteins. This indicates that Method B effectively clusters proteins with similar structural features.

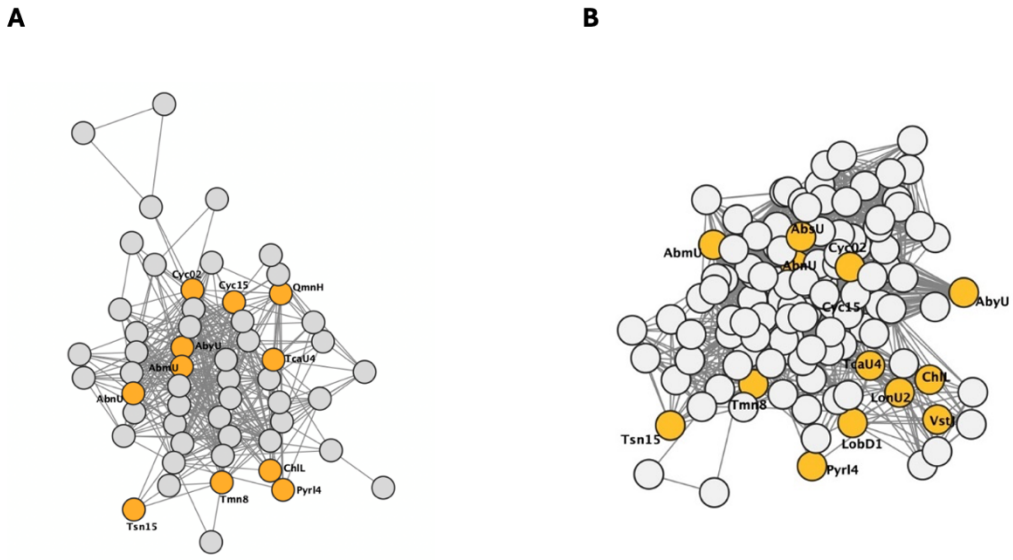


FIGURE 2- Comparison of Sequence Similarity Networks (SSNs) for Method A and Method B.

(A): SSN for Method A (Foldseek), displaying all positive controls clustered with other hits in a separate cluster. The positive controls identified in this network include: AbmU, AbsU, AbnU, Cyc15, AbyU, ChlL, TcaU4, LonU2, VstJ, LobD1, Pyrl4, Tmn8, and Tsn15. Hits within this cluster are considered relevant as they are grouped with the positive controls.

(B): SSN for Method B (AFDB Clusters), showing all positive controls clustered with other hits in a separate cluster. The positive controls in this network are: AbmU, AbnU, AbyU, Cyc15, QmnH, ChlL, TcaU4, Tmn8, and Tsn15. Hits within this cluster are considered relevant as they are grouped with the positive controls.

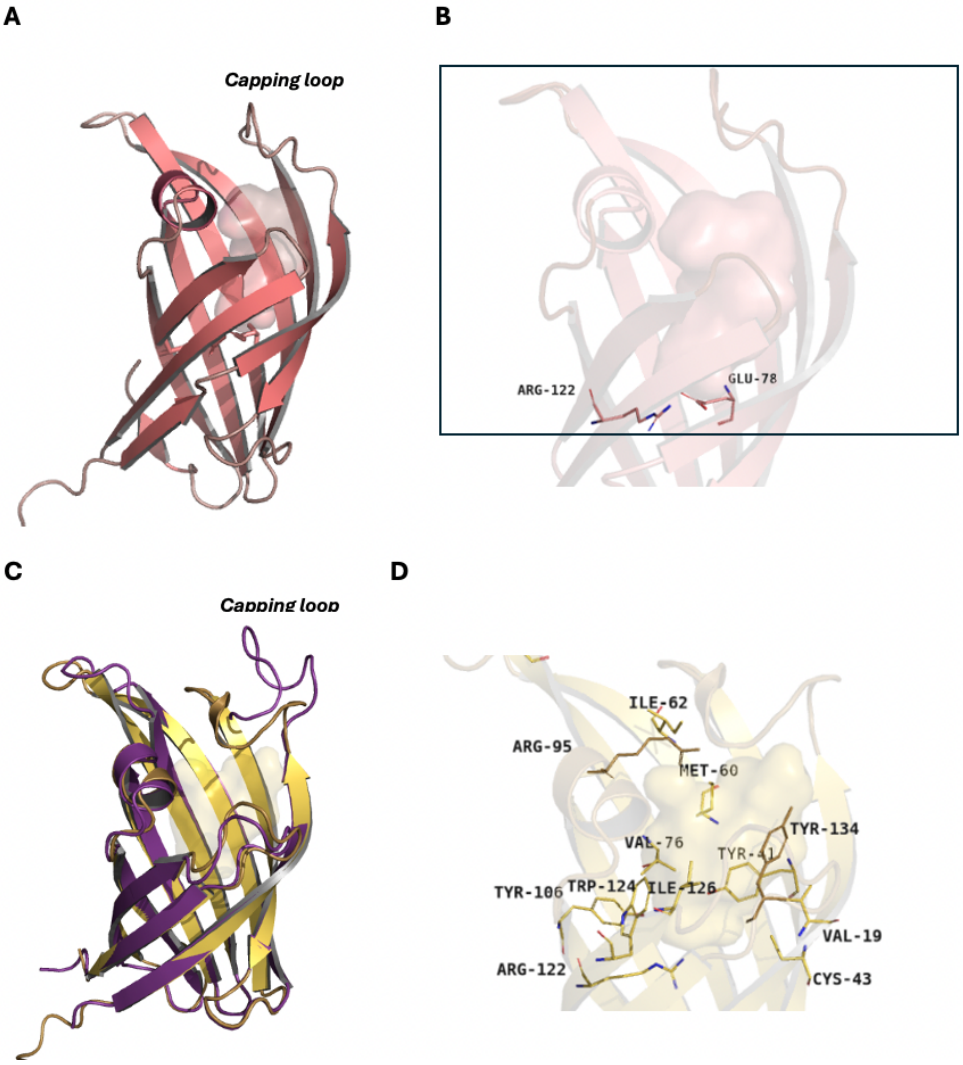


FIGURE 3- Structural visualization of AbyU and Cyc15 with emphasis on cavities and salt bridges.

(A): Complete structure of AbyU, highlighting the cavity delimited by the salt bridge between Glu 78 and Arg 122. The salt bridge is represented by sticks, and the position of the capping loop is indicated.

(B): Close-up view of the AbyU structure, focusing on the surface representation of the cavity and the salt bridge. This detailed view provides a closer look at the cavity's dimensions and the salt bridge.

(C): Comparison of Cyc15 in its open conformation (PDB 8OF7) and closed conformation (Alpha Fold prediction based on template UniProt A0A0K9XAS6). The closed conformation is shown in yellow and the open conformation in purple.

(D): Detailed close-up of Cyc15, highlighting the cavity-lining residues that define the cavity in the closed conformation.

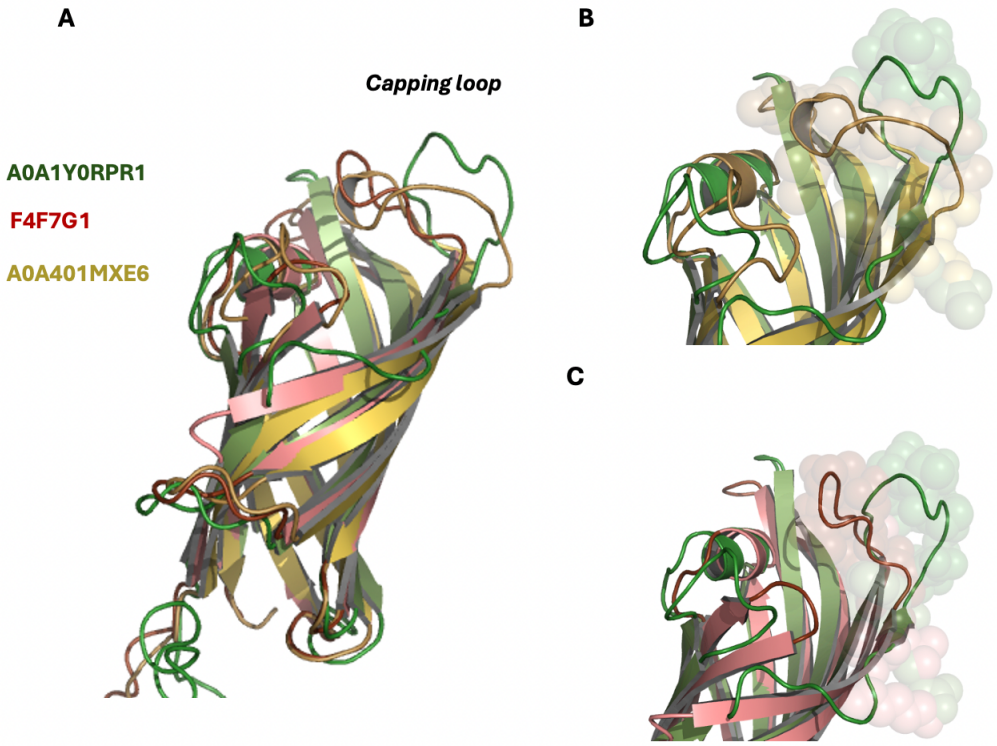


FIGURE 4- Closed conformation predictions using Cyc15 and AbyU.
(A): Alignment of A0A1Y0RPR1 with AbyU and Cyc15, highlighting the capping loop in sphere representation. This panel provides a comparative overview of the alignment and the positioning of the capping loop when using each template.
(B): Close-up view of the capping loop of A0A1Y0RPR1, showing that its conformation did not change when Cyc15 was used as a template. This panel demonstrates the stability of the capping loop conformation with the Cyc15 template.
(C): Close-up view of the capping loop of A0A1Y0RPR1, showing that its conformation did not change when AbyU was used as a template. This panel highlights the consistency of the capping loop conformation with the AbyU template.
The predictions with varying parameters were conducted after these attempts were unsuccessful, indicating the need for further adjustments.

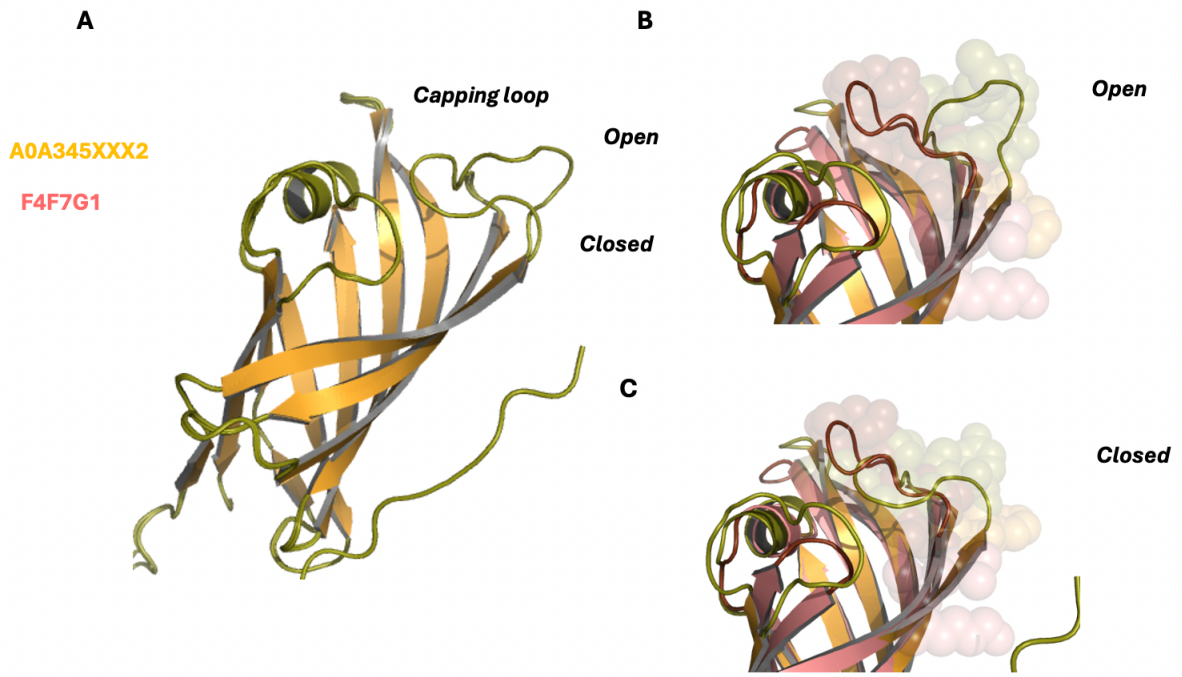


FIGURE 5- Successful closed conformation prediction of A0A345XXX2 using AbyU as a template.

(A): Structure of A0A345XXX2 aligned with AbyU, highlighting the capping loop. Initially, A0A345XXX2 had an open loop conformation. Using AbyU as a template with default parameters in ColabFold resulted in a successfully predicted closed conformation.

(B): Close-up view showing the open conformation of the capping loop before the alignment with AbyU. This panel illustrates the initial structure of the capping loop.

(C): Close-up view of the closed conformation obtained for A0A345XXX2 using AbyU as a template. This panel highlights the successfully predicted closed state of the capping loop.

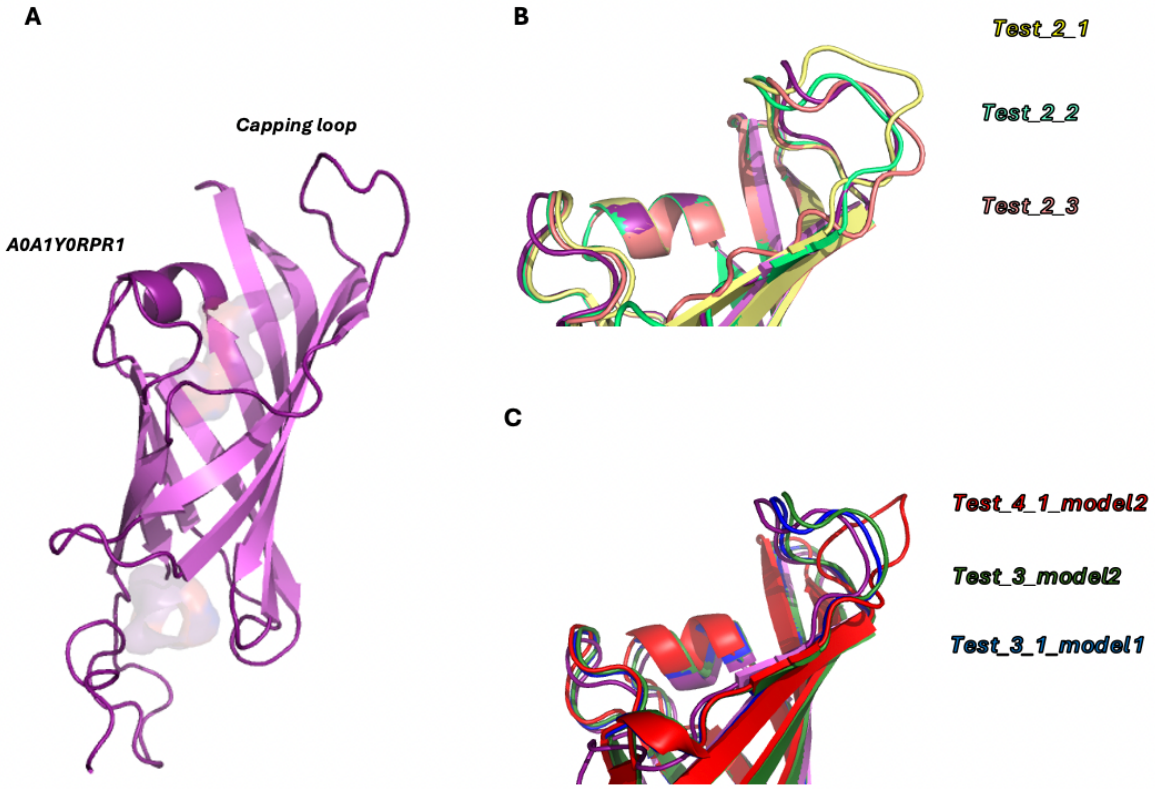


FIGURE 6- Prediction of the closed conformation of A0A1Y0RPR1 through variations in MSA depth and number of seeds.

(A): Structure of A0A1Y0RPR1 depicted in pink, with its cavity shown in surface representation. This view provides a general overview of the protein structure and its internal cavity.

(B): Close-up of the capping loop of A0A1Y0RPR1, aligned with the results from Test_2_1, Test_2_2, and Test_2_3. This panel illustrates the different conformations of the capping loop resulting from variations in MSA depth and the number of seeds used in the predictions.

(C): Results from Test_4_1_Model2, Test_3_Model2, and Test_3_1_Model1, where only the number of seeds was varied while the MSA depth remained constant. This panel demonstrates that the capping loop did not acquire significantly different conformations compared to the results in panel (B), highlighting the relative stability of the loop conformation with changes in the number of seeds.

Appendix

Supplementary Tables

Query Uniprot ID	Subject Uniprot ID	Identity (%)	Coverage (%)
A0A0F7N0P6	A0A3S8VII0	100	100
	A0A2K9FB81	100	100
A0A0M2RZN8	A0A0B6VM83	100	100
A0A6G3UFY1	A0A1H5HAZ2	100	100
	A0A6G2Z021	100	100
	A0A6I7XZ82	100	100
	A0A327URC2	100	100
	A0A6B3BH80	99	100
	A0A4S2UVU8	97	100
B5L6K7	A0A1S8XP27	99	100
A0A7Y0J7D7	A0A6G3VZ74	100	100
	A0A6I5GVW7	100	100
	A0A7K2LZ25	100	100
	A0A1V0QH49	100	100
	D6K898	99	100
F4F7G1	A0A097CSY6	100	100
M9T3W4	A0A5P9NYB6	100	100
	L7RSB5	97	100
	A0A6G9ENB7	96	100
A0A0L8NIA4	A0A0N0T4L8	100	100
A0A5S8WF63	A0A5C8PZY7	100	99
	A0A6C0QDS2	100	99
A0A5R8YLG6	A0A5D0QA99	100	100
A0A1B4FKD2	A0A7T2X384	100	96
L8EX63	A0A7K2RBB9	100	100
	A0A429JRF8	100	100
	A0A429I7L3	100	100
A0A5D0NLL8	A0A5D0UIN8	99	100
	A0A5D3FXF0	99	100
A0A2P1BT29	A0A4Q6VF37	99	100
A0A4U3LVR5	A0A7C9NN96	100	100
A0A1S1QVM7	A0A166S4A2	97	100
A0A2P4UBX2	A0A6I4WC81	97	100
A0A367F712	A0A367ECF2	97	100
A0A4R4MHI3	A0A7W7IEZ4	97	100
	A0A4R4TSZ3	97	100
A0A1C4PF23	A0A7K2G773	100	100
A0A0M8XYC3	A0A7K3G8J5	96	100
A0A4R5E5X7	A0A4R4TSZ3	96	100

TABLE S1- Variants in structural hits found by BlastP

This table presents an analysis of structural variants found in protein sequences identified as hits through BLASTP searches.

Cyclase name	Unitprot ID	Reference
AbyU	F4F7G1	
AbmU	A0A2P1BT29	(Li et al., 2022)
AbnU	A0A1S1R073	(Kashyap et al., 2021)
AbsU	A0A1V0QH49	(Xu and Yang, 2021)
Cyc15	A0A401MXE6	(Zorn et al., 2023)
Pyrl4	K7QVW7	(Tian et al., 2015)
LobD1	L7RSB5	(Yue et al., 2016)
Tsn15	A0A5S8WF63	(Fujiyama et al., 2021)
TcaU4	B5L6K7	(Fang et al., 2008)
VstJ	A0A0B6VRF8	(Hashimoto et al., 2015)
Tmn8	A7BEY9	(Demydchuk et al., 2008)
ChlL	Q0R4M0	(Jia et al., 2006)
QmnH	K4FDH3	(Jeon et al., 2017)

TABLE S2- Positive control

Cyclase name	Uniprot ID	Fasta
AbnU	A0A1S1R073	MSVVHEGIWEPQIRNEQNVNVADPQVGQIGSYYDEL YDSSRELLGITIGRYEIRYKKVGGAVLTYYSEDLFLRD GIIHAEGWADFNDVKNGVWVGYPAVGLDGVYRGLD GRREWRVIEPDQPVEARISLHG
AbyU	F4F7G1	MTERLETRPQALLIKVPTEIVVKVVDDVDVAAPAVGQ VGKFDDELYDEAGAQIGTSSGNFRIEYVRPTDGGLLT YYQEDITLSDGVIHAEGWADFNDVRTSKWVFYPATG VSGRYLGLTGFRQWRMTGVRKSAEARILLGE
Cyc15	A0A401MXE6	MGDTTTANDVVTVELVEKVTKKDLNESGSIEGFGPG MMATYWCDVFDTEGKHIGTTVGCMDILYADPESGHL VEHVAEQIRLPDGTIMAWGTMNRSDVLAQKWITYRC QGTSGRYAGLVGTRTWRIQSLEDESYPIVAKMELRG A
Pyrl4	K7QVW7	MTTPQIDERAMEAGAAALQETIVDPGPLDVTALAVAA ALAAGLHSAADDPAAALDKCIVLDELTEFAEKLVVHD RPGGIGTTVEYVEVYEDASGVRLGTATGNAVVLKME PHMWQFHQSVSELADGSFEAVGVIDCTAMLRRMTQ VLRVTGRSGRYAGKSGFMTLAISDPNQRPPHYSVQV VLC

TABLE S3- fasta sequences of queries AbnU, AbyU, Cyc15 and Pyrl4

Method	Minimum length of sequence	Maximum length of sequence	Alignment score threshold
Method A	120	625	7
Method B	120	625	7

TABLE S5- Parameters used for Sequence similarity Network visualization.

Query	Total hits per query
AbnU	23
AbyU	21
Cyc15	23
Pyrl4	18
Total hits	85

TABLE S6- Hits retrieved by method C

Query	Total hits
AbyU	10
List of hits	
Uniprot ID	**PDB**
A0A1S1R073	7DVI
A0A1V1FH01	6OOC
A0A2P1BT29	6LE0
A0A401MXE6	8OF7
A0A5S8WF63	6NNW
A0A9Y2YAD1	7X7Z
F4F7G1	5DYQ
K0V2D8	5K21
K7QVW7	5BTU
Q8H0N6	4H69

TABLE S7- Hits retrieved by method D

Supplementary Figures

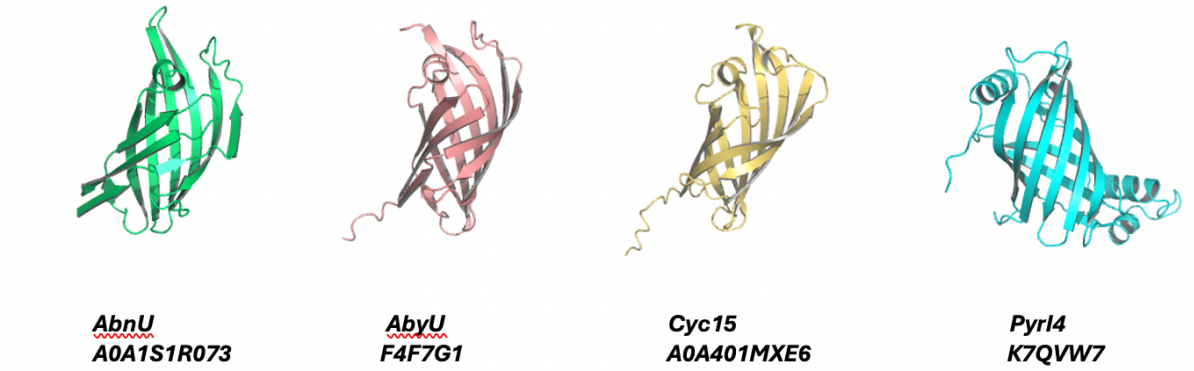


FIGURE S1- PDB structures of AbnU, AbyU, Cyc15 and Pyrl4

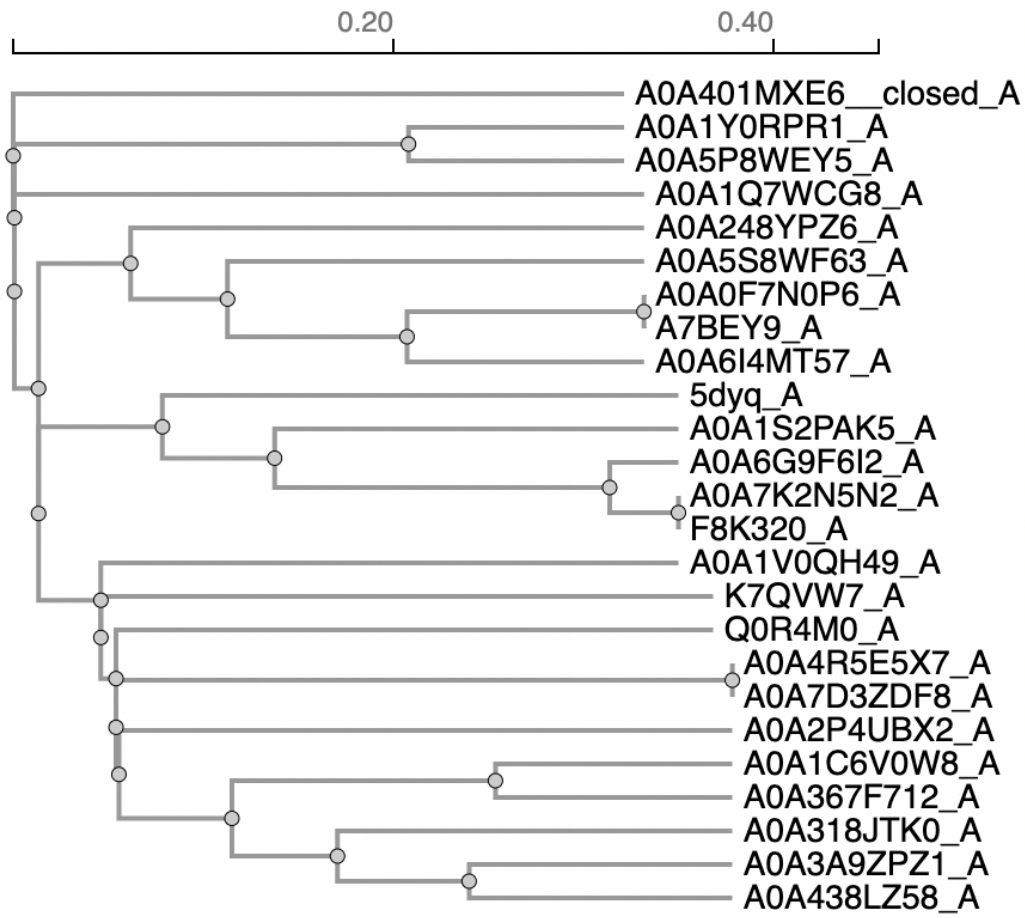
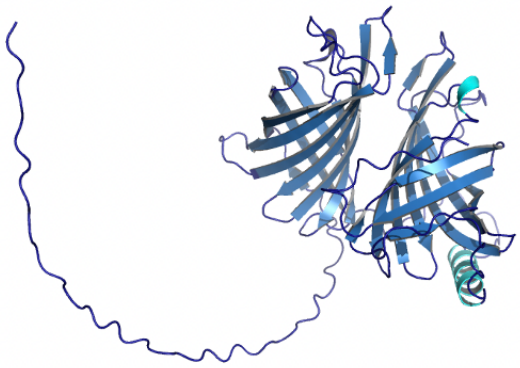
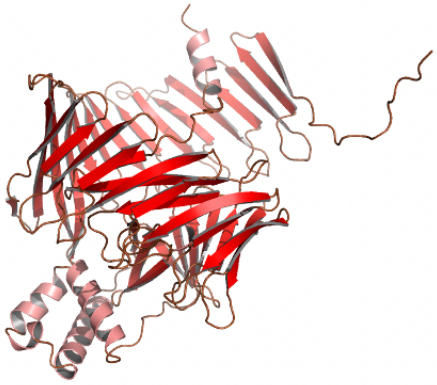


FIGURE S2- MSA visualized by Tree (Clustal Omega).

K4FDH3



B
A0A838IMS4



C
A0A522WLH4

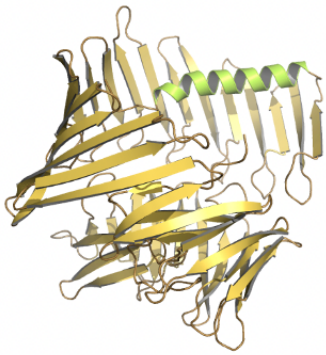


FIGURE S3- Relevant hits that do not display a B-barrel fold

1432

1433