Large Scale Discovery of Microbial Fibrillar Adhesins and Identification of Novel Adhesive Domain Families

Vivian Monzon[1] and Alex Bateman[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton. CB10 1SD.

Abstract

Fibrillar adhesins are bacterial cell surface proteins that mediate interactions with host cells during colonisation and with other bacteria during biofilm formation. These proteins are characterised by a stalk that projects the adhesive domain closer to the binding target. Fibrillar adhesins evolve quickly and thus can be difficult to computationally identify, yet they represent an important component for understanding bacterial host interactions.

To detect novel fibrillar adhesins we developed a random forest prediction approach based on common characteristics we identified for this protein class. We applied this approach to Firmicute and Actinobacterial proteomes, yielding over 4,000 confidently predicted fibrillar adhesins. To verify the approach we investigated predicted fibrillar adhesins that lacked a known adhesive domain. Based on these proteins, we identified 21 sequence clusters representing potential novel adhesive domains. We used AlphaFold to verify that 14 clusters showed structural similarity to known adhesive domains such as the TED domain.

Overall our study has made a significant contribution to the number of known fibrillar adhesins and has enabled us to identify novel adhesive domain families involved in the bacterial pathogenesis.

Keywords

Fibrillar adhesins
Host-pathogen interaction
RandomForest classification
Protein domain families
Adhesive domains
Structure prediction methods
AlphaFold2

Introduction

Fibrillar adhesins are an important class of bacterial surface proteins, which are expressed by a wide range of bacterial species to mediate binding interactions. Essential binding targets include different host surface structures, such as extracellular matrix proteins. A pathogenic colonization and infection can occur as a consequence of the binding interactions to host cells [1]. Single fibrillar adhesins have therefore been studied in depth and have been the focus point for anti-adhesion therapies [2, 3]. Fibrillar adhesins have also been described to mediate biofilm formation [4, 5].

Fibrillar adhesins are a recently defined class of proteins that led to a domain-based characterization and identification of these proteins across a wide range of bacterial species [6, 7]. Key characteristics of fibrillar adhesins are their large length with an adhesion region, a rod-like repetitive region and a cell surface anchor. The repetitive region contains repeating protein domains, also called stalk domains, which fold into a filamentous stalk. It has been suggested that the stalk projects the binding region closer to the binding target [8] and enables the adhesive region to be presented outside the cell by reaching beyond the surface layer. The repeats can vary in number leading to fibrillar adhesins with stalks of different lengths. Proteins with varying repeat number of stalk domains between related bacterial strains have recently been termed 'Periscope proteins' [9]. Whelan *et al.* propose that the varying length is used as a regulation mechanism to facilitate the binding to targets at different distances [9]. The varying length may also lead to the adhesive region being differentially displayed beyond the surface layer based on the number of repeats. Several adhesive domains have been identified when studying the binding regions of bacterial adhesins, of which most bind to protein ligands, some bind to carbohydrates and one adhesive domain is known to bind to ice crystals [10]. Nevertheless, undoubtedly a large number of adhesive domains remain undiscovered.

In our previous work, we have used the presence of known adhesive and stalk domains identified in fibrillar adhesins to detect more than 3,000 fibrillar adhesins-like (FA-like) proteins across all bacterial species based on profile hidden Markov models (HMM)-searches [7]. The limitation of this domain-based discovery approach is that only FA-like proteins with known adhesive and stalk domains are found. Not all adhesive and stalk domains are identified yet and finding novel binding proteins or domains is important for the understanding of emerging bacterial interactions. To overcome this limitation we studied the properties of FA-like proteins and developed a random-forest based discovery approach. The aim of this study is to enable the identification of FA-like proteins on a large-scale, including those lacking a known adhesive or stalk domain. We apply our newly developed machine learning approach on the Firmicute and Actinobacteria UniProt Reference Proteomes and verify the approach by predicting the structure of predicted Firmicute FA-like proteins lacking a known adhesive domain with AlphaFold [11]. Our approach facilitates the identification of relevant proteins during bacterial infection processes, enabling the investigation of novel adhesive domains leading to a better understanding of microbial interaction mechanisms.

Results

Random Forest classification

We decided to concentrate on the Firmicute and Actinobacteria phyla in this study. Fibrillar adhesins are best studied in Firmicutes and the cell surface composition and FA-like protein architecture of the Actinobacteria resembles those of the Firmicutes [2, 3, 7]. We created a

positive training set based on the FA-like proteins identified in our previous study [7]. For the negative training set we randomly selected non-FA-like proteins from eight Firmicute and Actinobacteria Reference Proteomes, in which FA-like proteins were detected in our previous study [7]. Although an individual proteome may contain several FA-like proteins, they are present at relatively low numbers per proteome (0% to 1.47%), thus randomly selecting the negative examples is unlikely to select many, if any, true FA-like proteins [7]. In total, the training set consists of 3,332 proteins, equally balanced between the positive and negative training sets. Using this training set, common properties for this protein class were determined.

Nearly all proteins of the positive training set (98%) have at least one adhesive and one stalk domain, whereas less than 0.1% of the proteins in the negative training set possess a known adhesive or stalk domain (Figure S1a,b). Hence, the adhesive and stalk domains are the strongest identification features for this protein class (Figure 1). To increase the chance to detect proteins with unidentified stalk domains, we selected the presence of tandem sequence repeats with a minimum of 70% sequence identity as an identification feature. We used a tool called T-REKS to identify these tandem sequence repeats [12]. In periscope proteins the stalk domains are described to be found in highly identical tandem repeats [9]. But compared to periscope proteins, not all stalk domains in FA-like proteins are found in highly identical tandem repeats and are therefore missed by T-REKS. We determined that 16% of the positive training data proteins are detected by T-REKS, and 99% of the proteins detected by T-REKS belong to the positive training data set (Figure S2a). We observed that the stalk domain region tends to have a biased sequence composition and tends to be predicted as disordered despite known structures being found in some of these regions. We used the fraction of predicted disordered residues by IUPRED2, as an additional feature [13]. 78% of the positive training data compared to 23% of the negative training data have a minimum of 10% of their protein region predicted as disordered (Figure S2b).

Fibrillar adhesins are attached to the bacterial cell surface. Known anchor domains or sortase motifs were found in 846 out of 1666 proteins of the positive training set. To extend the search for cell surface proteins we predict whether a protein is localized at the cell surface by using a pipeline called Inmembrane [14]. An additional 165 proteins of the positive training set were predicted as surface proteins by Inmembrane. Supplementary figure S3a,b represents the proportion of proteins belonging to the negative or positive training set with a surface anchor or predicted cell wall localization.

FA-like proteins are among the longest proteins in the protomes, thus the protein length turned out to be one of the strongest prediction features (Figure 1). Their high length facilitates FA-like proteins to cross the peptidoglycan layer, which can be around 20-50 nm wide depending on the bacterial species [15]. The average protein length of the positive training set is 1,196 residues compared to 300 residues for the negative training set. Hence, with a higher sequence length of a protein, the probability increases that the protein functions as a FA-like protein (Figure S4).

To characterize the protein sequence of FA-like proteins, the amount of charged as well as hydrophobic amino acids per protein were selected as additional features. The protein sequences of the positive training data tend to have a slightly lower fraction of charged amino acids and tend to have a lower fraction of hydrophobic amino acids compared to the negative training data set (Figure S5a,b).

Finally, we selected features related to the sequence composition. We calculated the fraction of each amino acid per sequence as well as the relative entropy describing the sequence composition bias. We observed that the relative entropy tends to be slightly higher in the positive training set and that threonine is 1.8-fold increased and leucine is 1.5-fold decreased in the positive training set compared to the negative training set (Figure S6, S7).

We implemented the selected features in a random forest classification approach and analysed

the feature importance in the classification prediction based on the training data (Figure 1). The adhesive and stalk domain features can yield a Receiver Operating Characteristics (ROC) Area Under the Curve (AUC) of 0.99 due to the fact that the training set is built upon FA-like proteins identified in our previous work using known adhesive and stalk domains (Figure 1). The low diversity of the training data, considering that the positive training data nearly solely consists of proteins with at least one adhesive and one stalk domain, is reflected in the reliability and precision-recall curve (Figure 2a,b). The reliability curve shows the high number of proteins of the negative training set predicted with a prediction score below 0.1 and the high number of proteins of the positive training set predicted with a score above 0.9. Even though the number of predicted proteins with a score between 0.2 and 0.8 are low, the ratio of false positives tends to increase in the prediction score ranges 0.5 to 0.7. To test if the training set leads to an overfitted model, we evaluated the random forest classification approach with an extreme case of FA-like proteins that we artificially present to have no adhesive or stalk domains (all other features of the proteins are retained), yielding a precision score of 0.83 and achieving a recall score of 0.74 due to missed true positives (Figure 2b). An important challenge for this work is to determine the random forest score threshold that will reliably identify novel FA-like proteins that potentially lack known stalk or adhesive domains.

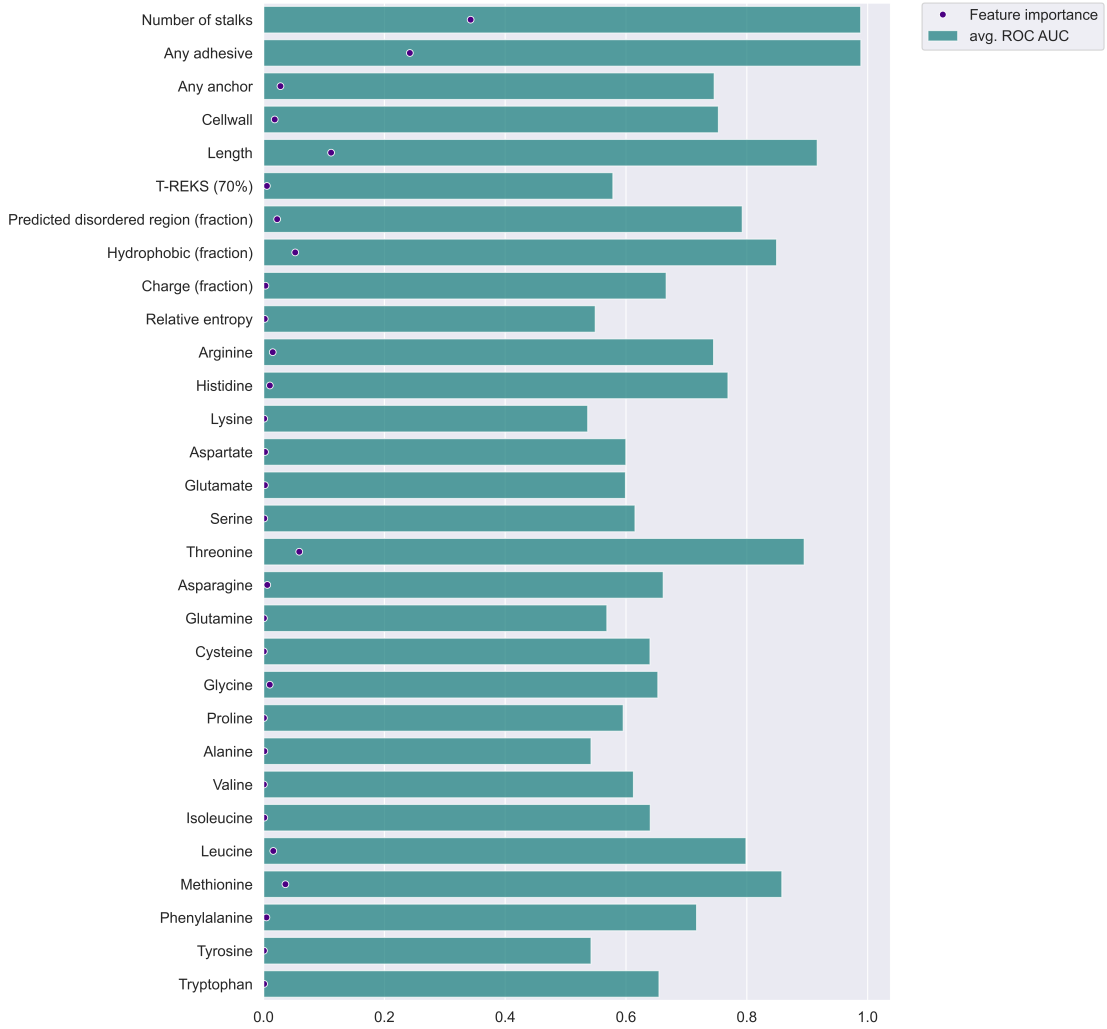

Figure 1: A bar plot showing the relative importance of prediction features: This plot visualizes the importance of each feature for the random forest classification. The bars show the calculated ROC AUC per feature when using it alone for classification. The dot represents the feature importance as calculated by the feature importance attribute implemented in the sklearn ensemble random forest classifier.

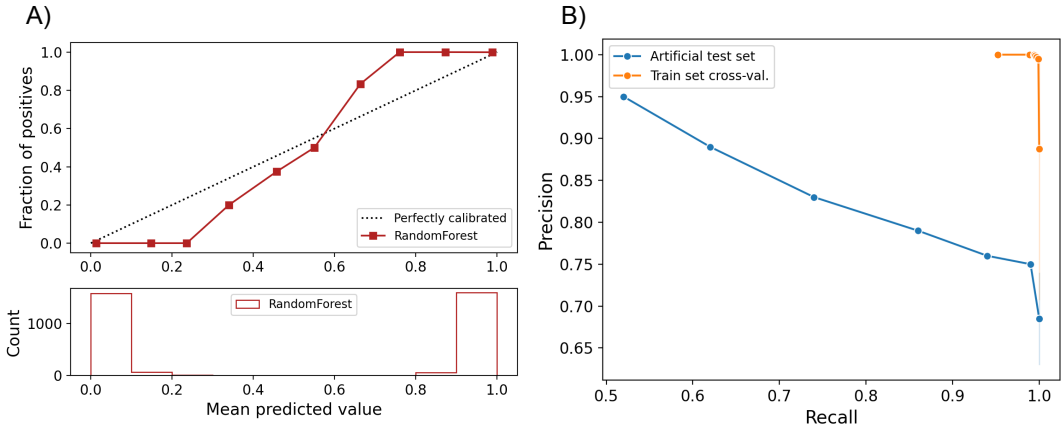

Figure 2: Validation of the trained random forest classifier: A) the subplot on the top represents the reliability curve showing the observed fraction of predicted proteins belonging to the positive training data set against the expected fraction of positives. The subplot below indicated the total number of proteins of the training set predicted per prediction score. B) Precision-Recall curves calculated with the training set using a cross-validation approach (orange) or using a test set with a positive set of FA-like proteins with adhesive and stalk domains artificially removed (blue).

Analysis of predicted FA-like proteins

When applying the classification approach on the Firmicute and Actinobacteria UniProt Reference Proteomes, 33,974 FA-like proteins with a prediction score above 0.5 were identified, 20,491 proteins in Firmicutes and 13,483 proteins in Actinobacteria (Table 1). These represent 0.41% and 0.2% of the total number of reference proteins respectively. Here we provide an analysis to help to determine a reasonable threshold to apply for downstream analysis and application of the classifier in general.

Table 1: Number of FA-like proteins discovered in Firmicutes and Actinobacteria per prediction score bin.

Score	0.5 – 0.6	0.6 – 0.7	0.7 – 0.8	0.8 – 0.9	0.9 – 1.0	1.0
Firmicutes	8,224	6,238	2,877	1,310	690	1,152
Actinobacteria	7,264	3,715	1,513	280	258	453

Studying the predicted FA-like proteins (at a threshold >0.5) showed that over 70% of the predicted proteins have known stalk domains. The stalk number per protein tends to increase with the prediction score with fewer proteins with more than seven stalk domains in proteins with a prediction score below 0.7 (Figure 3a,b).
In known fibrillar adhesins of Gram positive bacteria, the adhesive domain tends to be positioned at the distal end compared to the cell surface anchor position. We calculated the fraction of the protein sequence between the anchor and adhesive domain per predicted protein. The comparison is limited by the different numbers of predicted FA-like proteins with an adhesive domain as well as anchor between the prediction scores, because less than 5% of the predicted

FA-like proteins in total have both a known adhesive domain and anchor. As shown in figure 3d, the mean of the protein fraction tends to increase with the prediction score in Actinobacteria. In Firmicutes the protein fraction is slightly higher in proteins above a prediction score of 0.6 and strongly increased in proteins with a prediction score above 0.9 (Figure 3c). Nevertheless, there are exceptions of proteins with an adhesive domain positioned closer to the anchor. One example is the adhesive protein Adhesin_P1, which expresses two adhesive domains, one N-terminal folding towards the stalk and a second one in the middle, which is presented at the distal end in the final structural conformation [16]. Another example are proteins with repeating adhesive domains, e.g., SSURE (Pfam:PF11966), assumed to function as both a stalk and adhesive binding domain [7].

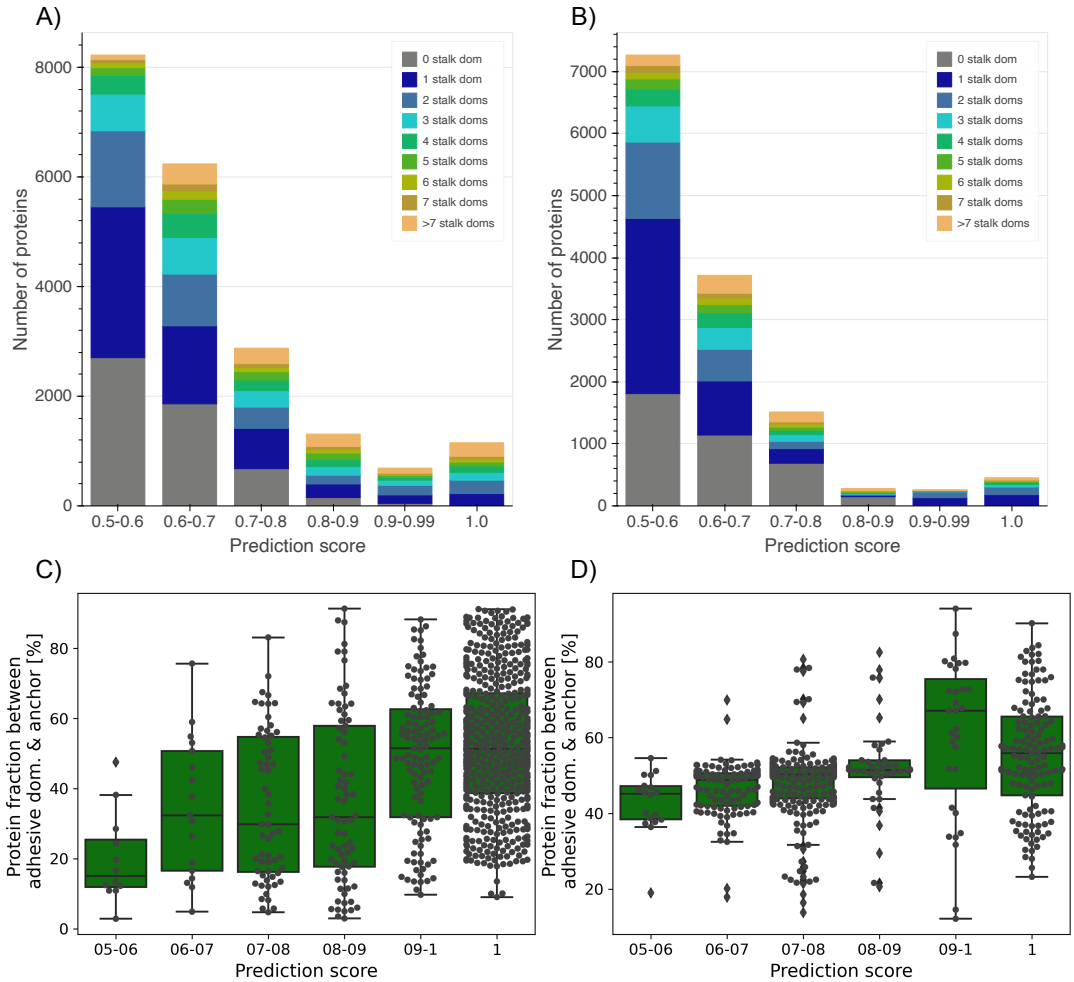

Figure 3: Investigation of the predicted FA-like proteins: This figure shows the results of the analysis of the proteins with the prediction scores between 0.5 and 1.0 for the Firmicutes (A, C) and Actinobacteria (B, D). Subfigures A and B show the amount of proteins per prediction score differentiated between the number of known stalk domains per protein and subfigures C and D visualise the percentage protein fraction between the adhesive domain and the cell surface anchor.

Even though most of the proteins (>90%) predicted with a score above 0.9 with known stalk domains have an adhesive domain, over 90% of the predicted FA-like proteins above a prediction score of 0.5 with known stalk domains have no known adhesive domain (Figure 4a,b). Given that our training set is biased towards FA-like proteins with both stalk and adhesive domains, this result may suggest that there exist a large number of FA-like proteins with as yet undiscovered or unannotated adhesive domains.

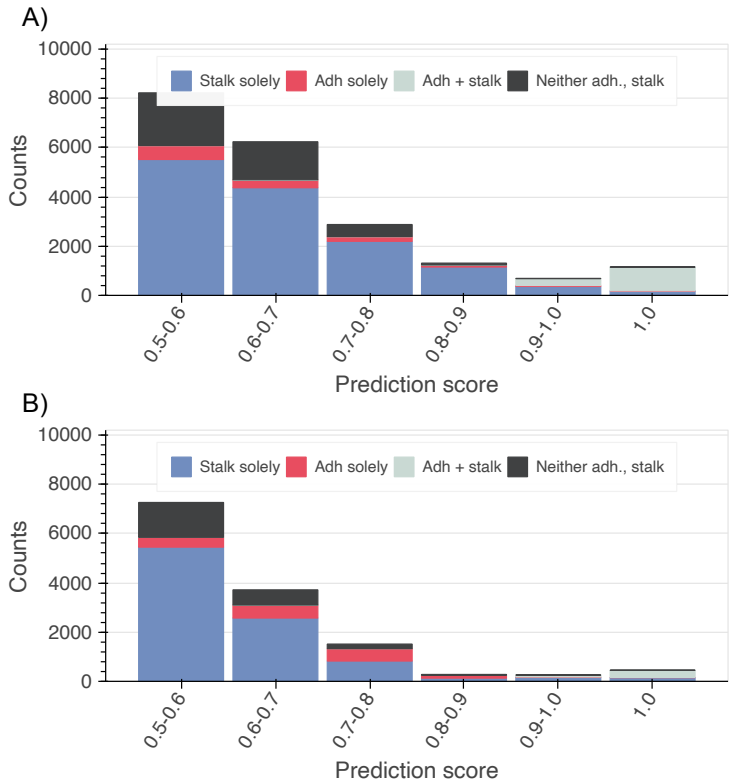

Figure 4: Overview of the presence and absence of known adhesive and stalk domains: The number of proteins per prediction score category is counted, differentiating between proteins with an adhesive and/or stalk domains or neither an adhesive nor a stalk domain. The results are shown for the predicted FA-like proteins in A) Firmicutes and B) Actinobacteria.

We investigated the predicted proteins missing an adhesive domain and found adhesive-domain-like sequences (Table 2). These were searched with the known adhesive domains using HMMER with a higher (less significant) E-value threshold of 1.0. Already over 70% of the proteins with a prediction score of 0.9 or above have a known adhesive domain (Table 2). Adding the number of proteins with distantly related adhesive domains leads to an increase up to 6.8% for a prediction score between 0.8 to 0.9 in Firmicutes (Table 2). These results suggest the presence of novel adhesive domain families distantly related to existing ones detected in the predicted FA-like proteins. The results also indicate the possible existence of potential novel adhesive domains, unrelated to known adhesive domains.

The reliability curve calculated based on the training set suggests a false positive rate of zero percent from a prediction score above 0.8 (Figure 2a). The analysis of the predicted FA-like pro-

Table 2: Distantly related adhesive domains: Percentage of the total predicted FA-like proteins with only known adhesive domains and additional with the distantly related adhesive domains found by using a less significant Evalue.

Score	0.5 – 0.6	0.6 – 0.7	0.7 – 0.8	0.8 – 0.9	0.9 – 1.0	1.0
Firmicutes (only known adh.)	8.99% (739)	6.48% (404)	9.63% (277)	9.08% (119)	72.32% (499)	99.57% (1,148)
Firmicutes (+ distantly related adh.)	13.12% (1079)	11.48% (716)	15.71% (452)	15.88% (208)	75.22 % (519)	99.74% (1150)
Actinobacteria (only known adh.)	8.95% (650)	18.84% (700)	40.98% (620)	56.43% (158)	99.22% (256)	100% (453)
Actinobacteria (+ distantly related adh.)	10.67% (775)	21.08% (783)	44.15% (668)	57.50% (161)	99.22% (256)	100% (453)

teins in Firmicutes and Actinobacteria indicate that predicted FA-like proteins above a predicted score of 0.7 show FA-like protein characteristics concerning the number of stalk domains, the protein architecture with the adhesive domains away from the cell anchor or the percentage of proteins with known or related adhesive domains. Taking the observed results into account, we suggest a high confidence scoring threshold of 0.8, since the predicted FA-like proteins with a scoring threshold below 0.8 might include false positives. Nevertheless, a scoring threshold of 0.7 can be used to find an extended set of FA-like proteins as long as a careful verification of the predicted proteins is carried out.

Detection of novel adhesive domains

To verify the random forest based discovery approach we further investigated the predicted FA-like proteins in the Firmicute Reference Proteomes. These include proteins without known adhesive or stalk domains. Here, we are interested in the proteins with known stalk domains that lack a known adhesive domain. We observed that many of them have a domain annotation gap at the N-terminus, distal to the cell surface anchor. Of these proteins we selected a subset of proteins with a minimum of 4 stalk domains. These were 1,500 proteins in Firmicutes with a prediction score above 0.5. Under the assumption that these annotation gaps might include an adhesive region of the proteins, we clustered the N-terminal sequences into homologous sequence clusters using blastp [17]. We selected the clusters with more than 5 sequences and with an average prediction score of 0.7 or above for further investigation, resulting in 21 clusters. To further investigate the clusters we chose one representative sequence per cluster (Table S1). We carried out two analyses (i) to search for overlapping known Pfam domains, using the highly sensitive iterative jackhmmer search, and (ii) to predict the structure of the sequence using AlphaFold [18, 11]. Using Jackhmmer can be considered a more sensitive version of our previous analysis where we used a less significant HMMER threshold for known Pfam adhesive domains. Using Jackhmmer we are able to detect even more distant similarities to known domains. To find out more about the function, particularly for sequences without overlapping Pfam domains, we searched with the predicted AlphaFold structure models against the PDB database for similar structures [19] (Suppl. Table S2). We aligned the sequences of each cluster and built profile-HMMs specific to each cluster. To understand the relative abundance of each of our clusters we then searched for homologous sequences in UniProtKB and UniProt Reference Proteomes as

well as in the metagenomic MGnify database [20, 21].

Clusters with sequence similarities to known adhesive domains

The jackhmmer search using the putative adhesive region indicated 5 out of the 21 clusters (cluster number 1, 9, 18, 19 and 20) are distantly related to known adhesive domains. These sequence similarities were confirmed by the DALI search with the AlphaFold predicted structure models (Figure 5a-f, S8a-f).

Table 3: Information about clusters with detectable sequence similarities to known Pfam adhesive domains. The distantly related Pfam adhesive domains found with jackhmmer are indicated under 'Domain overlap'. Other information are the homologous sequence hits per cluster in the UniProtKB, UniProt Reference Proteomes and MGnify database as well as the cluster size ('Seqs #') and the average protein prediction score per cluster ('Avg. Score').

Nr.	Avg. Score	Seqs. #	Domain overlap	UniProtKB	UniProt Ref. Proteomes	MGnify
1	0.92 ± 0.04	8	Big_8 (PF17961), Collagen_bind (PF05737)	487*	19*	650*
9	0.8 ± 0.14	8	LRR_4 (PF12799)	45,022	2,793	170,274
18	0.74 ± 0.09	28	TED (PF08341)	1,149	222	48,395
19	0.74 ± 0.04	47	TED (PF08341)	1,170	170	54,096
20	0.73 ± 0.08	6	Collagen_bind (PF05737)	276*	66*	1,174*

* C-terminal domain

Cluster 1 The predicted structure of cluster 1 representative sequence (UniProtKB:K8EVB1) shows that it contains two distinct domains (Figure 5b). For the N-terminal domain Jackhmmer identified a Big_8 domain (Pfam:PF17961), which is found in a variety of bacterial adhesins such as the *Staphylococcus aureus* proteins FnBPA, ClfA and ClfB. A DALI search with the N-terminal domain strengthens the jackhmmer results as the best hit is the N-terminal Big_8 domain of the FnBPA binding region (PDB:4b60:B). For the C-terminal end of the C-terminal domain of cluster 1 the jackhmmer search indicates an overlap to the Collagen_bind domain (Pfam:PF05737). The second domain structure yields with the DALI search the *Streptococcus gordonii* adhesin Sgo0707, where the structure superposes to the Sgo0707_N2 domain (Pfam:PF20623) (PDB:4igb:B). The subsequent DALI hits are the SdrG_C_C adhesive domain (Pfam:PF10425) (PDB:4jdz:A) (Figure S8b). The SdrG_C_C domain is also often found associated with the Big_8 domain in known adhesins and is likely to be homologous to the Big_8 domain. Cluster group 1 includes the *Enterococcus faecalis* protein with the gene name EF2505 or Fss2, which was described to bind to fibrinogen and to play an important role in the adherence

and virulence of *E. faecalis* [22]. Thus we propose that proteins in cluster 1 are likely to bind fibrinogen or other animal extracellular matrix proteins.

Cluster 20 For cluster 20 AlphaFold predicted a structure composed of two domains (Figure 5c). The DALI search indicates, that like cluster 1, for the N-terminal domain a Big_8 domain is the top hit (PDB:5cf3-A) and for the C-terminal domain a Collagen_bind domain is the best match (PDB:2z1p-A) (Figure S8c), thereby confirming the jackhmmer search results. The Big_8 domain functions together with the Collagen_bind domain as supra-domain, enabling the binding to collagen by the collagen hug binding mechanism [23].

Cluster 18 The jackhmmer search as well as the DALI search with the AlphaFold prediction indicate cluster 18 being a class II TED (Pfam:PF08341) adhesive domain (PDB:6fx6:A) (Figure 5d, S8d), which is known to bind to fibrinogen [24, 25]. The TED adhesive domain is categorized into a class I and class II TED domain, depending on an additional N-terminal indel forming an alpha helix or an additional C-terminal indel folding into a beta-sandwich, respectively [24]. Cluster group 18 includes the fibrinogen binding *E. faecalis* Fss3 protein (UniProtKB:Q833P7) as well as an *E. faecalis* protein (UniProtKB:Q831Z5) encoded by the virulence associated EF2347 gene [22].

Cluster 19 The jackhmmer search indicates cluster 19 to also be distantly related to the TED adhesive domain [25]. The DALI results with the best hit is a class II TED domain (PDB:6fwv) (Figure 5e, S8e) strengthens this hypothesis. To confirm that cluster 19 is a class II TED domain, we extended the representative sequence above the 400 residues limit by around 100 residues in order to include the characteristic C-terminal indel.

Cluster 9 Cluster group 9 includes the *Listeria monocytogenes* InternalinJ (UniProtKB:Q8Y3L4), for which the crystal structure of the adhesive domain exists (PDB:3bz5:A) [26], which was found using DALI with the predicted AlphaFold structure, reflecting the high accuracy of AlphaFold (Figure 5f, S8f). The adhesive domain is formed by a series of Leucine Rich Repeats that are not matched by the LRR_4 family (Pfam:PF12799) in Pfam. Proteins related to this class are the most prevalent that we found with over 175,000 homologues identified in the MGnify protein sequence database.

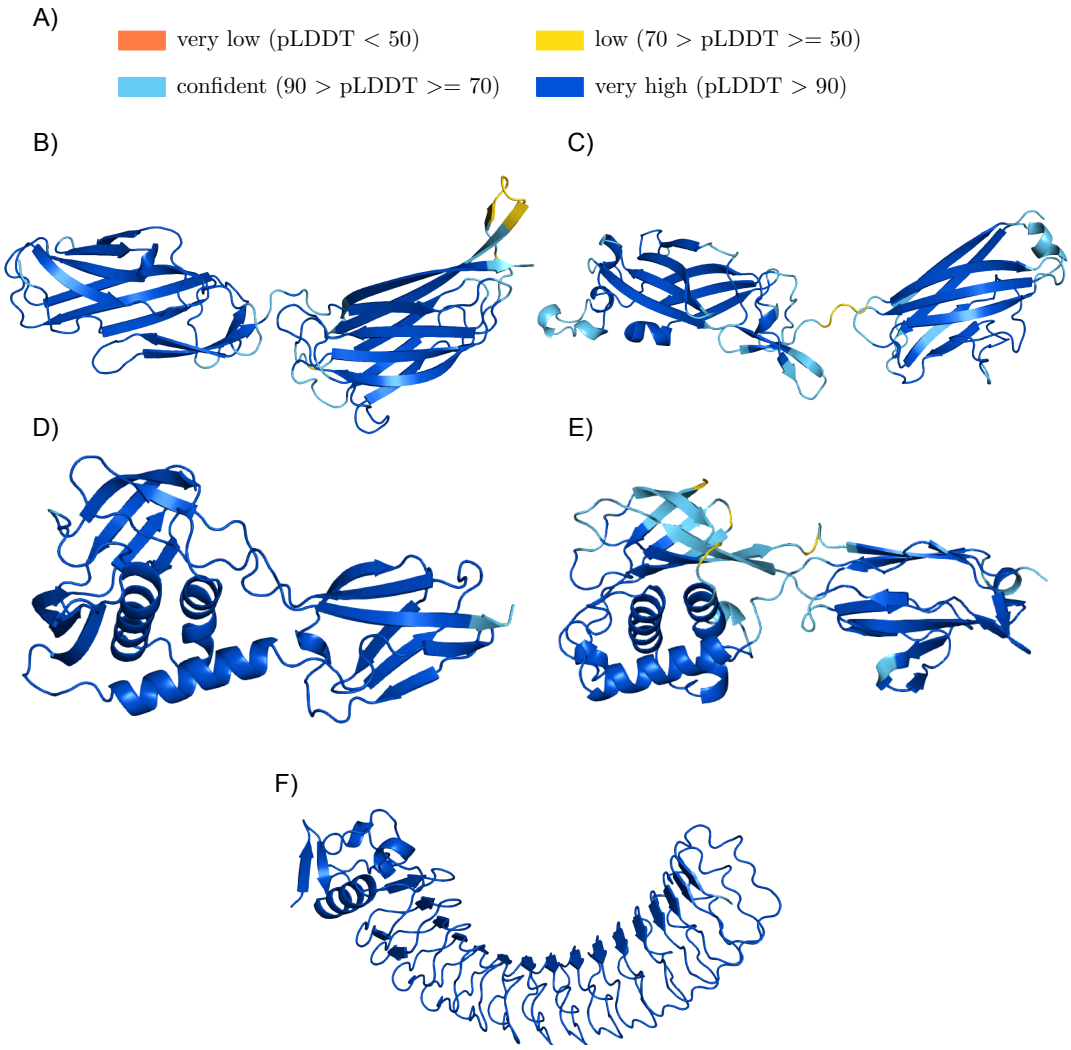

Figure 5: Structure models of cluster groups with distantly related adhesive domains found by jackhmmer: (A) Colour legend representing the quality of the AlphaFold models. Structure models of the potential adhesive domain of (B) cluster 1 (UniProt: K8EVB1/117-417), (C) cluster 20 (UniProt: V2XMF4/80-366), (D) cluster 18 (UniProt: R3TX93/40-276), (E) cluster 19 (UniProt: A0A2Z4U801/240-508) and (F) cluster 9 (UniProt: A0A4U7JL97/37-420). The figures were produced using Pymol [42].

Clusters with structure models indicating role in adhesion function

The cluster groups 2, 6, 7, 8, 10, 11, 15, 17 and 21 don't yield persuasive jackhmmer matches, but their predicted structures resemble structures of known adhesive domains or adhesion associated domains, suggesting that these clusters might be novel domains with potential binding functions.

Table 4: Investigation of the potential novel adhesive domains: This table lists information regarding overlapping known Pfam domain families found with jackhmmer (Domain overlap) and the abundancy in the UniProtKB, UniProt Reference Proteomes and MGnify database for the N-terminal sequence clusters. The 'Seqs #' column represents the cluster size and the 'Avg. Score' column the average prediction score of the proteins per cluster.

Nr.	Avg. Score	Seqs. #	Domain overlap	UniProtKB	UniProt Ref. Proteomes	MGnify
2	0.9 ± 0.12	7		256	38	569
6	0.81 ± 0.08	11	MBG (PF17883) (C-terminal)	53	24	125
7	0.81 ± 0.07	9		502	122	1,124
8	0.81 ± 0.05	5		89	17	89
10	0.8 ± 0.08	8		145	32	157
11	0.79 ± 0.09	5		103	16	87
15	0.75 ± 0.09	10		2,359	551	6,668
17	0.74 ± 0.12	8		444	47	1,180
21	0.71 ± 0.07	6	Cthe_2159 (PF14262)	495	65	1,495

Clusters with jelly-roll resembling structure predictions The structure predictions for cluster 6, 8, 10 and 11 are jelly-roll like structures (Figure 6a-e). The jackhmmer search for cluster group 6 showed an overlap with the MBG stalk domain at the C-terminus of the sequence, which was trimmed off, leaving a significant N-terminal sequence region for which the structure was predicted. The top DALI matches are related to surface adhesins with the best match being a *Clostridium perfringens* pilin protein (PDB:5xcb:A) with the GramPos_pilinBB domain (Pfam:PF16569) (Figure S9b).

The majority of the DALI hits for cluster 8 are related to bacterial adhesins, with the best hit being the Big_8 domain of the functional region of the fibronectin binding FnBPA protein (PDB:4b5z:A) (Figure 6c, S9c). A disordered region is found C-terminal to the cluster 8 domain (UniProtKB:J9W466). The disordered region presumably acts as a flexible linker between the adhesive domain and the stalk leading to an extended radius the adhesive domain can find the binding target as well as to a higher flexibility regarding the orientation of the adhesive domain to its binding target (Suppl. figure S12).

For cluster group 10 the best DALI hit was the GramPos_pilinBB domain of the RrgB pilus backbone (PDB:2x9x:A). (Figure 6d, S9d).

The top DALI hit indicated for cluster group 11 a similarity to an human integrin alpha-V (PDB:

6avu:A), followed by an A-macroglobulin receptor binding domain in cattle (PDB:1ayo:B, Pfam: PF07677) (Figure 6e, S9e), suggesting a protein binding function. Similar to cluster 8, the cluster 11 potential adhesive domain is followed by a disordered region (UniProtKB:X0SH12) potentially functioning as a flexible linker between the adhesive domain and the stalk in order to increase the binding ability (Suppl. Figure S12).

The Evolutionary Classification Of protein Domains (ECOD) databases classifies the Gram-Pos_pilinBB domain, Big_8 domain (ECOD:EUF07753) and A-macroglobulin receptor binding domain together under the same topology named 'Common fold of diphtheria toxin/transcription factors/cytochrome f' [27]. This category also includes the adhesive domains SdrG_C_C or Collagen_bind. The Collagen_bind adhesive domain is a well studied jelly-roll structure [28], which is composed of two antiparallel beta-sheets and two short alpha-helices. All of the AlphaFold structures for cluster 6, 8, 10 and 11 fold into jelly-roll like structures and show a high similarity to the Collagen_bind domain (PDB:1amx-A). The structure surfaces of cluster 6, 8 and 10 seem to provide a groove on the beta-sheets, indicating a potential collagen binding site [28]. However, the Collagen_bind domain alone has a 10-fold lower collagen binding affinity compared to the collagen hug binding mechanisms based on the Big_8:Collagen_bind supra-domain [23]. The similarity to the Collagen_bind structure, but also the N-terminal protein position distal to the cell surface anchor strongly suggests adhesive function for clusters 6, 8, 10 and 11.

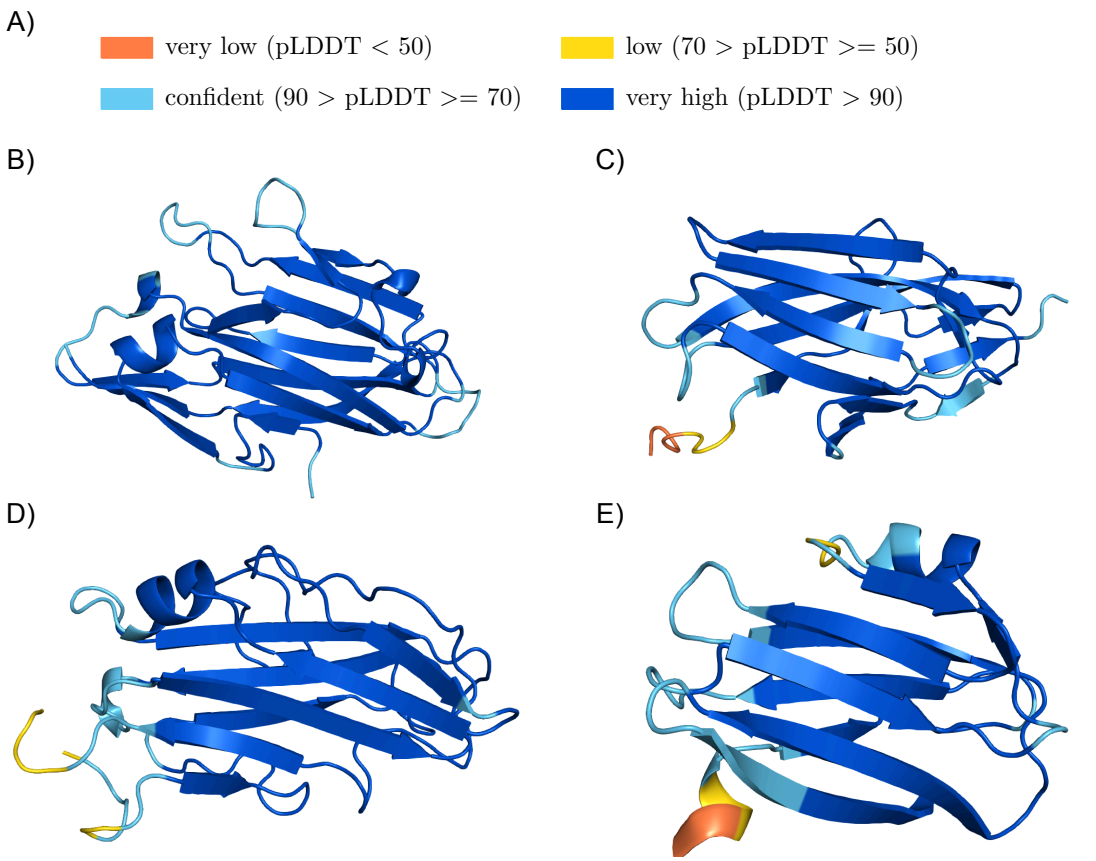

Figure 6: Potential novel adhesive domains with jelly-roll like predicted structures: (A) Colour legend representing the quality of the AlphaFold models. Structure models of the potential adhesive domain of (B) cluster 6 (UniProt: A0A5R8Q9T8/68-240), (C) cluster 8 (UniProt: J9W466/30-163), (D) cluster 10 (UniProt: B0S3M8/153-316) and (E) cluster 11 (UniProt: X0SH12/37-142). The figures were produced using Pymol [42].

Beta-solenoid fold cluster structure models Clusters 2, 15 and 21 are predicted to fold into beta-solenoid structures (Figure 7a-d). The predicted structures for cluster group 2 as well as 21 are most similar to the binding region of the serine-rich repeat protein (SRRP) from *Lactobacillus reuteri* strain 100-23C (PDB:5ny0:A) (Figure S10b,c), being described to bind to epithelial cells, pectic acids and to play a role in the biofilm formation [29]. The jackhmmer search for cluster 21 already indicated after the second iteration a distant relation to the carbohydrate binding Cthe_2159 (Pfam:PF14262) domain which is part of the Pectate Lyase superfamily, whereas the DALI search clearly indicated the highest similarity to the *L. reuteri* SRRP adhesive region [29]. The SRRP protein is not part of any existing Pfam family. Although we limited the clustering sequence to 400 residues we investigated whether the domains were longer with AlphaFold and extended it in the case of cluster 2 to about 800 residues. Interestingly, cluster 2 is found on a *Staphylococcus epidermidis* protein with a SasG_G5-E stalk (UniProtKB:A0A3G1RMM4), which was so far only found associated with the Bact_lectin adhesive domain (Pfam:PF18483) in *S. epidermidis* and *S. aureus* SasG homologues [30]. In our previous study we discussed

the possibility that an adhesive domain can function with any arbitrary stalk [7]. The described example underlines this hypothesis and furthermore indicates a possible transfer of the adhesive domain onto a given stalk increasing the adhesins variability. The SasG_G5-E stalk is also described to promote biofilm formation [30].
Cluster group 15 resembles an ice binding domain (PDB:4nuh:A) (Figure S10d), where the representative protein sequence (UniProt:A0A0F7RLJ7) is identical to a *Bacillus anthracis* protein (UniProt:A0A384LNE7) with the gene name BA_0871 or BASH2_04951, which was described to be collagen binding and to be linked to the bacterial pathogenicity [31].

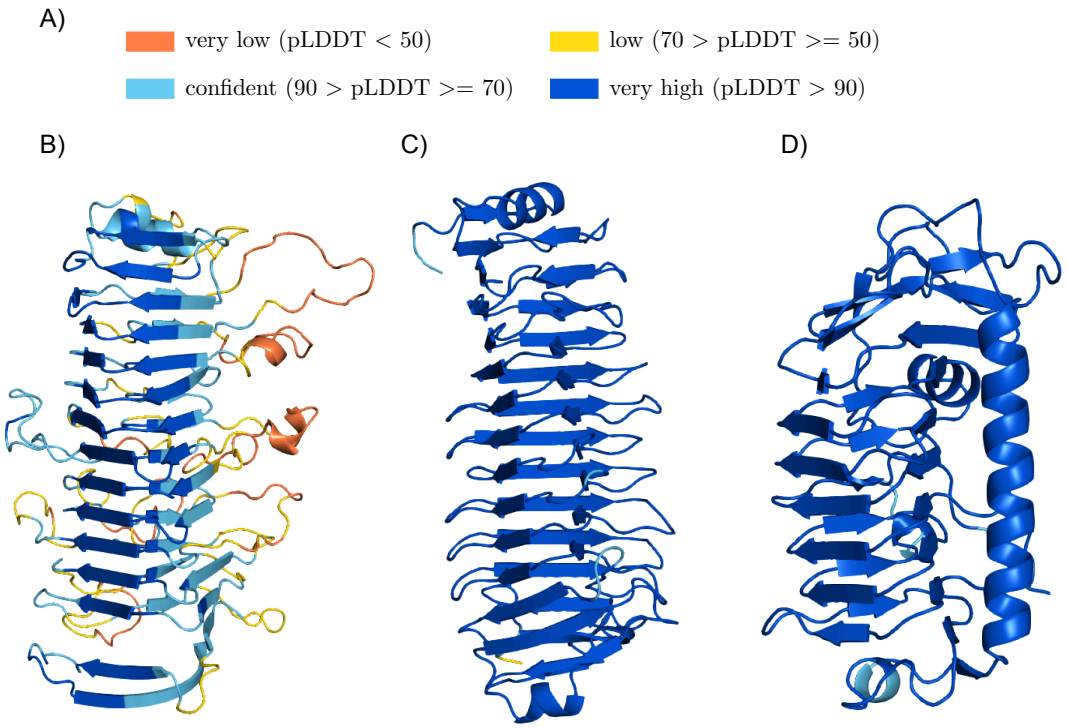

Figure 7: Predicted structures for clusters with potential novel adhesive domains, whose structure, but not sequence, seem to be related to known adhesive domains: (A) Colour legend representing the quality of the AlphaFold models. Structure models of the potential adhesive domain of (B) cluster 2 (UniProt: A0A2Z6T9E9/185-718), (C) cluster 21 (UniProt: A0A099WCN8/48-417) and (D) cluster 15 (UniProt: A0A0F7RLJ7/46-326). The figures were produced using Pymol [42].

Remaining clusters with structure models indicating potential adhesion function The best DALI hit for the cluster 7 structure model (Figure 8b) is a N-terminal helical domain of a group B streptococcus immunogenic bacterial adhesin named BibA (PDB:6poo:A) (Figure S11b), which superposes with the N-terminal alpha helices of the structure model [32]. Running the DALI search for the domain in the middle of the structure model separately, results in the *S. aureus* SdrD adhesive protein (PDB:4jdz:A), where the structure model superposes with the SdrD_B stalk domain (Pfam:PF17210). The cluster 7 includes one *Streptococcus merionis* protein (UniProtKB:A0A239SMH4), which is encoded by the bca gene. The bca gene has been

shown to be involved in the initial stage of Group B Streptococcus infection [33], suggesting adhesion function.

The best DALI match for cluster 17 is the human integrin alpha-5 protein (PDB:7nxd:A), followed by the best bacterial match being the N-termini of the *S. gordonii* adhesins Sgo0707 (PDB:4igb:B) (Figure 8b, S11c). Here, the structure model aligns to the Sgo0707_N2 domain. The cluster includes a *E. faecalis* protein (UniProt: Q82YW8) encoded by the EF3314 gene, which was described to contribute to the virulence properties of this pathogen [34].

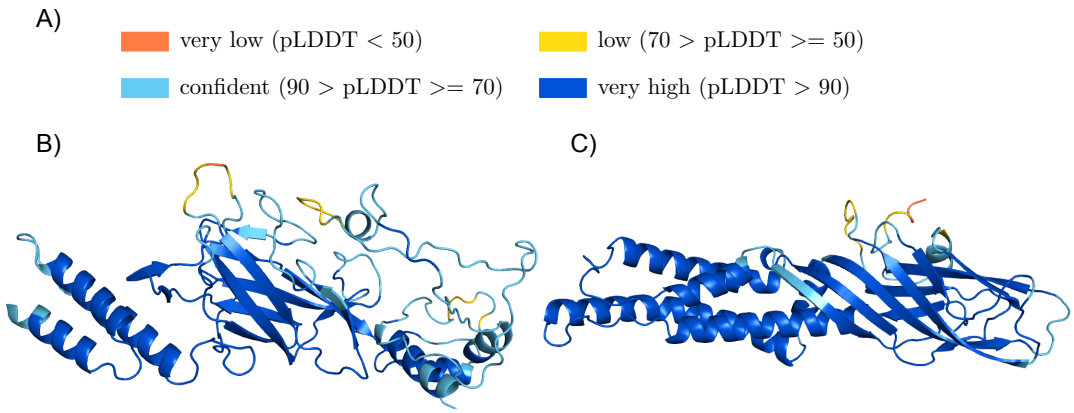

Figure 8: Clusters with AlphaFold structures showing ambiguous adhesion function: (A) Colour legend representing the quality of the AlphaFold models. Structure models for (B) cluster 7 (UniProt: A0A1Q8E8C7/76-420) and (C) cluster 17 (UniProt: A0A069CUH0/64-357). The figures were produced using Pymol [42].

We created new putative adhesive pfam domain families for the cluster 2, 6, 8, 10, 11, 15, 17 and 21. Clusters 2 and 21 were combined into a single cluster. The Pfam identifiers can be found in supplementary table S2. The common domain architectures of proteins with these potential novel adhesive domains are shown in supplementary figure S12.

Clusters unlikely to be adhesive domains

The jackhmmer and DALI search indicated cluster group 3, 4, 5, 12, 13, 14 and 16 to be unlikely to function as binding domains (Suppl. figure S13) and all, except cluster 5 and 12, to be similar to structures with catalytic functions. Indeed, it is challenging to differentiate between fibrillar adhesins and as we call them 'fibrillar enzymes'. Fibrillar enzymes are also composed out of repeating domains, but have an enzymatic domain instead of an adhesive domain. Nevertheless, the enzymatic region can be able to have binding functions, for example, cluster group 13 is similar to the *Streptomyces* pectate lyase (PDB:4pex:B), which is suggested it may have sugar binding function [35].

The structure predictions for cluster 5 and 12 yielded DALI matches without catalytic functions. The best DALI hit for the cluster 12 structure model is a human integrin (PDB:6v4p:A). For cluster 5, AlphaFold predicted a structure composed of two domains. The best DALI hits for both domains are all related to bacterial surface adhesins with the top DALI match being each time the BcpA pilus subunit of *Bacillus cereus* (PDB:3rkp:A or 3kpt:B) (Figure S13d) [36]. The model superposes to the XNA domain of the major subunit of BcpA, whose isopeptide bonds within the

XNA domain are critical for bundle and pili formation [36]. The jackhmmer search yielded the DUF11 stalk domain. This is strengthened by the predicted structures, which strongly resemble stalk domains, nevertheless leaving open whether these stalk domains can be involved in the binding function [7]. One possibility is that the function of these proteins is not adhesive, but to act as a steric regulator altering the access of other adhesive proteins to binding partners. One example is the *S. aureus* periscope protein SasG, which is suggested to block the binding of proteins located closer to the cell surface from interacting with host cell fibrinogen [9].

Table 5: Investigation of clusters with unlikely adhesion function: Information regarding overlapping known Pfam domain families found with jackhmmer (Domain overlap) and the abundancy in the UniProtKB, UniProt Reference Proteomes and MGnify database are listed for the N-terminal sequence clusters. The 'Seqs #' column represents the cluster size and the 'Avg. Score' column the average prediction score of the proteins per cluster.

Nr.	Avg. Score	Seqs. #	Domain overlap	UniProtKB	UniProt Ref. Proteomes	MGnify
3	0.83 ± 0.08	5	FlgD_ig (PF13860)	485	145	1,310
4	0.82 ± 0.09	5	Laminin_G_3 (PF13385)	1,935	422	22,544
5	0.82 ± 0.07	5	DUF11 (PF01345)	554*	98*	2,833*
12	0.77 ± 0.11	11	BNR_4 (PF15892)	7,916	1,148	37,575
13	0.77 ± 0.07	5	fn3 (PF00041), Pectate_lyase_3 (PF12708)	3,469	950	6,897
14	0.76 ± 0.08	7	PSII_BNR (PF14870)	15,047	2,460	86,845
16	0.74 ± 0.13	8	Beta_helix (PF13229)	6,056	1,463	66,262

* N-terminal domain

Discussion

Novel pathogens are emerging constantly with uncharacterized host cell interaction mechanisms. Homologous virulence associated proteins with known adhesive domains are the first step towards an understanding of the pathogenicity of these bacteria. But adhesive domains evolve quickly and are highly variable. Hence, detection mechanisms independent from known adhesive domains are important. In this study we developed a random forest based discovery approach to detect FA-like proteins. We applied the approach on the Firmicute and Actinobacteria UniProt Reference Proteomes, yielding over 4,000 confidentially predicted FA-like proteins. With the characterization of FA-like proteins we could identify a variety of notable features which

we could use for machine learning. The known stalk and adhesive domains are the strongest feature in the classification decision approach. This bias is due to the positive training data, which was selected from the prior domain-based discovery approach [7]. Other strong features were the protein length, which is required to overcome the bacterial cell surface, and the amino acid composition of the protein sequences. Here, particularly threonine was strongly over-represented in the positive training data set compared to the negative training data set, raising the question about what role it plays in bacteria-host interactions. One explanation could be that despite the most commonly phosphorylated amino acids being histidine and aspartate, serine/threonine/tyrosine phosphorylation in bacteria was shown [37, 38]. Additionally, phosphorylation during the adherence step was described, connecting it to the regulation of the bacterial virulence [38, 39]. Implementing the selected identification properties in the random forest classification approach and applying it to the Firmicute and Actinobacteria UniProt Reference Proteomes led to over 4,000 confidentially detected FA-like proteins. This indicates that more than 2,400 of them were missed by the domain based discovery approach detected in our previous study [7]. More importantly, with our new machine learning discovery approach FA-like proteins are predicted that lack known adhesive or stalk domains enabling us to discover novel protein domain families.

To verify the random forest prediction approach we further studied the predicted FA-like proteins lacking a known adhesive domain, but with known stalk domains. When investigating the sequence clusters representing an annotation gap N-terminal to known stalk domains, related Pfam domain sequence matches could be found for many of the described clusters using jackhmmer. This suggests that the Pfam domain families can be expanded to include these sequences or novel domain families related to the overlapping domain families can be created. We have taken advantage of the recent release of the AlphaFold2 software to validate our machine learning approach as well as use it to refine predictions of adhesive domains in our predicted fibrillar adhesins. Given that the predicted structures confirm the jackhmmer results, highlights the high accuracy of the structure prediction method AlphaFold2. We see many new opportunities to use large scale structure predictions to identify and investigate the components of the bacterial cell surface that are likely to interact with the host.

While further investigating the described N-terminal sequence clusters the difficulty to differentiate between fibrillar adhesins and the newly discovered class of fibrillar enzymes was shown. Given that fibrillar enzymes can play an important role in the bacterial pathogenesis as well, they have comparable characteristics as fibrillar adhesins in terms of being long surface proteins with a stalk and that several of them can have binding functions. This impedes the differentiation of these two protein classes by our identification features and so far we have no property to differentiate between adhesive and enzymatic domains. Nevertheless, the prediction score and the cluster size can together give an assessment about the reliability. The analysed clusters with the higher sequence number or higher prediction score are mostly potential adhesive domains. The potential novel adhesive domains as well as the clusters related to known adhesive domains found with jackhmmer verify the random forest based discovery approach. The high number of homologous sequences of these domains in the metagenomic MGnify database and in known pathogenic genera in the UniProt database underline their relevance. For the potential novel adhesive domains discovered in the course of this study the predicted structure models and the DALI search results give a first understanding of their function and potential binding partners. AlphaFold2 and the recent announcement of AlphaFold-Multimer opens up further ways to predict the structures of fibrillar adhesins-target protein complexes [40]. We believe that we are at the beginning of a new age of discovery where computational analyses will lead to fundamental improvements in our understanding of microbial host interactions.

Materials and Methods

Training data selection

We selected as positive training data the FA-like proteins of Actinobacteria and Firmicutes discovered with the domain-based detection approach in our previous study [7]. Additionally, we included 25 additional FA-like proteins, which don't have either a known adhesive and stalk domain. As negative training data we selected randomly non FA-like proteins in reference proteomes, in which FA-like proteins could be detected with the domain-based discovery approach. These are from the following nine organisms: *Bifidiobacterium subtile*, *Olsenella sp. oral*, *Slackia exigua*, *Streptomyces coelicolor*, *Staphylococcus aureus*, *Lactococcus lactis*, *Streptococcus gordonii*, *Listeria monocytogenes* and *Enterococcus faecalis*. The training set consists of a total of 3,332 proteins, of which half belong to the positive and the other half to the negative training data set. The training data can be found in the GitHub repository (see below).

Identification features calculation and random forests classification

To search in the protein sequences for known adhesive, stalk and anchor domains, the collection of Pfam domain HMMs from our previous study was used [7]. Additional, the adhesive domain GspA_SrpA_N (Pfam:PF20164) and the stalk domains aRib (Pfam:PF18938), RibLong (Pfam:PF18957), SasG_E (Pfam:PF17041), GA-like (Pfam:PF17573), YDG (Pfam:PF18657), Lipoprotein_17 (Pfam:PF04200) and IgG_binding_B (Pfam:PF01378) were used. These HMMs were run against the protein sequences using the HMMER tool (version 3.1b2) with the gathering (GA) threshold option. Using regular expression, we searched within the C-terminal 50 residues of the protein sequences for the following sortase anchor motifs: 'LPxTG', 'LPxTA', 'LPxTN', 'LPxTD', 'LPxGA', 'LAxTG', 'IPxTG', 'NPxTG', 'NPQTM' ('x' can be any amino acid).
To identify highly similar tandem sequence repeats that may represent potential unknown stalk domains we applied the T-REKS software on the sequences using as parameters a minimum of 70% sequence identity, 50 residues as minimum length of the repeat region and 5 residues as minimum seed length [12].
Disordered regions were predicted using IUPred (IUPred2a) with the IUPred2 type 'long' for predicting long disordered regions [13]. Each residue with an IUPred score above 0.5 was counted as predicted disordered. The predicted disordered fraction was calculated using the percentage of predicted disordered residues from the total protein length.
The cellular localisation was predicted using Inmembrane (version 0.95.0) with the default settings [14]. The proteins predicted by Inmembrane as potential surface exposed (PSE) proteins at the cellwall ('PSE-cellwall') were selected for the identification feature 'Cellwall'.
The proportion of charged or hydrophobic amino acids per protein sequence was calculated using as charged amino acids glutamic acid (E), aspartic acid (D), lysine (K), arginine (R) and as hydrophobic amino acids alanine (A), isoleucine (I), leucine (L), methionine (M), phenylalanine (F), tryptophan (W), tyrosine (Y), valine (V).
The residue length was counted per the complete UniProt protein sequence.
The proportion for each amino acid per protein sequence was calculated and to evaluate the amino acid composition bias the relative entropy (Kullback-Leibler (KL) divergence) was calculated per protein sequence (S). Here, we quantify the difference between the observed frequency (P) per amino acid (i) compared to equally frequent amino acids, being 0.05 for 20 amino acids.

$$\mathrm{KL(S)} = \sum_{i}^{20} \mathrm{P(i)} \log \frac{P(i)}{0.05}$$

We calculated the identification features for the protein sequences of the training data. With the calculated feature data we trained a random forest classifier from sklearn ensemble methods with 50 trees with maximum 3 features per tree and random state 2 [41]. The Random Forest method takes the 30 features as input and outputs a score per protein between 0 and 1 with FA-like proteins scoring closer to 1.
The reliability curve was calculated for the applied random forest model on the training data set using calibration_curve from the sklearn calibration module and a 10-fold cross-validation approach.
For calculating the precision and recall of the model and generating the precision-recall curve, we generated a testing data set of 258 proteins, composed of 128 FA-like proteins and 130 non FA-like proteins. We artificially adapt the features of the testing set to have no adhesive or stalk domains, whereby all other features are retained. For these calculations the proteins of the testing set were excluded from the training data set. The precision and recall of the model as well as the precision-recall curve was calculated using the macro-average method to determine how the random forest model performs overall across the two classes: FA-like, non FA-like proteins. The precision-recall curve was also calculated using a cross-validation approach with the training set.
To use the random forest discovery approach, we provide the code in our GitHub repository (see below).

FA-like proteins prediction for Firmicute and Actinobacteria UniProt Reference Proteomes

To apply our machine learning method against known Firmicute and Actinobacterial proteins we first gathered available sequences. The UniProt proteome identifier for all Firmicute and Actinobacteria Reference Proteomes were searched for in the UniProt website (release 2020_04). We collected the relevant sequences for these identifiers by searching in the knowledgebase under the bacterial reference proteomes (release 2020_03) for the identifier.
As described in the subsection 'Identification features calculation and random forest classification' we calculated the identification features for the Firmicute and Actinobacteria reference protein sequences and applied the trained random forest classification approach to score each protein.

Analysing predicted FA-like proteins

We further analysed the predicted FA-like proteins by differentiating the prediction scores. The number of stalk domains per protein were counted regardless whether these are stalk domains of the same or different domain families.
The fraction of the protein's length between the adhesive domain and the anchor were determined using the middle point per domain or motif.
Using profile HMM-search (version 3.1b2) with the gathering (GA) threshold option the Pfam database (version 33.1) was run against the sequences of the predicted FA-like proteins.

Selecting potential functional sequences

To verify the machine learning approach, we were particularly interested in the predicted proteins with an annotation gap at the N-terminus, which might contain a missing functional domain. We focused on the N-terminus, because we showed in our previous study that the adhesive domain

in FA-like proteins in Firmicutes and Actinobacteria is mostly found at the N-terminus [7]. We selected proteins with at least four known stalk domains, which lack a known adhesive domain and with no Pfam domain annotations within the first 20% of the protein length. Before finding homologous sequence groups, we deleted the selected protein's first 20 residues to avoid clustering based on a potential signal peptide. We cut these sequences N-terminal to the first domain annotation, but no longer than 400 residues in order to try to avoid clustering based on potential stalk domains. We clustered those excised sequences into homologous sequence cluster groups using blastp all against all with an E-value threshold of 0.001, requiring a coverage threshold of 85% and an identity threshold of 25% [17]. For each cluster we calculated the reliability by averaging the random forest prediction scores of the proteins per cluster. We sorted the resulting clusters by average prediction score as well as sequences per cluster. We further investigated the 21 cluster groups with at least 5 homologous sequences.

To investigate the potential function of these sequence clusters, we chose one representative protein per cluster. To do so we aligned the N-terminal sequences per cluster and manually selected one representative sequence per cluster, which was used for the following investigations. For each representative sequence we searched the whole UniProtKB with jackhmmer using the HMMER website (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer) to find domain families related to the sequence clusters [18]. For cluster 6 we found a distant related stalk domain overlapping with the C-terminus of the representatives sequences, we trimmed off the sequence with the domain annotation and continued with the N-terminal sequence.

The structure for each representative sequence was predicted with AlphaFold2 using the Google colab repository provided by DeepMind (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb) [11]. Based on the predicted structure we selected the domain boundaries and cut the structure as well as sequence of each cluster accordingly (Suppl. Table S1). In most cases, we cut off disordered regions. In single cases, for cluster 10 and 11, we optimized the structure by cutting off a stalk domain-like C-terminal to focus on the potential adhesive domain and rerun the AlphaFold structure prediction. For cluster 2 and 19, we extended the sequence to include the whole domain.

To assess the quality of the models, AlphaFold stores the pLDDT confidence in the B-factor field of the output PDB files, which were used to colour the structure models by quality using Pymol [42]. To find out more about the function of the clusters, we searched with the predicted structure models, optimizised to the domain boundaries, for similar structures in the PDB database using DALI. [43].

We created a HMM from the sequences per cluster based on the detected domains using hmmbuild [18]. With these HMMs we searched against the metagenomic MGnify (release 2019_05), UniProt Reference Proteomes and UniProtKB (release 2021_01) databases for homologous sequences using an domain E-value threshold of 0.01 [20, 21]. From the UniProt website the Retrieve ID/Mapping tool was used to collect the organisms information to the UniProtKB matches.

Data availability

We provide a GitHub repository (https://github.com/VivianMonzon/FAL_prediction), which includes the AlphaFold structure models and the code to run the RandomForest based FA-like protein prediction on a sequence of interest.

Acknowledgment

We thank Aleix Lafita for his advice and his great knowledge on bacterial stalk domains.

Competing interests

The authors of this work are supported by the core EMBL funding and declare that they have no competing interests.

References

[1] Vaca DJ, Thibau A, Schütz M, Kraiczy P, Happonen L, Malmström J, et al. Interaction with the host: the role of fibronectin and extracellular matrix proteins in the adhesion of Gram-negative bacteria. Medical Microbiology and Immunology. 2020 Jun;209(3):277–299. Available from: http://link.springer.com/10.1007/s00430-019-00644-3.

[2] Krachler AM, Orth K. Targeting the bacteria–host interface: Strategies in anti-adhesion therapy. Virulence. 2013 May;4(4):284–294. Available from: http://www.tandfonline.com/doi/abs/10.4161/viru.24606.

[3] Foster TJ, Geoghegan JA, Ganesh VK, Höök M. Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus. Nature Reviews Microbiology. 2014 Jan;12(1):49–62. Available from: http://www.nature.com/articles/nrmicro3161.

[4] Geoghegan JA, Corrigan RM, Gruszka DT, Speziale P, O'Gara JP, Potts JR, et al. Role of Surface Protein SasG in Biofilm Formation by *Staphylococcus aureus*. Journal of Bacteriology. 2010 Nov;192(21):5663–5673. Available from: https://journals.asm.org/doi/10.1128/JB.00628-10.

[5] Valle J, Latasa C, Gil C, Toledo-Arana A, Solano C, Penadés JR, et al. Bap, a Biofilm Matrix Protein of Staphylococcus aureus Prevents Cellular Internalization through Binding to GP96 Host Receptor. PLoS Pathogens. 2012 Aug;8(8):e1002843. Available from: https://dx.plos.org/10.1371/journal.ppat.1002843.

[6] Back CR, Higman VA, Le Vay K, Patel VV, Parnell AE, Frankel D, et al. The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture. Journal of Biological Chemistry. 2020 May;295(19):6689–6699. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925817485267.

[7] Monzon V, Lafita A, Bateman A. Discovery of fibrillar adhesins across bacterial species. BMC Genomics. 2021 Dec;22(1):550. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07586-2.

[8] Rich RL, Demeler B, Ashby K, Deivanayagam CCS, Petrich JW, Patti JM, et al. Domain Structure of the *Staphylococcus aureus* Collagen Adhesin. Biochemistry. 1998 Nov;37(44):15423–15433. Available from: https://pubs.acs.org/doi/10.1021/bi981773r.

[9] Whelan F, Lafita A, Gilburt J, Dégut C, Griffiths SC, Jenkins HT, et al. Periscope Proteins are variable-length regulators of bacterial cell surface interactions. Proceedings of the National Academy of Sciences. 2021 Jun;118(23):e2101349118. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.2101349118.

[10] Vance TDR, Bayer□Giraldi M, Davies PL, Mangiagalli M. Ice□binding proteins and the ‘domain of unknown function’ 3494 family. The FEBS Journal. 2019 Mar;286(5):855–873. Available from: https://onlinelibrary.wiley.com/doi/10.1111/febs.14764.

[11] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–589. Available from: https://www.nature.com/articles/s41586-021-03819-2.

[12] Jorda J, Kajava AV. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics. 2009 Oct;25(20):2632–2638. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp482.

[13] Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research. 2018 Jul;46(W1):W329–W337. Available from: https://academic.oup.com/nar/article/46/W1/W329/5026265.

[14] Perry AJ, Ho BK. Inmembrane, a bioinformatic workflow for annotation of bacterial cell-surface proteomes. Source Code for Biology and Medicine. 2013 Dec;8(1):9. Available from: https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-9.

[15] Radkov AD, Hsu YP, Booher G, VanNieuwenhze MS. Imaging Bacterial Cell Wall Biosynthesis. Annual Review of Biochemistry. 2018 Jun;87(1):991–1014. Available from: https://www.annualreviews.org/doi/10.1146/annurev-biochem-062917-012921.

[16] Heim KP, Crowley PJ, Long JR, Kailasan S, McKenna R, Brady LJ. An intramolecular lock facilitates folding and stabilizes the tertiary structure of Streptococcus mutans adhesin P1. Proceedings of the National Academy of Sciences. 2014 Nov;111(44):15746–15751. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1413018111.

[17] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct;215(3):403–410. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

[18] Eddy SR. Profile hidden Markov models. Bioinformatics. 1998 Oct;14(9):755–763. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/14.9.755.

[19] wwPDB consortium, Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research. 2019 Jan;47(D1):D520–D528. Available from: https://academic.oup.com/nar/article/47/D1/D520/5144142.

[20] The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021 Jan;49(D1):D480–D489. Available from: https://academic.oup.com/nar/article/49/D1/D480/6006196.

[21] Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research. 2019 Nov;p. gkz1035. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1035/5614179.

[22] Larsen J, Schønheyder HC, Singh KV, Lester CH, Olsen SS, Porsbo LJ, et al. Porcine and Human Community Reservoirs of *Enterococcus faecalis* , Denmark. Emerging Infectious Diseases. 2011 Dec;17(12):2395–2397. Available from: http://wwwnc.cdc.gov/eid/article/17/12/10-1584_article.htm.

[23] Zong Y, Xu Y, Liang X, Keene DR, Höök A, Gurusiddappa S, et al. A 'Collagen Hug' Model for Staphylococcus aureus CNA binding to collagen. The EMBO Journal. 2005 Dec;24(24):4224–4236. Available from: http://emboj.embopress.org/cgi/doi/10.1038/sj.emboj.7600888.

[24] Miller OK, Banfield MJ, Schwarz-Linek U. A new structural class of bacterial thioester domains reveals a slipknot topology: Thioester Domains. Protein Science. 2018 Sep;27(9):1651–1660. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pro.3478.

[25] Walden M, Edwards JM, Dziewulska AM, Bergmann R, Saalbach G, Kan SY, et al. An internal thioester in a pathogen surface protein mediates covalent host binding. eLife. 2015 Jun;4:e06638. Available from: https://elifesciences.org/articles/06638.

[26] Bublitz M, Holland C, Sabet C, Reichelt J, Cossart P, Heinz DW, et al. Crystal Structure and Standardized Geometric Analysis of InlJ, a Listerial Virulence Factor and Leucine-Rich Repeat Protein with a Novel Cysteine Ladder. Journal of Molecular Biology. 2008 Apr;378(1):87–96. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022283608002039.

[27] Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECOD database: ECOD Manual Classification Strategies. Proteins: Structure, Function, and Bioinformatics. 2015 Jul;83(7):1238–1251. Available from: https://onlinelibrary.wiley.com/doi/10.1002/prot.24818.

[28] Symersky J, Patti JM, Carson M, House-Pompeo K, Teale M, Moore D, et al. Structure of the collagen-binding domain from a Staphylococcus aureus adhesin. Natural Structural Biology. 1997 Oct;4(10):833–838. Available from: http://www.nature.com/doifinder/10.1038/nsb1097-833.

[29] Sequeira S, Kavanaugh D, MacKenzie DA, Šuligoj T, Walpole S, Leclaire C, et al. Structural basis for the role of serine-rich repeat proteins from *Lactobacillus reuteri* in gut microbe–host interactions. Proceedings of the National Academy of Sciences. 2018 Mar;115(12):E2706–E2715. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1715016115.

[30] Gruszka DT, Wojdyla JA, Bingham RJ, Turkenburg JP, Manfield IW, Steward A, et al. Staphylococcal biofilm-forming protein has a contiguous rod-like structure. Proceedings of the National Academy of Sciences. 2012 Apr;109(17):E1011–E1018. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1119456109.

[31] Xu Y, Liang X, Chen Y, Koehler TM, Höök M. Identification and Biochemical Characterization of Two Novel Collagen Binding MSCRAMMs of Bacillus anthracis. Journal of Biological Chemistry. 2004 Dec;279(50):51760–51768. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820676098.

[32] Robledo Retana T, Bradley□Clarke J, Croll T, Rose R, Hoti I, Stagg AJ, et al. Lig v 1 structure and the inflammatory response to the Ole e 1 protein family. Allergy. 2020 Sep;75(9):2395–2398. Available from: https://onlinelibrary.wiley.com/doi/10.1111/all.14351.

[33] Li J, Kasper DL, Ausubel FM, Rosner B, Michel JL. Inactivation of the C protein antigen gene, bca, by a novel shuttle/suicide vector results in attenuation of virulence and immunity in group B Streptococcus. Proceedings of the National Academy of Sciences. 1997 Nov;94(24):13251–13256. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.94.24.13251.

[34] Creti R, Fabretti F, Koch S, Huebner J, Garsin DA, Baldassarri L, et al. Surface Protein EF3314 Contributes to Virulence Properties of *Enterococcus faecalis*. The International Journal of Artificial Organs. 2009 Sep;32(9):611–620. Available from: http://journals.sagepub.com/doi/10.1177/039139880903200910.

[35] Bianchetti CM, Takasuka TE, Deutsch S, Udell HS, Yik EJ, Bergeman LF, et al. Active Site and Laminarin Binding in Glycoside Hydrolase Family 55. Journal of Biological Chemistry. 2015 May;290(19):11819–11832. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820484960.

[36] Hendrickx APA, Poor CB, Jureller JE, Budzik JM, He C, Schneewind O. Isopeptide bonds of the major pilin protein BcpA influence pilus structure and bundle formation on the surface of Bacillus cereus: BcpA pilus structure. Molecular Microbiology. 2012 Jul;85(1):152–163. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2012.08098.x.

[37] Deutscher J, Saier Jr MH. Ser/Thr/Tyr Protein Phosphorylation in Bacteria – For Long Time Neglected, Now Well Established. Journal of Molecular Microbiology and Biotechnology. 2005;9(3-4):125–131. Available from: https://www.karger.com/Article/FullText/89641.

[38] Macek B, Mijakovic I, Olsen JV, Gnad F, Kumar C, Jensen PR, et al. The Serine/Threonine/Tyrosine Phosphoproteome of the Model Bacterium Bacillus subtilis. Molecular & Cellular Proteomics. 2007 Apr;6(4):697–707. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1535947620317485.

[39] Cozzone AJ. Role of Protein Phosphorylation on Serine/Threonine and Tyrosine in the Virulence of Bacterial Pathogens. Journal of Molecular Microbiology and Biotechnology. 2005;9(3-4):198–213. Available from: https://www.karger.com/Article/FullText/89648.

[40] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. Bioinformatics; 2021. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034.

[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

[42] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.3; 2015.

[43] Holm L. DALI and the persistence of protein shape. Protein Science. 2020 Jan;29(1):128–140. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pro.3749.

Supplementary

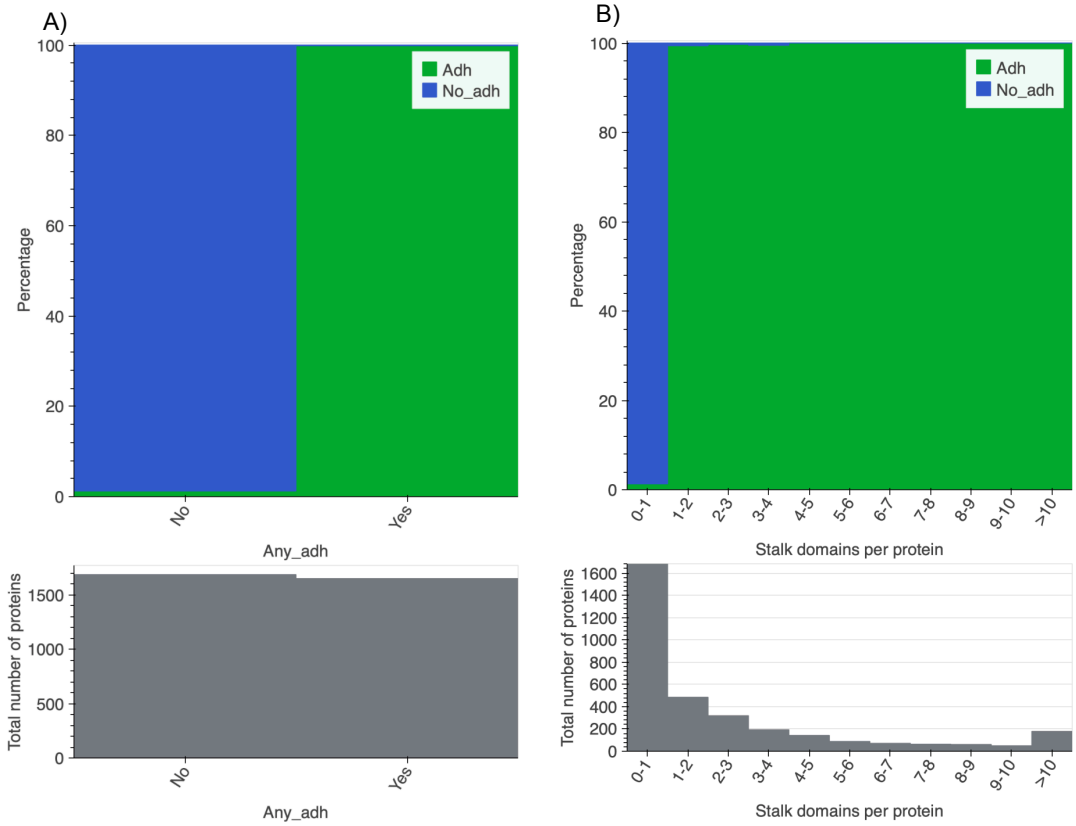

Figure S1: Comparison of positive and negative training set regarding known adhesive and stalk domains: These feature comparison plots are composed out of two subplots. The plots below represent the total number of training data proteins (y-axis) found per feature category (x-axes) and the plots above shows the percentage of proteins (y-axis) of the positive (green) compared to the negative (blue) training data found per feature category (x-axes). The two subplots shown in this figure represent the differentiation between the positive and negative training data regarding (a) the presence of a known adhesive domains, whereby the x-axes category represents whether or not any adhesive domain is present and (b) the number of stalk domains found per proteins, whereby the x-axes is categorized into 0 up to over 10 stalk domains per protein.

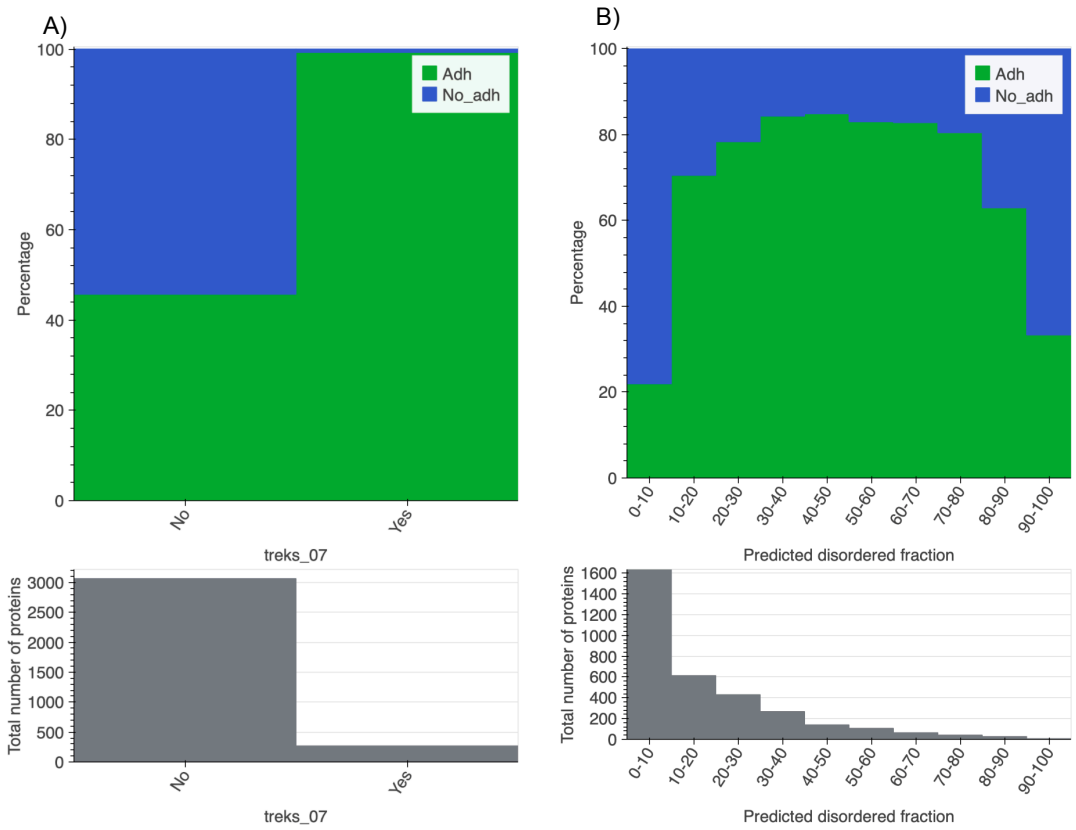

Figure S2: Comparison of positive and negative training data regarding the detection of unknown stalk domains: These plots represent feature comparison plots as described in supplementary figure S1. (a) Comparison regarding high identical tandem repeat domains found by T-REKS with the x-axes category into whether or not the proteins were detected by T-REKS. (b) Comparing the positive and negative training data regarding their predicted disordered regions by IUPred. The x-axes represents the percentage of sequence fraction predicted as disordered by IUPRED split into ten categories ranging from 0 to 100%.

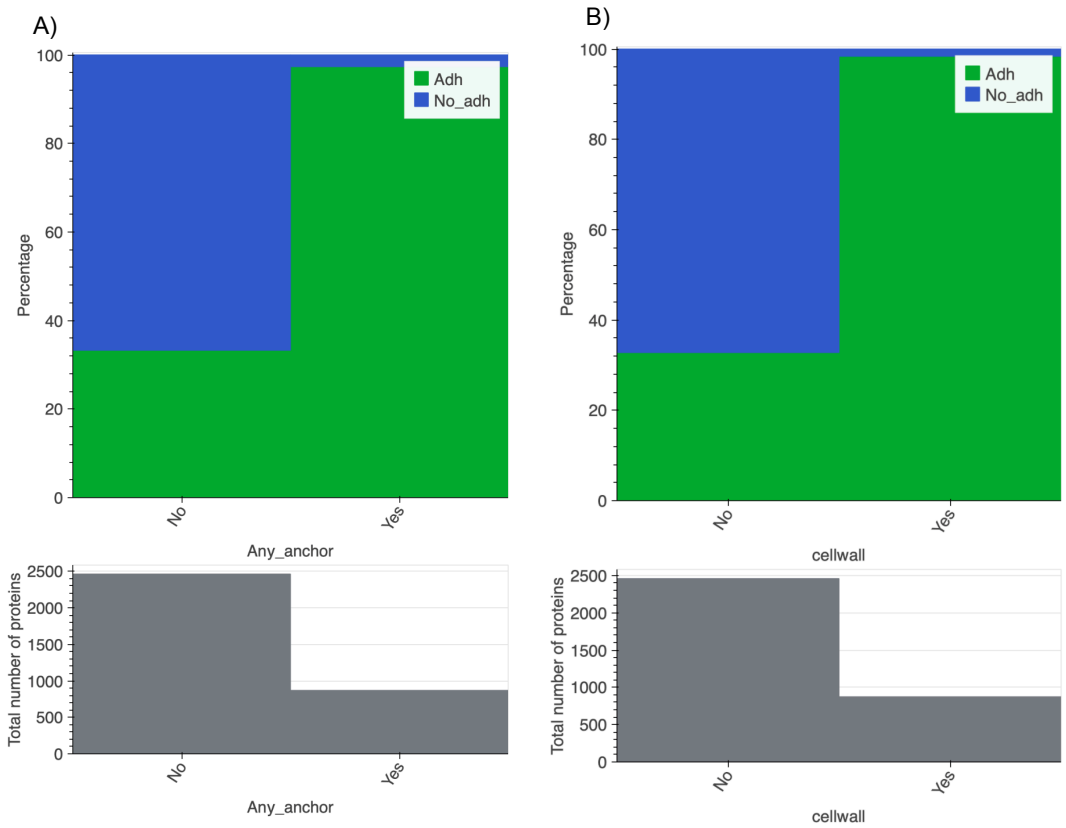

Figure S3: Comparison of positive and negative training data regarding the detection of surface proteins: These plots represent feature comparison plots as described in supplementary figure S1. a) Comparing the negative and positive training data set regarding whether or not (a) any anchor domain or motif can be found per protein or (b) the protein's localisation is predicted by Inmembrane to be at the cell surface.

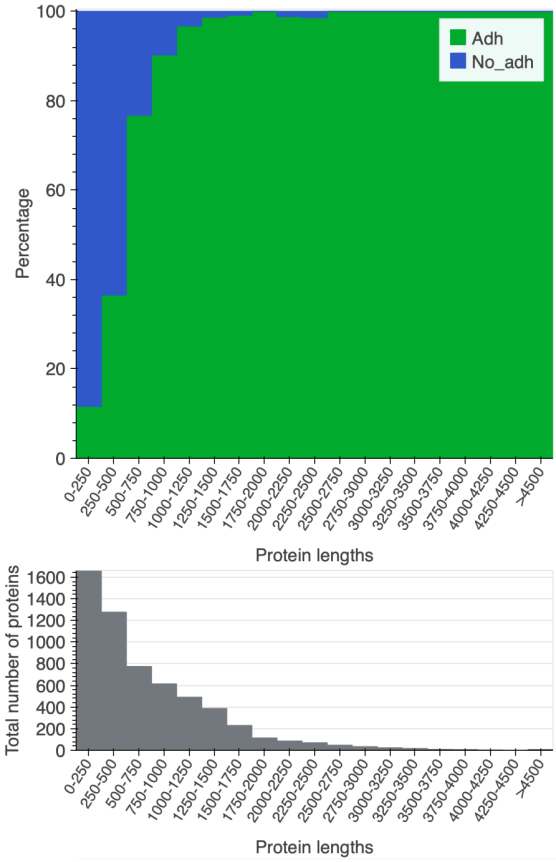

Figure S4: Comparison of positive and negative training data regarding the protein sequence length: This plot represents a feature comparison plot as described in supplementary figure S1. Here, the protein length of the positive and negative training data proteins is compared. The x-axes represent different length categories.

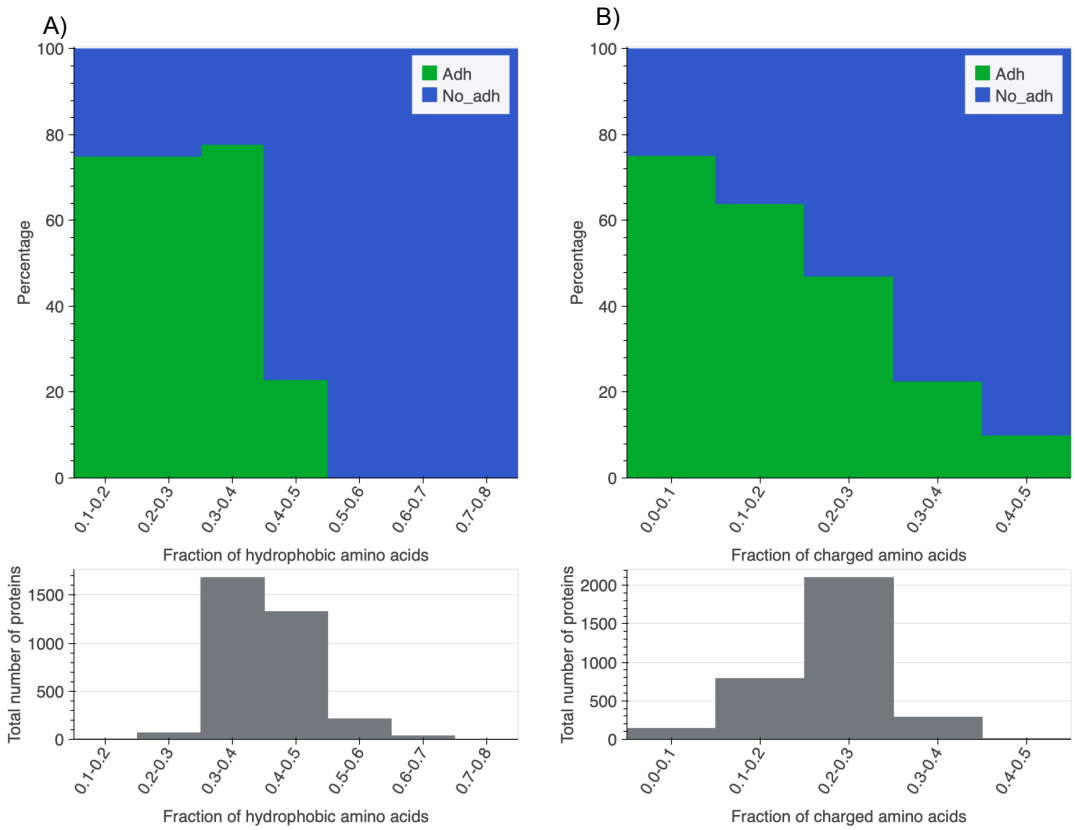

Figure S5: Comparison of positive and negative training data regarding the fraction of hydrophobic and charged amino acids: These plots represent feature comparison plots as described in supplementary figure S1. In these plots the fraction of (a) hydrophobic and (b) charged amino acids per protein sequence are compared between the positive and negative training data. The x-axes represent different fraction categories.

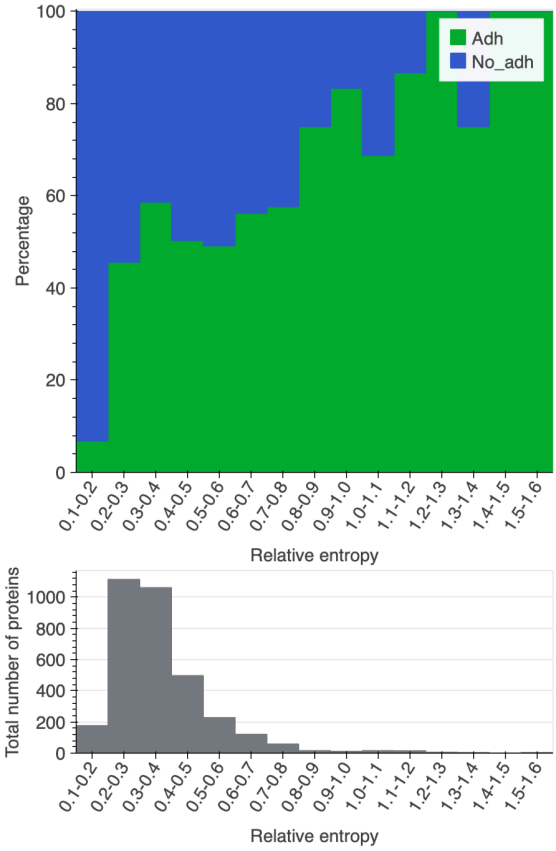

Figure S6: Comparison of positive and negative training data regarding the sequence composition bias (relative entropy): This plot represents a feature comparison plot as described in supplementary figure S1. Here, the relative entropy of the positive and negative training data protein sequences is compared. The x-axes represent different relative entropy value categories.

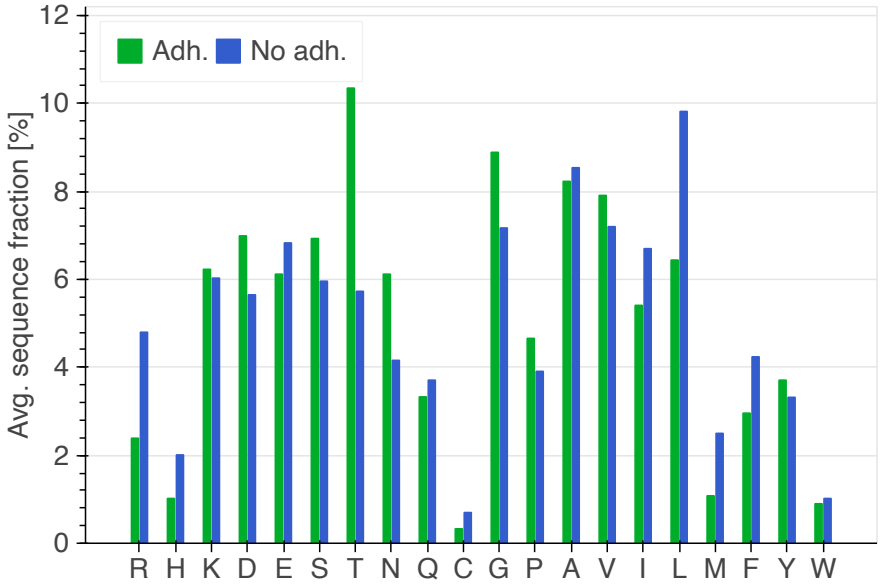

Figure S7: Average amino acid fraction: This bar plot shows for each amino acid the average fraction calculated per positive (Adh.) and negative (No adh.) training data set.

Table S1: Representative sequences: this table lists the representative sequences and the regarding domain positions based on the AlphaFold structure prediction for the N-terminal annotation gaps of predicted FA-like proteins with minimum 4 stalk domains, lacking a known adhesive domain. The DALI search was conducted with the sequence regions listed here.

Nr.	UniProt identifier	Domain info	Start position	End position
1	K8EVB1	N-term	117	256
		C-term	257	417
2	A0A2Z6T9E9		185	718
3	E3GH89		31	227
4	A0A329MSD5		48	279
5	A0A3D9IIT5	N-term	39	172
		C-term	175	309
6	A0A5R8Q9T8		68	240
7	A0A1Q8E8C7		76	420
8	J9W466		30	163
9	A0A4U7JL97		37	420
10	B0S3M8		153	316
11	X0SH12		37	142

12	A0A494X6E7		38	377
13	A0A559K5W2		120	420
14	C6CUY3		40	363
15	A0A0F7RLJ7		46	326
16	R6V4J4		37	420
17	A0A069CUH0		64	357
18	R3TX93		40	276
19	A0A2Z4U801		240	508
20	V2XMF4	N-term	80	239
		C-term	240	366
21	A0A099WCN8		48	417

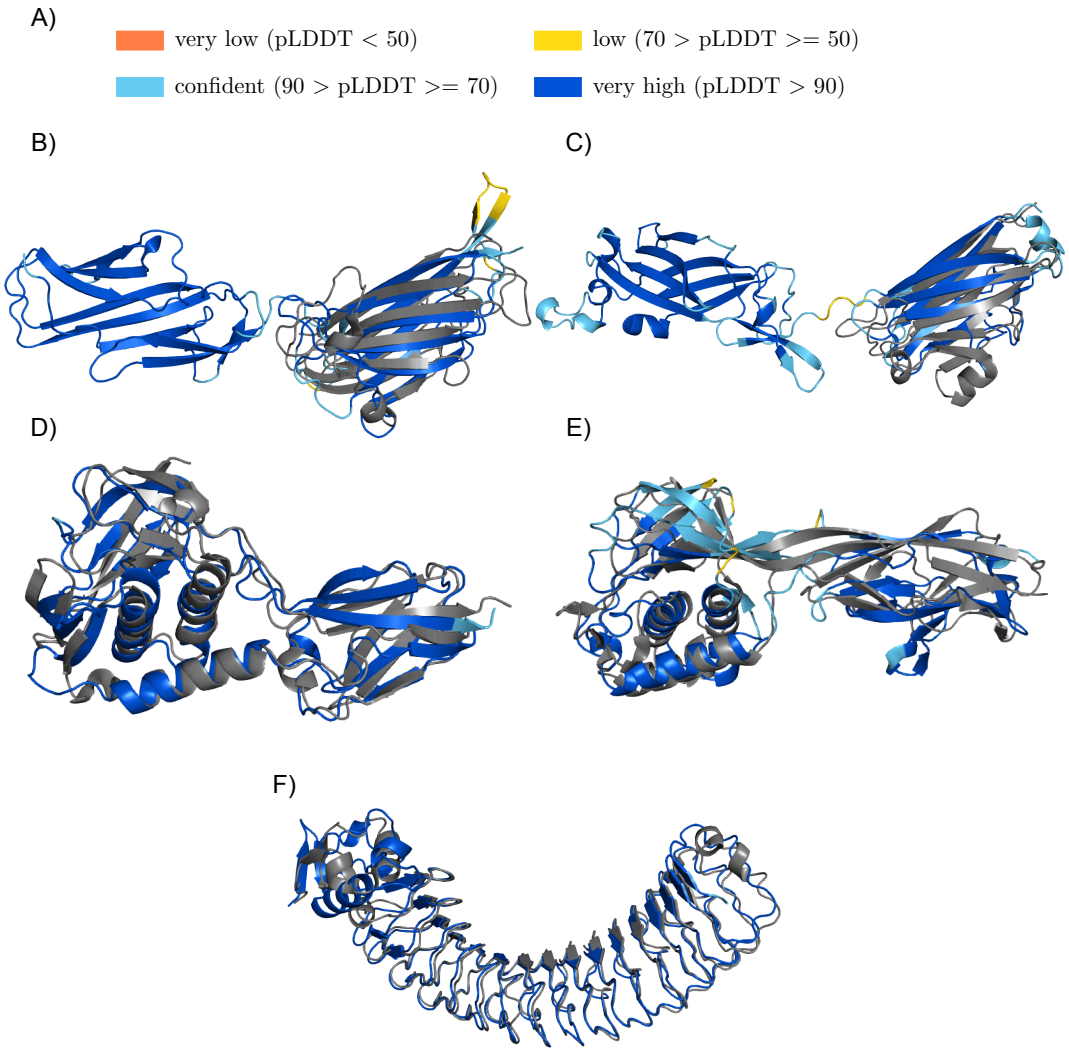

Figure S8: Superposition of structure models shown in figure 5 with distantly related adhesive domains found by jackhmmer: a) Colour legend representing the quality of the AlphaFold models. b) Superposition of the structure model of cluster group 1 with the SdrG_C_C binding domain (PDB:4jdz:A). c) Superposition of Collagen_bind domain (PDB:2z1p:A) to the C-terminal domain of the cluster group 20 structure model. d) Predicted structure model for cluster group 18 superposed with S. aureus TED adhesive domain (PDB:6fx6:A). e) Predicted structure model of group 19 superposed with the Bacillus anthracis TED adhesive domain (PDB:6fwv:B). f) Predicted structure model for cluster group 9 superposed with the functional region of Internalin J (PDB:3bz5:A).

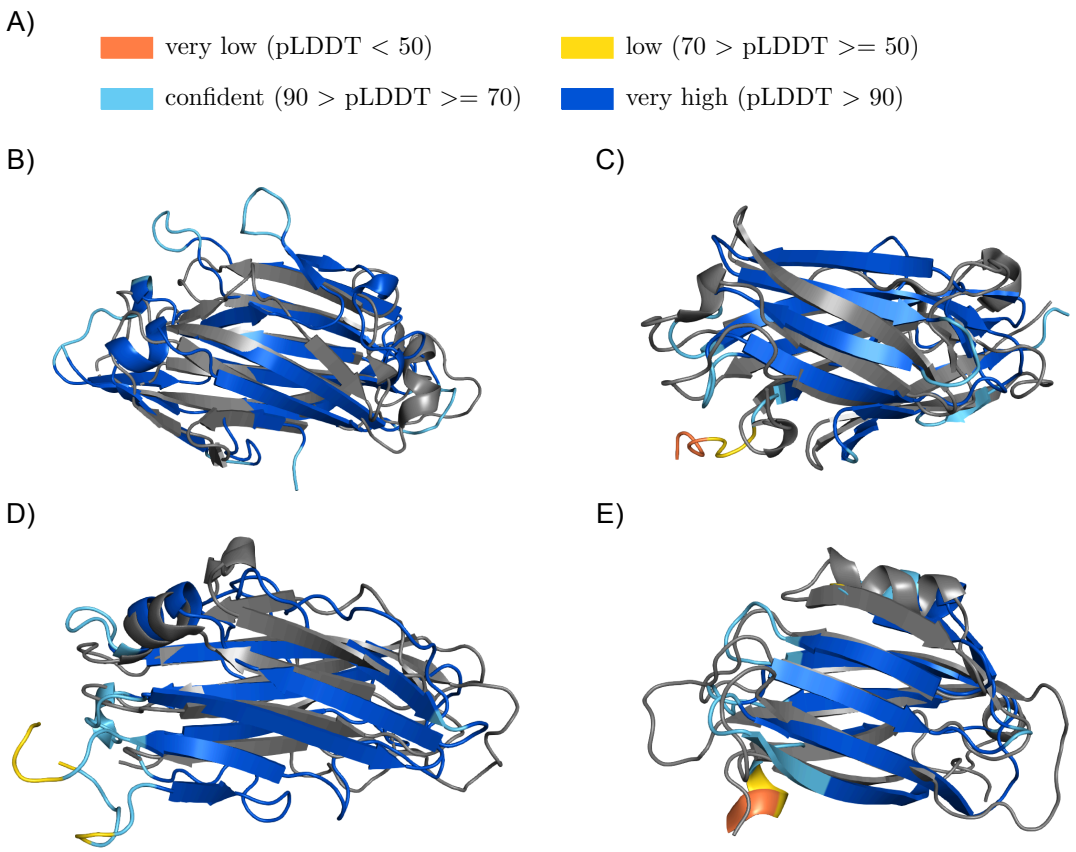

Figure S9: Superposition with the DALI detected PDB structures similar to the clusters showing a jelly-roll like structure. a) AlphaFold colour legend for the structure prediction confidence. b) Predicted structure model for cluster group 6 superposed with the GramPos_pilinBB domain (PDB:5xcb:A). c) Predicted structure model for cluster group 8 superposed with the Big_8 domain (PDB:4b5z:A). d) Superposition of the structure model for cluster group 10 with the GramPos_pilinBB domain (PDB:2x9x:A). e) The predicted structure for cluster group 11 superposed with the A-macroglobulin receptor binding domain in cattles (PDB:1ayo:B).

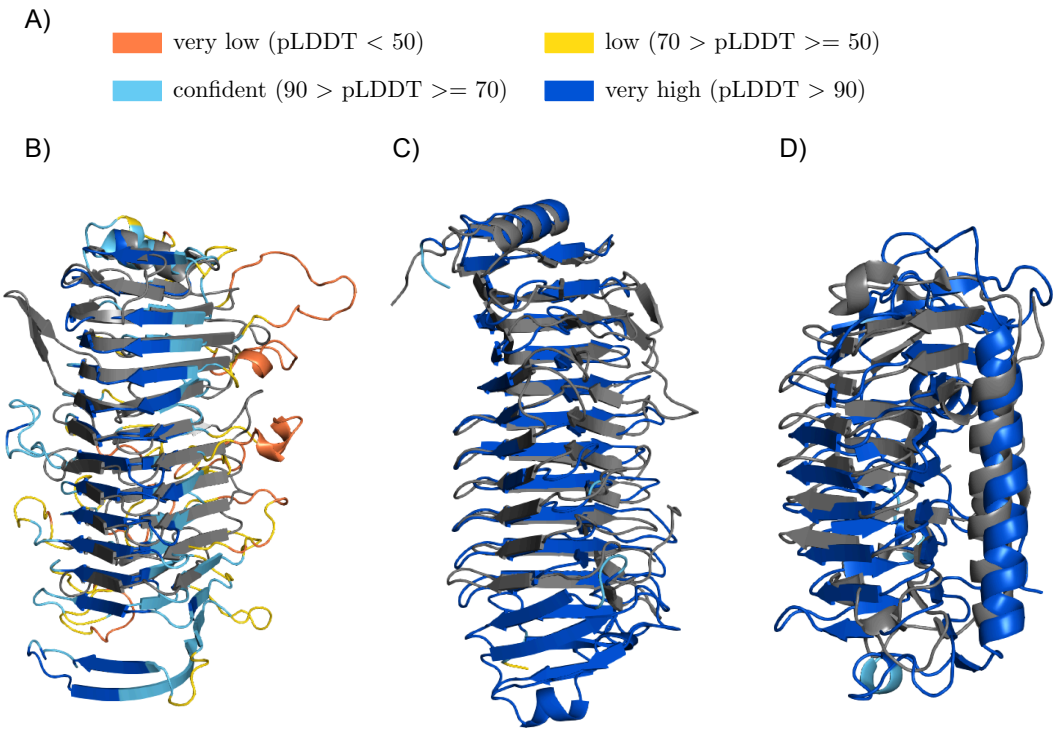

Figure S10: Beta-solenoid predicted structure superposed with the top DALI matches in the PDB database. a) AlphaFold prediction confidence colour legend. Superposition with the adhesive region of the serine-rich Lactobacillus reuteri adhesin (PDB:5ny0:A) with the b) predicted structure model for cluster group 2 and c) predicted structure for cluster group 21. d) Superposition of the structure model for cluster group 15 with the Ice_binding domain (PDB:4nuh:A).

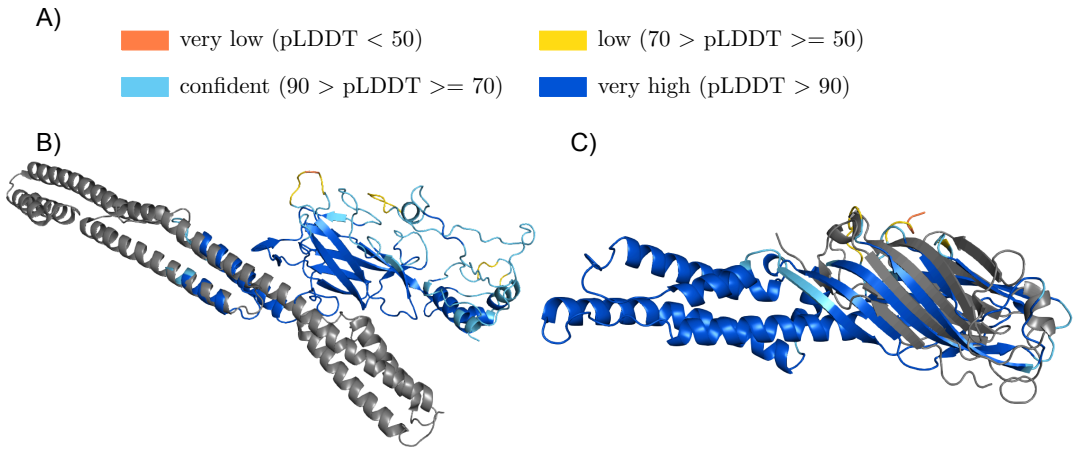

Figure S11: Superposition of clusters with ambiguous adhesion function with DALI hits: a) Confidence colour legend for AlphaFold predicted structures b) Cluster 7 structure model superposed with the immunogenic adhesin BibA (PDB:6poo:A). c) Cluster 17 structure model superposed with the S. gordonii Sgo0707_N2 domain (PDB:4igb:B).

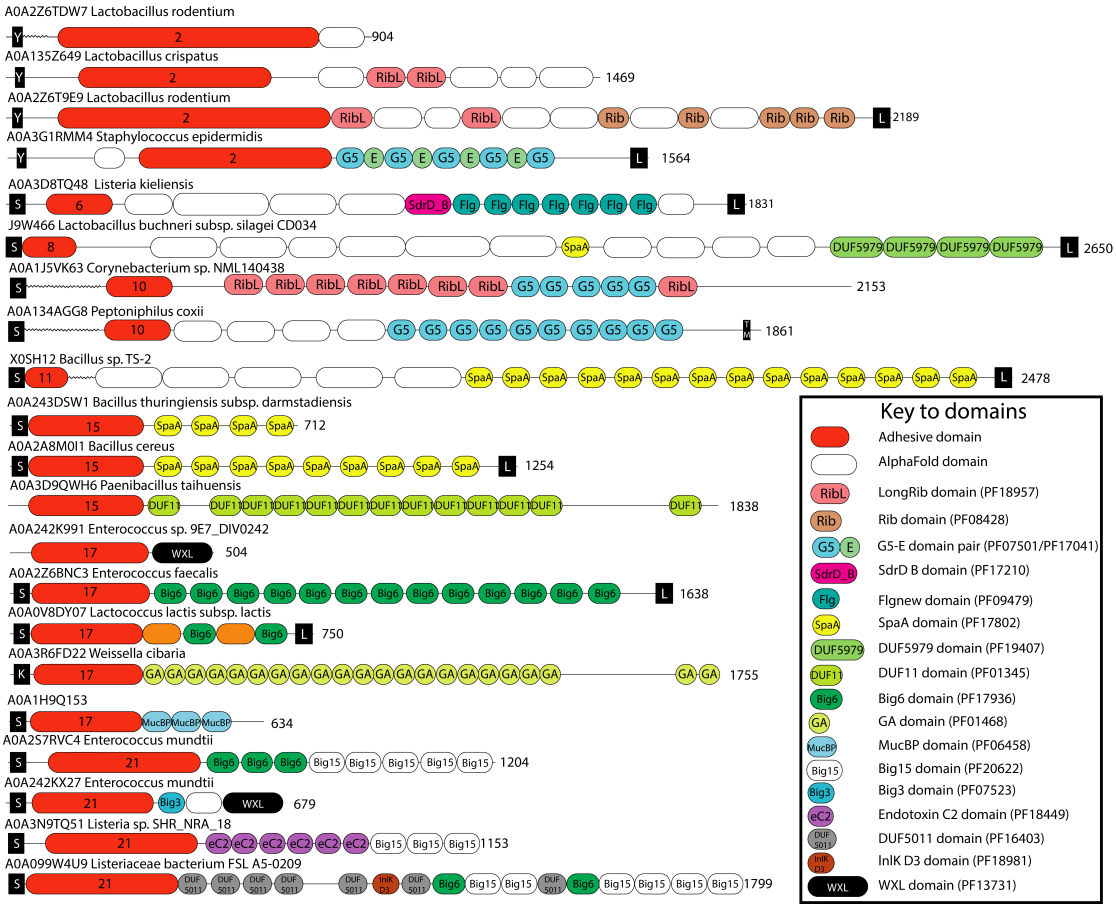

Figure S12: Examples of protein architectures in which the potential novel adhesive domains can be found: The potential novel adhesive domains are annotated in red, whereby the number indicates the regarding cluster number.

Table S2: Pfam families created based on this study

Cluster	Pfam accession	Pfam identifier
2+21	PF20585	Pectate_lyase_5
6	PF20595	pAdhesive_6
8	PF20602	pAdhesive_8
10	PF20592	pAdhesive_10
11	PF20596	pAdhesive_11
15	PF20597	pAdhesive_15
17	PF20609	pAdhesive_17

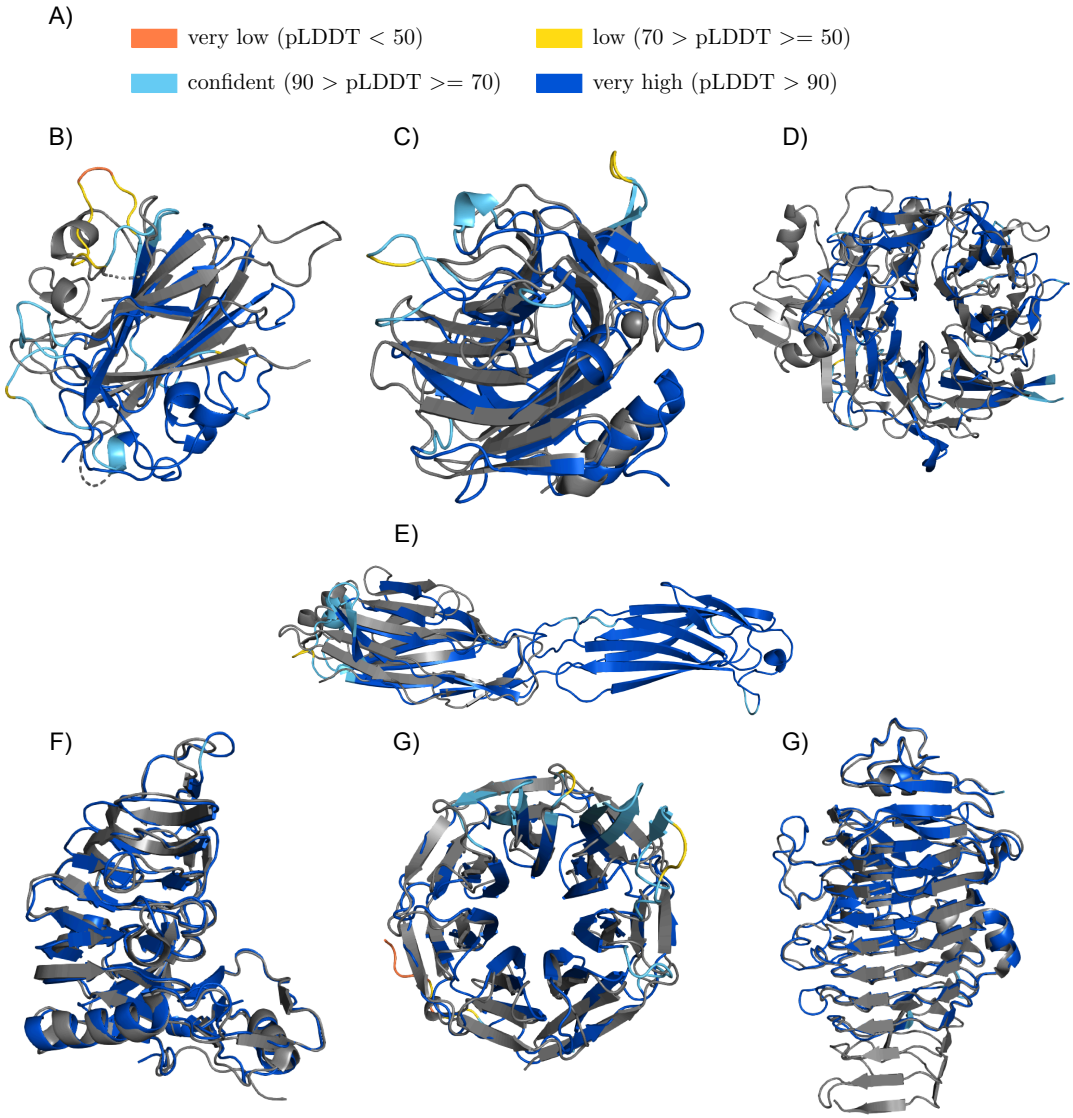

Figure S13: Predicted structures for clusters with unlikely adhesion function: (a) Colour legend for AlphaFold models quality. (b) Superposition of cluster 3 structure model with the Calpain_III domain (Pfam:PF01067) of a calpain, which are Ca^{2+}-dependent cysteine proteases (PDB:1qxp:B). (c) Cluster 4 structure model superposed with an extracellular arabinase (PDB:5hp6:A). (d) Cluster 12 structure model aligned with an human integrin alpha-IIb/beta-3, which binds among others to fibrinogen (PDB: 6v4p:A). (e) Cluster 5 predicted structure superposed with the Bacillus cereus BcpA pilus subunit (PDB:3rkp:A). (f) Cluster 13 structure model superposed with a glycoside hydrolase (PDB:4pex:B). (g) Superposition of cluster 14 structure model with a virginiamycin B lyase (PDB:2z2o:B). (h) Cluster 16 structure model superposed with an Eubacterium lacto-N-biosidase (PDB:6kqs:A).

Table S3: Structure model DALI results info: The table lists the top DALI hits of the AlphaFold models for the N-terminal sequence clusters with the regarding Z-scores and similarity information when the structures are superposed using pymol. The Notes column gives information about the similarities to structures alternative to the top DALI hit, which were described in the main text.

Nr.	Domain	Top DALI hit	Z-score	Superposition (RMSD)	Notes
1	N-term	4b60:B	11.5	3.42 Å over 128 res.	
	C-term	4igb:B	13.7	3.27 Å over 128 res.	4jdz:A -> 3.74 Å over 136 res.
2		5ny0:A	27.3	5.07 Å over 272 res.	
3		1qxp:B	6.1	4.92 Å over 104 res.	
4		5hp6:A	16.2	3.55 Å over 176 res.	
5	N-term	3rkp:A	13.6	3.95 Å over 128 res.	
	C-term	3kpt:B	12.6	3.63 Å over 120 res.	
6		5xcb:A	10.8	4.52 Å over 128 res.	
7		6poo:A	10.5	3.24 Å over 64 res.	4jdz:A -> 3.89 Å over 104 res.
8		4b5z:A	10.6	5.09 Å over 128 res.	
9		3bz5:A	37.7	3.00 Å over 352 res.	
10		2x9x:A	7.5	4.43 Å over 112 res.	
11		6avu:A	10.6	4.94 Å over 104 res.	1ayo:B -> 4.08 Å over 104 res.
12		6v4p:A	28.9	4.64 Å over 272 res.	
13		4pex:B	44.0	1.27 Å over 296 res.	
14		2z2o:B	31.1	3.04 Å over 272 res.	
15		4nuh:A	15.2	4.30 Å over 176 res.	
16		6kqs:A	52.8	1.42 Å over 352 res.	
17		7nxd:A	10.1	6.96 Å over 72 res.	4igb:B -> 4.88 Å over 72 res.
18		6fx6:A	20.8	3.78 Å over 224 res.	
19		6fwv:B	13.6	7.61 Å over 224 res.	
20	N-term	5cf3:A	11.1	4.57 Å over 128 res.	
	C-term	2z1p:A	11.7	3.50 Å over 120 res.	
21		5ny0:A	31.1	3.39 Å over 280 res.	