**Accurate contact-based modelling of repeat proteins predicts the structure of new repeats protein families.**

Claudio Bassot\* and Arne Elofsson\*

\* Science for Life Laboratory and Dep of Biochemistry and Biophysics, Stockholm University.

**Abstract**

Repeat proteins are abundant in eukaryotic proteomes. They are involved in many eukaryotic specific functions, including signalling. For many of these proteins, the structure is not known, as they are difficult to crystallise. Today, contact-predictions based on direct coupling analysis and lately, deep learning often provides sufficient information to predict the structure of many protein families. However, the unique sequence features present in repeat proteins have been a challenge to use direct coupling analysis for predicting contacts. Here, we show that deep learning-based methods (DeepMetaPsicov (DMP) or PconsC4) are more useful for predicting both intra- and inter-unit contacts among a comprehensive set of repeat proteins. In a benchmark dataset of 815 repeat proteins about one third can be correctly modelled and among 49 PFAM families lacking a protein structure, we produce models of eleven families with estimated high accuracy.

**Author Summary**

Repeat proteins are widespread among organisms and particularly abundant in eukaryotic proteomes. Their primary sequence presents repetition in the amino acid sequences that origin structures with repeated folds/domains. Although the repeated units often can be recognised from the sequence alone, often structural information is missing. Here, we used contact prediction for predicting the structure of repeats protein directly from their primary sequences. We benchmark our method on a dataset comprehensive of all the known repeated structures. We evaluate the contact predictions and the obtained models for different classes of repeat proteins. Further, we develop and benchmark a quality assessment (QA) method specific for repeat proteins. Finally, we used the prediction pipeline for all PFAM repeat families without resolved structures and found that eleven of them could be modelled with high accuracy.

**Introduction**

Repeat proteins contain periodic units in the primary sequence that are likely the result of duplication events at the genetic level [[1]](https://paperpile.com/c/NMPUUL/7BHY). Repeat proteins emerge through replication slippage [[2]](https://paperpile.com/c/NMPUUL/Gkvz) and double-strand break repair [[3]](https://paperpile.com/c/NMPUUL/Ind2). This protein class is present in all genomes but is more frequent in eukaryotic organisms [[4–6]](https://paperpile.com/c/NMPUUL/SgoR+RDz1+2HFJ) where they are involved in a wide range of functions [[7]](https://paperpile.com/c/NMPUUL/5Fjy). In particular, due to their extended structures, repeat proteins often behave as molecular scaffolds in protein signalling or for protein complexes as WD40 domain [[8]](https://paperpile.com/c/NMPUUL/nKVc), or ankyrin repeats [[9,10]](https://paperpile.com/c/NMPUUL/HLpY+WWdy). Repeat proteins are usually conserved among orthologs [[4,11]](https://paperpile.com/c/NMPUUL/W73P+SgoR) while exhibiting a more accelerated evolution and divergence among paralogs [[11]](https://paperpile.com/c/NMPUUL/W73P).

A classification of repeat proteins was proposed by Kajava [[12,13]](https://paperpile.com/c/NMPUUL/AdXS+lEFg) based on the length of the repeat units and the tertiary structure of the repeat units. According to Kajava’s classification, there are five classes of repeat proteins. However, in this study, we ignore class I and II because there are no available structures for class I, and class II structures are folded in a coiled-coil structure possible to predict using other methods. Moreover, the extreme amino acid compositional bias of many of these proteins makes it difficult to identify the coevolving residues in these two classes.

The dataset used in our study contains three classes of proteins divided into 20 subclasses by their secondary structure, according to RepeatsDB [[14]](https://paperpile.com/c/NMPUUL/z3tG), Figure 1. The three types are; class III extended repeats (e.g. ɑ and β solenoids); class IV closed repeats structures (e.g. TIM and β barrels and β-propeller), and class V where the units appear as separate domains on a string. Further, in class V the repeat units are longer than in the other classes.

***Figure1. Repeats proteins classification.*** *Representation of the repeats classes and subclasses as classified in repeatsDB 2.0* [*[14]*](https://paperpile.com/c/NMPUUL/z3tG).

The solenoid structures (subclasses III.1, III.2 III.3) dominate Class III [[13]](https://paperpile.com/c/NMPUUL/lEFg), and these proteins contain a wide range of repeated units (from 4 to 38), Figure 1. The length of the individual unit is also quite variable (from 10 to 50 residues) [[14]](https://paperpile.com/c/NMPUUL/z3tG), with β-solenoids having significantly shorter repeats compared with α and α/β solenoid [[13]](https://paperpile.com/c/NMPUUL/lEFg).

Members of class IV are constrained in variability by the closed fold. Indeed, despite ten subclasses of different units, the number of units varies between 3 and 16, and proteins with more than ten repeat units are rare. Even in this class, the length of the repeats units varies between 10 to 50 residues [[13]](https://paperpile.com/c/NMPUUL/lEFg). Finally, class V proteins are made up of the extended repeat units, often longer than 40 residues [[14]](https://paperpile.com/c/NMPUUL/z3tG), where each unit folds into proper domains, and they only have few inter-unit contacts.

Many repeat protein families lack a resolved structure. For these protein families, residue-residue contact prediction is the best method to obtain structural information [[15]](https://paperpile.com/c/NMPUUL/27Hch). Contact prediction methods use residue-residue co-evolution from multiple sequence alignment and identify the evolutionary constraints of the residues imposed by the tertiary protein structure [[16]](https://paperpile.com/c/NMPUUL/gK7oZ). Nevertheless, repeat proteins are a difficult target for contact prediction; the internal symmetry introduces artefacts in the contact map at a distance corresponding to the repeated units [[17]](https://paperpile.com/c/NMPUUL/4uCf).

Here, we benchmark the deep-learning-based contacts prediction programs PconsC4 [[18]](https://paperpile.com/c/NMPUUL/kw3D) and DeepMetaPsicov [[19]](https://paperpile.com/c/NMPUUL/AtyZ) against the GaussDCA [[20]](https://paperpile.com/c/NMPUUL/LRiE) on a comprehensive dataset generated from RepeatsDB [[14]](https://paperpile.com/c/NMPUUL/z3tG). The predicted contacts were then used as constraints to generate protein models, and the model quality was evaluated using Pcons [[21]](https://paperpile.com/c/NMPUUL/hxH5). Based on the benchmark, we propose models for the protein structures of PFAM protein families missing resolved structures.

**Results and Discussion**

**General contact prediction analysis in repeat proteins**

To assess the quality of the contacts predictions among repeat protein classes, we generate a dataset of proteins using the reviewed entries of RepeatsDB [[14]](https://paperpile.com/c/NMPUUL/z3tG) and then clustered at 40% sequence identity. For each repeats region in the dataset, we also extracted a representative repeat unit and a pair of repeats, obtaining in this way three datasets: i) a single unit datasets; ii) a double unit datasets; iii) complete repeat region datasets.

For all the three sets, multiple sequence alignments (MSA) and secondary structure predictions were generated. Subsequently using the MSA as input for PconsC4, DeepMetaPsicov, and GaussDCA contacts were predicted for each family. The performance of the contact predictions was then evaluated for each subclass separately. As expected, PconsC4 outperforms GaussDCA in all the three sets and all the classes of repeat proteins, and DeepMetaPsicov, the most recent method outperforms both, Figure 2.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure2.png

*Figure2.* ***The precision of contact predictions.*** *Positive Predictive Value (PPV) for the GaussDCA (red), Pconsc4 (Blue), and DeepMetaPsicov (green). For all three methods results are shown for the three datasets, in light colour the single unit dataset, in intermediate colour the double units dataset, and in the darker colour the complete region dataset.*

Here, it should be remembered that PconsC4 and DeepMetaPsicov, in addition to other information, use DCA predictions as an input and then learn to recognise specific patterns [[18]](https://paperpile.com/c/NMPUUL/kw3D). Therefore, artefacts present in the DCA predictions might propagate into these methods.

In general, the predictions for the full-length regions give better results than when splitting the proteins into smaller units, Figure 2. However, in class V, which is composed of bigger units forming repeats of the *“beads on a string*” type, the splitting in units sometimes helps to reach better contact predictions, as discussed later.

In Figure 3, selected contact maps are shown as examples. The GaussDCA predictions contain periodic artefacts of wrong predictions (red dots) forming diagonal lines, occurring between equivalent positions in the repeat unit. PconsC4 and DeepMetaPsicov appear efficient in removing the artefacts seen in GaussDCA. Here, it can be noted that there is only limited overlap between our repeat protein set and the training set of PconsC4 and DeepMetaPsicov, 25 out of 2856 and 29 out of 3456 proteins are identical respectively. Further, the accuracy for the shared proteins does not show a higher precision than the other proteins, Supplementary Figure 1. Therefore, we are convinced that the results can be extrapolated to all repeat proteins.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%203.png

*Figure 3.* ***GaussDCA, PconsC4 and DeepMetaPsicov contact maps.*** Co*ntact map for predictions obtained with GaussDCA, PconsC4 and DeepMetaPsicov. In grey, the real contacts from the structure, in green, the corrected predicted contacts, and in red, the falsely predicted contacts.*

It is well known that the quality of the prediction is directly correlated with the number of sequences in the starting MSA, especially for DCA methods [[18]](https://paperpile.com/c/NMPUUL/kw3D). This trend is also confirmed for protein repeats where the repeats with a smaller MSA are predicted with less accuracy, Figure 4.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure4.png

*Figure 4.* ***The relation between Precision and the effective number of sequences in the MSA.*** *Positively Predicted Value for GaussDCA in red PconsC4 in Blue and DeepMetaPsicov in orange on the Neff value (the effective number of sequences length weighted with the length of the protein). The single dots correspond to each protein in the datasets, and the line is the running average on (n=50).*

**Differences among repeat classes in contacts prediction**

Figure 2 shows variations in the fraction of correctly predicted contacts among different protein repeat classes and subclasses. To clarify the origin of these differences, we investigated, more in-depth, the source of the predicted contacts. One central aspect that affects the difficulty of prediction is due to the pattern of the contacts [[22]](https://paperpile.com/c/NMPUUL/V7Po). In general, contacts that are parts of larger interaction areas or close in the sequence are predicted more accurately. Therefore, we compared the intra-unit and inter-unit contacts predicted by DeepMetaPsicov, Figure 5a. Here, we obtained the number of predicted intra and inter-unit contacts from the PDB structures and selected the same number of predicted intra- and inter-units contacts. The PPV was finally calculated using the number of correctly predicted contacts divided with the number of contacts.

On average the intra-units contacts are predicted with higher accuracy than the inter-unit contacts. However, this is not true for all protein classes, due to the differences of the structures among the classes: Class III units are short, and the residues are mostly in contact with the neighbour units; in class V, on the contrary, the units are long, folded in independent domains and the contacts are predominantly inside the units with few inter-unit contacts; class IV is halfway between class III and V. The inter-unit contacts in class III and (partially) class IV appear to be easier to predict than class V inter-unit contacts. It can be seen that they form detectable patterns in the contact maps. On the contrary, the intra-unit contacts of class V are predicted better than intra-unit contacts in class III and IV for the same reason.

In general, the PPV versus the ratio of the inter-unit contacts over the total number of contacts of each protein shows an inverse relation, Figure 5. The intra-unit PPV is low for proteins with an inter-contact ratio lower than 20i.e.i.e the vast majority of class V proteins, Figure 5c. In contrast, the inter-unit PPV is high for proteins with an inter/intra contact ratio of 80% or higher. The families with high inter/intra ration correspond largely to solenoid structures and TIM barrels that have 80%-100% larger interaction surfaces between different units than within a single repeat unit. In Figure 5d, we divide the proteins by their secondary structure class. In DeepMetaPsicov prediction, there is no significant variation between the secondary structure composition, while some difference can be seen in the other methods.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%205.png

**Figure 5. Predicted contacts analysis.** *Positive Predicted Value (PPV) obtained by DeepMetaPsicov for different types of contacts. a) Examples of inter- and intra- unit contacts. b) In red, the PPV for intra-units contacts in blue PPV for inter-units contact. c) Repeats subclasses; in red, the PPV for intra-units contacts, and in blue PPV for inter-units contact, colours and shapes in the scatter plot indicate different protein subclasses. d) Secondary structure, coloured in red, the overall PPV, in blue, the ɑ-helical subclasses, in green, the ɑ-helix/β-strand subclasses, and in orange, the β-strand subclasses.*

**Protein model generation**

Protein models were generated using CONFOLD [[23]](https://paperpile.com/c/NMPUUL/GtDi) using contact predictions from either PconsC4 or DeepMetaPsicov and combined with secondary structure predictions from PSIPRED. In Figure 6, we compare the TM-score between the first ranked model by CONFOLD and the corresponding PDB protein structure.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure6.png

**Figure 6. Protein model quality.** *TM-score for the subfamilies; Models from PconsC4 in blue and DeepMetaPsicov in yellow. The single unit is shown in a light colour, the double in an intermediate colour and the full-length protein in dark colours.*

Although the best contact predictions were, on average, obtained using the entire regions, splitting the structure lead in some cases to a better model; this is true in particular for class V where the prediction of the inter-unit interactions is challenging.

DeepMetaPsicov generates models with higher accuracy complete or double unit regions than PconsC4. In particular, in the “propeller” subclasses class IV: III.3 ɑ solenoid, IV.4 β propeller, IV.8 ɑ/β propeller, IV.5 ɑ/β prism, and IV.10 aligned prism. All these subclasses, but ɑ-solenoid, have a low ratio of inter-units contacts (below 50%), Figure 5b.

**Quality assessment of the models**

To evaluate the model quality, we compare the TM-scores of the models with the quality assessment scores from Pcons [[21]](https://paperpile.com/c/NMPUUL/hxH5) and QmeanDisCo [[24]](https://paperpile.com/c/NMPUUL/lax1). From Figure 7a and 7b, it is clear that both methods fail to rank a significant number of models properly.

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%207.png

***Figure 7. TM-score versus QA methods.*** *a) TM-score versus Pcons-score for complete region models generated with DeepMetaPsicov. b) TM-score versus* QmeanDisCo *-score for full region models created from DeepMetaPsicov contacts.*

To overcome the limitation in quality estimation, we developed a Random Forest Regression method using data from multiple sources (Pcons, QmeanDisCo, protein features, Confold). Five-fold cross-validation was performed on the complete region dataset. The method obtained an average accuracy of 86.1%, and an average absolute error of 0.06 TM-score, Figure 8a. The Random Forest Regression predicts the TM-score significantly better than Pcons and QmeanDisCo alone (Figure8b). We found that eight of the features were helpful for the prediction of the TM-score, Supplementary Figure 2. The most important features are the Pcons score, the local QmeanDisCo score, and the percentage of the dihedral angle restraints satisfied in the model obtained by Confold.

<https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%208.png>

Figure 8 *a) Real TM-score versus Random Forest Predicted TM-score for complete region models generated with DeepMetaPsicov. b) Pearson correlation coefficient between the TM-score and the QA methods.*

**Modelling of repeat protein families without known structures.**

We selected 49 PFAM repeat-families without resolved structure and fed them through the structure prediction pipeline described above. For all families, except Curlin (see below), the estimated TM-score was higher for DeepMetaPsicov than for PconsC4.

Here, we discuss the 11 models with a predicted TM-score higher than 0.55, Table 1. For eight of these families, we found a template with a GMQE score [[25]](https://paperpile.com/c/NMPUUL/3xrz) higher than 0.4 using Swissmodel [[26]](https://paperpile.com/c/NMPUUL/yDw0). In these cases, homology models were generated for comparison with the contact based models. We compared the similarity of the contact-based and homology-based models with the predicted TM-score for the contact-based model. For five families, the models obtained by homology agree with the predicted TM-score, the difference between the TM-scores is below 0.1, i.e. the estimated TM-score agrees with what would be estimated if the homology model was identical to an experimental structure. However, for the other three families, there are significant differences, see Fig 9.

In the model of *WD40\_alt* (PF00400) the position of the loops differs significantly, while in *Morn2* (PF07661) a bend in the homology model template is not present in the contact-based model. Finally, the two models of *Bacterial tandem repeat domain* (PF17660) differ in many aspects. Here, it can be noted that both the contact-based and homology-based models of *Bacterial tandem repeat domain* have predicted qualities just slightly above the threshold used. The remaining three PFAM (DUF2963, SPW, Curlin) families do not have suitable templates, and, therefore, we cannot compare their models with a homology-based model, but they are discussed below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *PFAM family* | *PFAM ID* | *Representative*  *protein*  *Uniprot ID* | *TM prediction* | *Template*  *PDB ID* | *GMQE* | *Identity* | *TM-score between*  *The contact model*  *and Homology model* |
| *WD40\_alt* | *PF00400* | *G3VIY2* | *0.58* | *5mzh\_A* | *0.67* | *31.1%* | *0.38* |
| *Plasmodium*  *repeat MYXSPDY* | *PF05593* | *A0A1G0MXS8* | *0.59* | *6h6e\_F* | *0.63* | *22.2%* | *0.50* |
| *MORN 2* | *PF07661* | *Q8RH85* | *0.68* | *1muf\_A* | *0.62* | *21.4%* | *0.39* |
| *FG-GAP\_2* | *PF14312* | *W4LGN0* | *0.60* | *3fcs\_A* | *0.58* | *26.6%* | *0.51* |
| *RTTN\_N* | *PF14726* | *W5P499* | *0.71* | *4v3o\_A* | *0.51* | *17.6%* | *0.65* |
| *DUF5122* | *PF17164* | *A0A1Z4C3E9* | *0.58* | *6d69\_A* | *0.46* | *19.5%* | *0.53* |
| *Bacterial tandem*  *repeat domain 1* | *PF17660* | *A0A252E8A5* | *0.56* | *5x3j\_A* | *0.41* | *16.4%* | *0.19* |
| *SBBP* | *PF06739* | *U5QIU9* | *0.57* | *6ske\_A* | *0.40* | *20.6%* | *0.49* |
| *DUF2963* | *PF11178* | *Q6YQH3* | *0.62* | *5e9u\_A* | *0.38* | *16.7%* | *N/A* |
| *SPW* | *PF03779* | *A0A2A3HD64* | *0.65* | *None* | ***-*** | *-* | *-* |
| *Curlin* | *PF07012* | *Q8EIH3* | *0.67* | *None* | ***-*** | *-* | *-* |

**Table 1 The models of the PFAM families with a predicted TM-score above 0.55.** In the columns: the family name, the PFAM ID, the Uniprot ID of the sequence used for the modelling, the predicted TM-score, the best template PDB ID, the Swismodel GMQE score, the identity between the target/template alignment, the TM-score between the contact-based model and the homology model

https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%209.png

Figure 9 **Comparison between the contact-based model and homology modelling.** The superposition between the contact-based model (red) and the homology model (blue) and respective TM-score.

**SPW family (PF03779)**

According to the PFAM database [[27]](https://paperpile.com/c/NMPUUL/Udag), the SPW family is present in Bacteria and Archaea, and each protein consists of one or two repeat units. Furthermore, some members also contain an additional domain, either a Vitamin K epoxide reductase (PF07884) or an NAD-dependent epimerase/dehydratase (PF01370). Each repeat unit is formed by two transmembrane alpha-helices and is characterised by an SPW motif [[28]](https://paperpile.com/c/NMPUUL/mub6). According to our model, the repeated motif is buried in the membrane symmetrically located close to the extracellular side, Figure *10*. PFAM architectures show many proteins with only a single SPW motif however a more careful analysis of these sequences shows that in many cases they contain a second degenerate SPW unit with the proline residue conserved (Figure S3).

The Tryptophan is on the outer side of the protein, facing the bilayer, while the proline is on the inner side of the protein, promoting the formation of a kink in the transmembrane helix [[29]](https://paperpile.com/c/NMPUUL/J0Ig). The protein contains a ser-pro motif, which is rare among TM-proteins and most likely increases the bending effect of proline significantly due to their hydrogen bond pattern [[30]](https://paperpile.com/c/NMPUUL/fMtW).

# **DUF2963 (PF11178)**

DUF2963 is a protein family of unknown function present only in Mollicutes bacteria [[27]](https://paperpile.com/c/NMPUUL/Udag). Our model results in a Class III.5 “anti-parallel β layer / β hairpins” protein, Figure 10.

**Curlin repeats family (PF07012)**

The curlin repeats family is the only family where PconsC4 outperforms DeepMetaPsicov. Here, the PconsC4 model has a higher predicted TM score (Supplementary Table1) and agrees better with information from the available literature [[31]](https://paperpile.com/c/NMPUUL/NMtO). Curlin is predicted to have a β-solenoid structure, see Figure 10. DeBenedictis et al. presented ab-initio models for two members of the Curlin repeat family, CsgA and CsgB [[31]](https://paperpile.com/c/NMPUUL/NMtO). The structure of their best models is visually in agreement with our model (a direct comparison is difficult as the coordinates are not available for their models). Our model is also in agreement with the partial structure of the repeat units of CsgA published by Perov et al. [[32]](https://paperpile.com/c/NMPUUL/gXnj). This model contains two parallel β-sheets with individual units situated perpendicular to the fibril axis (corresponding PDB IDs are 6G8C, 6G8D, 6G8E).

### Additional families

Next, we examined the models with predicted TM-score lower than 0.55 manually. Among them, we found the representative of the UCH-protein repeats (PFAM family *PF13446)*  interesting despite a predicted TM-score of only 0.41. The low score is due to the difficulties in predicting the inter-unit fold while at the unit level, the models appear to be quite reliable. Our model suggests that this repeat region is a member of Class V.1 ɑ-beads, with four helical domains separated by a flexible linker.

UCH-protein repeats family is a repeat domain found in Ubiquitin carboxyl-terminal hydrolase. Despite UCH-proteins being widespread among eukaryotes, the repeated domain is present only in yeasts in a variable number of units. According to PFAM [[27]](https://paperpile.com/c/NMPUUL/Udag), the UCH-protein repeats could be involved in the formation of the complex of UCH with Rsp5 and Rup1.

<https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure%2010.png>

Figure 10 **Well model without homologous proteins.** The three high scoring models for SPW, DUF2963 and Curlin and the medium scoring UCH-protein are shown, with the different protein units are coloured in red and blue.

**Conclusion.**

<https://github.com/ElofssonLab/Claudio/blob/master/Repeats/Figures/Figure11.png>

***Figure 11. Datasets comparisons.*** *a) Neff scores comparison between the two datasets. b) The variation in the membership to the three domains of life between the PFAM families of the “Unresolved Proteins Dataset” and the “PDB dataset”.*

Here, we performed a comprehensive coevolution analysis on repeat protein families, and we show that Deep Meta Psicov contact-predictions method overcomes the traditional difficulties of DCA methods for this class of proteins. We investigated the modelling of repeat units, and we developed a novel quality assessment method for repats proteins. Finally, we tested the pipeline on PFAM families without protein structures showing its usefulness in providing new structural information.The modelling of the unknown PFAM families is challenging, but 22% of the datasets had a predicted TM-score equal or higher to 0.55; compared with 56% of the benchmark dataset. However, the families in the “Unknown protein families” are significantly smaller, Figure 11a. Moreover, in the unknown set, there are more eukaryotic-specific protein families, see Figure 11b, indicating that it might be possible to model these when more eukaryotic genomes become available.

Despite the significant improvement brought by deep-learning in contact prediction, there is still room for improvement. The accuracy of inter-domain contacts is often lower than for the intra-units one, and the development and future development could focus on that. Furthermore, the folding part of the pipeline is a limiting step, in particular for longer proteins where it seems not to generate compact structures.

**Materials and Methods**

**Datasets generation**

The repeat protein dataset was generated starting from the 3585 reviewed entries in RepeatsDB [[14,33]](https://paperpile.com/c/NMPUUL/z3tG+unQK),. The proteins of class I and II were removed, and then the dataset was homology reduced using CD-HIT [[34]](https://paperpile.com/c/NMPUUL/WPGJ) at 40% identity resulting in 819 repeat regions. From this “complete region dataset” two other datasets were generated. First, a “single unit” dataset with one repeat unit from each family, and secondly a “double unit” dataset with two. In the two derived datasets, the representative units were selected, avoiding or at least minimising, the presence of insertions.

The non-resolved repeats protein family dataset was generated, collecting all the repeat proteins families with missing structural information present in PFAM [[27]](https://paperpile.com/c/NMPUUL/Udag) as of May 2019 and removing domains with a significant overlap with the disorder prediction. It results in 49 protein families. The representative sequence for each family of repeat was chosen for matching these criteria: 1) select the most common architecture; 2) Include when possible at least three repeat units.

**Multiple sequence alignment (MSA)**

The multiple sequence alignments (MSA) were carried out using HHblits [[35]](https://paperpile.com/c/NMPUUL/wdaQ) using an E-value cutoff of 0.001 against the Uniclust30\_2017\_04 database [[36]](https://paperpile.com/c/NMPUUL/Orcg). The number of effective sequences of the alignment, expressed as Neff-score, was calculated by HHblits and used for subsequent analysis. More detail about the Neff calculation can be found at <https://github.com/soedinglab/hh-suite/wiki> [[35]](https://paperpile.com/c/NMPUUL/wdaQ)

**Contact prediction and models generation**

The protein models were generated following the PconsFold2 protocol [[37]](https://paperpile.com/c/NMPUUL/OKfq). The secondary structure of the repeat regions was predicted by PSIpred [[38]](https://paperpile.com/c/NMPUUL/GaV2). Protein contacts were predicted using DeepMetaPsicov [[39]](https://paperpile.com/c/NMPUUL/rXIQ), or PconsC4[[18]](https://paperpile.com/c/NMPUUL/kw3D) and together with the secondary structure predictions used as input to Confold [[23]](https://paperpile.com/c/NMPUUL/GtDi). The modelling used the top scoring 1.5 L contacts (where L is the length of the modelled regions).

**Contacts analysis**

A protein contact was defined as two residues having a beta carbon distance equal to or lower than 8Å in the PDB structure and farther than five residues in the sequence. Using this definition, we assess the number of correctly predicted contacts (the Positively Predicted Value (PPV)) taking into account the top-scoring 1.5 L contacts.

In the intra/inter-unit contacts analysis, the predicted contacts of each protein were divided into i) intra-unit contacts, if between residues inside the same unit; ii) inter-units if the residues are in different repeat units. The units mapping was taken from the RepeatsDB database [[14]](https://paperpile.com/c/NMPUUL/z3tG). In this analysis, we calculate the number of intra- and inter-unit contacts in the PDB structure, and then we selected the same number of predicted intra- and inter-units contacts. The PPV was then calculated as the fraction of correct predictions.

**Template search and homology modelling**

The template search and the homology models were generated from the representative sequences using the default options from Swissmodel [[26]](https://paperpile.com/c/NMPUUL/yDw0).

**Protein models analysis**

The model quality, expressed in TM-score, was assessed using a random forest regression model using the python module Sklearn. The random forest regression was optimized to include 240 estimators and a maximum depth of 60. The “complete region” benchmark set was used as a training set. The label of the training set was the TM-score of each model [[40]](https://paperpile.com/c/NMPUUL/8si2). To ensure that the protein structure and the model were aligned correctly, the TMalign option -I was used, providing a local alignment of the two sequences.

For training, five cross-validation sets were generated. Several inputs were used for the random forest, described briefly below and in Supplementary Table 1. The Confold and QmeanDisco inputs were obtained from analysing the first ranked model.Pcons was run using the option -d using all the models in the stage2 folder generated by Confold. Among the different set of features tried, we select eight features that all improve the prediction of the random forest regression, see Supplementary Table 1.

## **Data availability**

All the protein models, contact prediction, and Multiple Sequence alignments are available at 10.6084/m9.figshare.9995618

## Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 823886. AE and OS are funded by grants from the Swedish Natural Science Research Council No VR-NT 2016-03798.

**Bibliography**

1. [Heringa J. Detection of internal repeats: how common are they? Curr Opin Struct Biol. 1998;8: 338–345.](http://paperpile.com/b/NMPUUL/7BHY)

2. [Strand M, Prolla TA, Liskay RM, Petes TD. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature. 1993;365: 274–276.](http://paperpile.com/b/NMPUUL/Gkvz)

3. [Pâques F, Leung W-Y, Haber JE. Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. Molecular and Cellular Biology. 1998. pp. 2045–2054. doi:](http://paperpile.com/b/NMPUUL/Ind2)[10.1128/mcb.18.4.2045](http://dx.doi.org/10.1128/mcb.18.4.2045)

4. [Schaper E, Gascuel O, Anisimova M. Deep conservation of human protein tandem repeats within the eukaryotes. Mol Biol Evol. 2014;31: 1132–1148.](http://paperpile.com/b/NMPUUL/SgoR)

5. [E.M. Marcotte, M. Pellegrini, T.O. Yeates, D. Eisenberg. A census of protein repeats. J Mol Biol. 1999;293: 151–160.](http://paperpile.com/b/NMPUUL/RDz1)

6. [Björklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. PLoS Comput Biol. 2006;2: e114.](http://paperpile.com/b/NMPUUL/2HFJ)

7. [Andrade MA, Perez-Iratxeta C, Ponting CP. Protein Repeats: Structures, Functions, and Evolution. Journal of Structural Biology. 2001. pp. 117–131. doi:](http://paperpile.com/b/NMPUUL/5Fjy)[10.1006/jsbi.2001.4392](http://dx.doi.org/10.1006/jsbi.2001.4392)

8. [Stirnimann CU, Petsalaki E, Russell RB, Müller CW. WD40 proteins propel cellular networks. Trends Biochem Sci. 2010;35: 565–574.](http://paperpile.com/b/NMPUUL/nKVc)

9. [Li J, Mahajan A, Tsai M-D. Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry. 2006;45: 15168–15178.](http://paperpile.com/b/NMPUUL/HLpY)

10. [Mosavi LK, Cammett TJ, Desrosiers DC, Peng Z-Y. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci. 2004;13: 1435–1448.](http://paperpile.com/b/NMPUUL/WWdy)

11. [Persi E, Wolf YI, Koonin EV. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. Nat Commun. 2016;7: 13570.](http://paperpile.com/b/NMPUUL/W73P)

12. [Kajava AV. Review: Proteins with Repeated Sequence—Structural Prediction and Modeling. Journal of Structural Biology. 2001. pp. 132–144. doi:](http://paperpile.com/b/NMPUUL/AdXS)[10.1006/jsbi.2000.4328](http://dx.doi.org/10.1006/jsbi.2000.4328)

13. [Kajava AV. Tandem repeats in proteins: From sequence to structure. Journal of Structural Biology. 2012. pp. 279–288. doi:](http://paperpile.com/b/NMPUUL/lEFg)[10.1016/j.jsb.2011.08.009](http://dx.doi.org/10.1016/j.jsb.2011.08.009)

14. [Paladin L, Hirsh L, Piovesan D, Andrade-Navarro MA, Kajava AV, Tosatto SCE. RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. Nucleic Acids Res. 2017;45: 3613.](http://paperpile.com/b/NMPUUL/z3tG)

15. [Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins. 2018;86 Suppl 1: 97–112.](http://paperpile.com/b/NMPUUL/27Hch)

16. [Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 1997;271: 511–523.](http://paperpile.com/b/NMPUUL/gK7oZ)

17. [Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. Capturing coevolutionary signals inrepeat proteins. BMC Bioinformatics. 2015;16: 207.](http://paperpile.com/b/NMPUUL/4uCf)

18. [Michel M, Hurtado DM, Elofsson A. PconsC4: fast, accurate, and hassle-free contact predictions. Bioinformatics. 2018. doi:](http://paperpile.com/b/NMPUUL/kw3D)[10.1093/bioinformatics/bty1036](http://dx.doi.org/10.1093/bioinformatics/bty1036)

19. [Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun. 2019;10: 3977.](http://paperpile.com/b/NMPUUL/AtyZ)

20. [Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, et al. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS One. 2014;9: e92721.](http://paperpile.com/b/NMPUUL/LRiE)

21. [Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci. 2001;10: 2354–2362.](http://paperpile.com/b/NMPUUL/hxH5)

22. [Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol. 2014;10: e1003889.](http://paperpile.com/b/NMPUUL/V7Po)

23. [Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins. 2015;83: 1436–1449.](http://paperpile.com/b/NMPUUL/GtDi)

24. [Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics. 2020;36: 2647.](http://paperpile.com/b/NMPUUL/lax1)

25. [Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42: W252–8.](http://paperpile.com/b/NMPUUL/3xrz)

26. [Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46: W296–W303.](http://paperpile.com/b/NMPUUL/yDw0)

27. [El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432.](http://paperpile.com/b/NMPUUL/Udag)

28. [Yeats C, Bentley S, Bateman A. New knowledge from old: in silico discovery of novel protein domains in Streptomyces coelicolor. BMC Microbiol. 2003;3: 3.](http://paperpile.com/b/NMPUUL/mub6)

29. [von Heijne G. Proline kinks in transmembrane alpha-helices. J Mol Biol. 1991;218: 499–503.](http://paperpile.com/b/NMPUUL/J0Ig)

30. [Deupi X, Olivella M, Govaerts C, Ballesteros JA, Campillo M, Pardo L. Ser and Thr Residues Modulate the Conformation of Pro-Kinked Transmembrane α-Helices. Biophysical Journal. 2004. pp. 105–115. doi:](http://paperpile.com/b/NMPUUL/fMtW)[10.1016/s0006-3495(04)74088-6](http://dx.doi.org/10.1016/s0006-3495(04)74088-6)

31. [DeBenedictis EP, Ma D, Keten S. Structural predictions for curli amyloid fibril subunits CsgA and CsgB. RSC Adv. 2017;7: 48102–48112.](http://paperpile.com/b/NMPUUL/NMtO)

32. [Perov S, Lidor O, Salinas N, Golan N, Tayeb-Fligelman E, Deshmukh M, et al. Structural Insights into Curli CsgA Cross-β Fibril Architecture Inspired Repurposing of Anti-amyloid Compounds as Anti-biofilm Agents. doi:](http://paperpile.com/b/NMPUUL/gXnj)[10.1101/493668](http://dx.doi.org/10.1101/493668)

33. [Hirsh L, Paladin L, Piovesan D, Tosatto SCE. RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins. Nucleic Acids Res. 2018;46: W402–W407.](http://paperpile.com/b/NMPUUL/unQK)

34. [Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006. pp. 1658–1659. doi:](http://paperpile.com/b/NMPUUL/WPGJ)[10.1093/bioinformatics/btl158](http://dx.doi.org/10.1093/bioinformatics/btl158)

35. [Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2012. pp. 173–175. doi:](http://paperpile.com/b/NMPUUL/wdaQ)[10.1038/nmeth.1818](http://dx.doi.org/10.1038/nmeth.1818)

36. [Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45: D170–D176.](http://paperpile.com/b/NMPUUL/Orcg)

37. [Bassot C, Menendez Hurtado D, Elofsson A. Using PconsC4 and PconsFold2 to Predict Protein Structure. Curr Protoc Bioinformatics. 2019; e75.](http://paperpile.com/b/NMPUUL/OKfq)

38. [McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000. pp. 404–405. doi:](http://paperpile.com/b/NMPUUL/GaV2)[10.1093/bioinformatics/16.4.404](http://dx.doi.org/10.1093/bioinformatics/16.4.404)

39. [Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins. 2019;87: 1092–1099.](http://paperpile.com/b/NMPUUL/rXIQ)

40. [Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309.](http://paperpile.com/b/NMPUUL/8si2)

**Supporting Information**

**Table S1. Unknown protein family dataset.** In the columns are reported respectively: the UniProt ID of the modelled sequence, the PFAM family, the predicted TM score by random forest regression.

**Figure S1 Performance expressed in PPV for the set of proteins contained in the training set of the contact predictions methods.** In light green the Deep Meta Psicov (DMP) average precision for the proteins of the benchmark set non overlapping with the DMP training set, in dark green the DMP average precision for the proteins benchmark set present in the DMP training set, in light blue the PconsC4 average precision for the proteins of the benchmark set non overlapping with the PconsC4 training set, in dark blue the PconsC4 average precision for the proteins benchmark set present as well in the PconsC4 training set.

**Figure S2 Amino Acid frequency of the single domain architecture sequences**. From the logo is possible recognize two SPW domains, one of them degenerated (in particular the first Serine in the second motif) that is not recognized by PFAM.