

# Accurate contact predictions for thousands of protein families using PconsC3.

Marcin J. Skwark<sup>a,1</sup>, Mirco Michel<sup>b,1</sup>, David Menéndez Hurtado<sup>b</sup>, Magnus Ekeberg<sup>c</sup>, and Arne Elofsson<sup>b,2</sup>

<sup>a</sup>Department of Chemistry, Vanderbilt University, Nashville, TN, USA; <sup>b</sup>Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Stockholm 10691, Sweden; <sup>c</sup>Royal Institute of Technology, Stockholm 10691, Sweden

This manuscript was compiled on October 7, 2016

**Protein structure prediction was for decades one of the grand unsolved challenges in bioinformatics. A few years ago it was shown that by using a maximum entropy approach to describe couplings between columns in a multiple sequence alignment it was possible to significantly increase the accuracy of residue contact predictions. For very large protein families with more than 1000 effective sequences the accuracy is sufficient to produce accurate models of proteins as well as complexes. Today, for about half of all Pfam domain families no structure is known, but unfortunately most of these families have at most a few hundred members, i.e. are too small for existing contact prediction methods. To extend accurate contact predictions to the thousands of smaller protein families we present PconsC3, an improved method for protein contact predictions that can be used for families with as little as 100 effective sequence members. We estimate that PconsC3 provides accurate contact predictions for up to 4646 Pfam domain families. In addition, PconsC3 outperforms previous methods significantly independent on family size, secondary structure content, contact range, or the number of selected contacts. This improvement translates into improved de-novo prediction of three-dimensional structures. PconsC3 is available as a web server and downloadable version at <http://c3.pcons.net>. The downloadable version is free for all to use and licensed under the GNU General Public License, version 2.**

protein structure prediction | contact prediction | direct coupling analysis  
| protein family

In recent years great progress has been made in the area of residue contact prediction. The vast amount of available sequence data is utilized by direct coupling analysis (DCA) methods to predict contacts between residues with unprecedented quality [1, 2]. This has enabled accurate blind predictions of the structure of soluble proteins [3–5], membrane proteins [6–8], and protein complexes [9, 10]. However, the widespread use of such methods has been limited to protein families with more than 1000 members [11, 12]. Unfortunately, the structure of at least one member of most large families is known (see Fig. S1). This limits the practical usefulness of DCA methods [13] and strongly suggests that methods that accurately predict residue contacts for smaller protein families would be of much greater utility.

Before the advent of DCA methods there has been a long-standing effort in using machine learning techniques to predict residue contacts [14–16]. These methods utilize covariance-based evolutionary information (e.g. mutual information), as well as knowledge based constraints as inputs to a machine learning algorithm. The best non-DCA methods are less dependent on the size of the protein family and although their predictive quality is easily outperformed by DCA on large families, they perform significantly better on smaller families, Figure 1 (a). We have earlier used an iterative ma-

chine learning approach, built on the observation that contacts are not randomly distributed, to improve the performance of DCA based contact prediction methods when we developed PconsC2 [17]. Here we propose a way to substantially improve the predictive power, by including state-of-the-art non-DCA predictors among the method's inputs.

For about half (53%) of the protein families in the Pfam database [18] no structure that covers most of the length can be found in the protein data bank (PDB) [19] (Figure 1b). The distribution of family sizes in the Pfam database shows that the median size of families with known structure (680 effective sequences) is significantly (rank sum p-value  $< 2.2 \times 10^{-16}$ ) larger than that of families without a known structure (134). The number of potential target families (sufficiently many members, but without known structure) would increase more than three-fold from 1528 to 4973 if accurate predictions could be made from a family with 100 effective sequences instead of 1000 (Supplementary Fig. S1).

PconsC3 combines two DCA methods with contacts predicted using a non-DCA machine learning approach. PconsC3 utilizes the iterative pattern recognition approach introduced in PconsC2 [17]. We find that PconsC3 significantly outperforms earlier methods independent on protein family size, Figure 1 (a), as well as it yields better structural models. The increased accuracy for small proteins leads to a substantial increase in the number of potential targets, from 12% of all Pfam families with unknown structure when using the best DCA

## Significance Statement

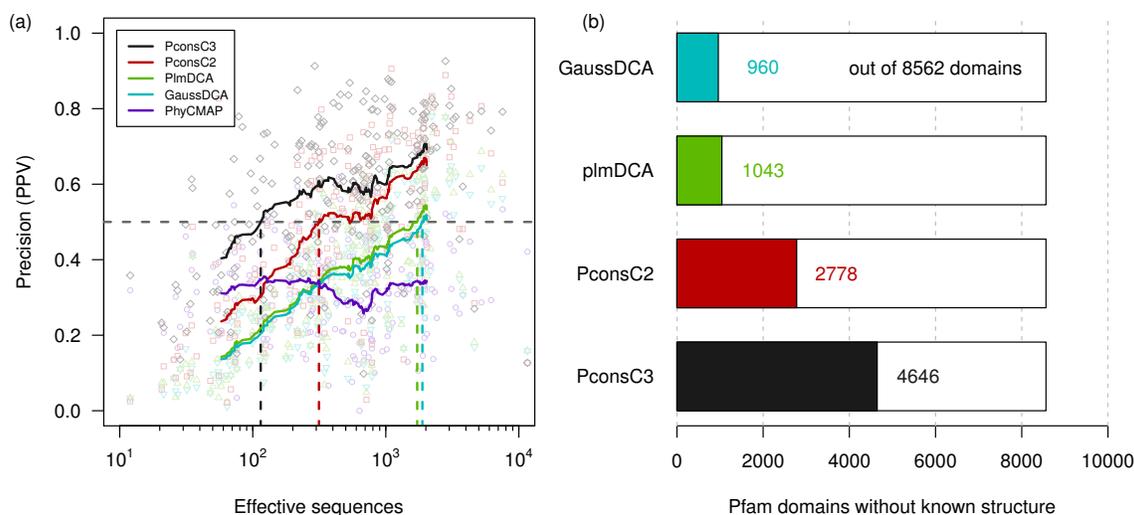
One of the longest standing challenges in structural bioinformatics is the protein structure prediction problem, i.e. to predict a protein structure from its sequence. It has recently been shown that, by using accurate residue contact prediction it is possible to predict the structure of many proteins with high accuracy. However, all direct-coupling based contact prediction methods introduced in the last few years require thousands of protein sequences for accurate predictions, but many protein families are smaller. Here, we introduce PconsC3, a method that accurately predicts contacts for families with as few as 100 effective members. When we apply PconsC3 to Pfam we estimate that for 4646 (54%) families of unknown structure we have sufficient coverage for accurate contact predictions.

Author contributions: PconsC3 was mainly developed by M.J.S. with help from all the other authors. M.M. primarily did analysis with input from A.E., D.M.-H. and M.J.S. M.M. and A.E. wrote the manuscript with input from the other authors.

The authors declare no conflict of interest.

<sup>1</sup> M.J.S and M.M contributed equally to this work.

<sup>2</sup> To whom correspondence should be addressed. E-mail: arnebioinfo.se



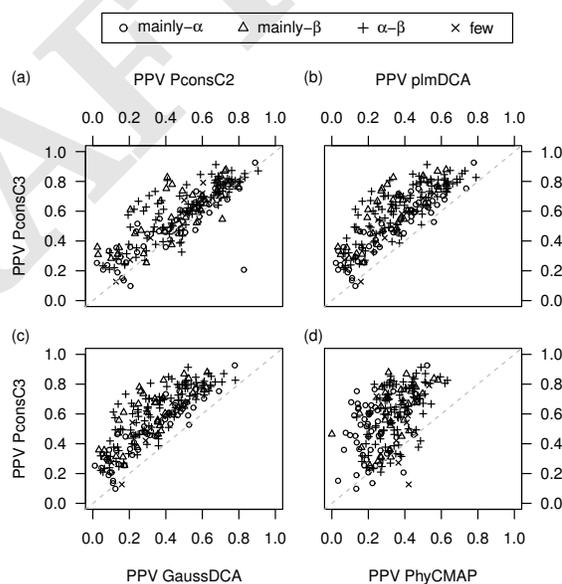
**Fig. 1.** (a) Contact predictor performance on the benchmark dataset measured in Positive Predictive Value (PPV or precision). Performance of the top  $N/2$  ranked contacts against protein family size measured in effective sequences, where  $N$  denotes the number of contacts observed in the native structure. A native contact is defined as a pair of  $C_{\beta}$ -atoms within a spatial distance of 8Å. The horizontal dashed line marks a precision of 0.5. The vertical lines illustrate at which family size each method reaches this threshold on average. (b) Number of Pfam 29.0 families with unknown structure. A family is defined to have a known structure if there is a significant hit to an entry in PDB that covers more than 75% of the sequence length of the family. In color are shown the numbers of potential target families for each method. These families have at least the number of effective sequences at which the corresponding method reaches an average PPV of 0.5 on our benchmark dataset.

method to 54% using PconsC3 instead. These predictions are publically available at <http://c3.pcons.net>.

## 1. Results and Discussion

**Improvement over all protein family sizes.** The precision of both DCA methods, plmDCA [20] and GaussDCA [21] as well as that of PconsC2 [17] is strongly dependent on family size. The average Precision (PPV) for PconsC2 for  $N/2$  ( $N$  being the number of native contacts) increases from 0.3 to 0.56 when the average effective family size increases from 100 to 1000 sequences. In contrast, the performance of PhyCMAP [15] is approximately 0.3 for families with between 50 and 2000 effective sequences, Fig. 1 (a). When including PhyCMAP as well as other improvements (see methods) into PconsC3 the performance increases significantly for small families. The average PPV for a 100 effective sequence protein family is 0.47, and increases to 0.60 for a 1000 member family. We have noted that on average a PPV of 0.5 is needed for accurate modeling using the PconsFold [22] pipeline. This average precision is never reached for PhyCMAP, for plmDCA and GaussDCA more than 1700 effective sequences are needed, for PconsC2 314 and for PconsC3 only 115. Even below 100 effective sequences 23% of the benchmark proteins have a PPV larger than 0.5 when using PconsC3.

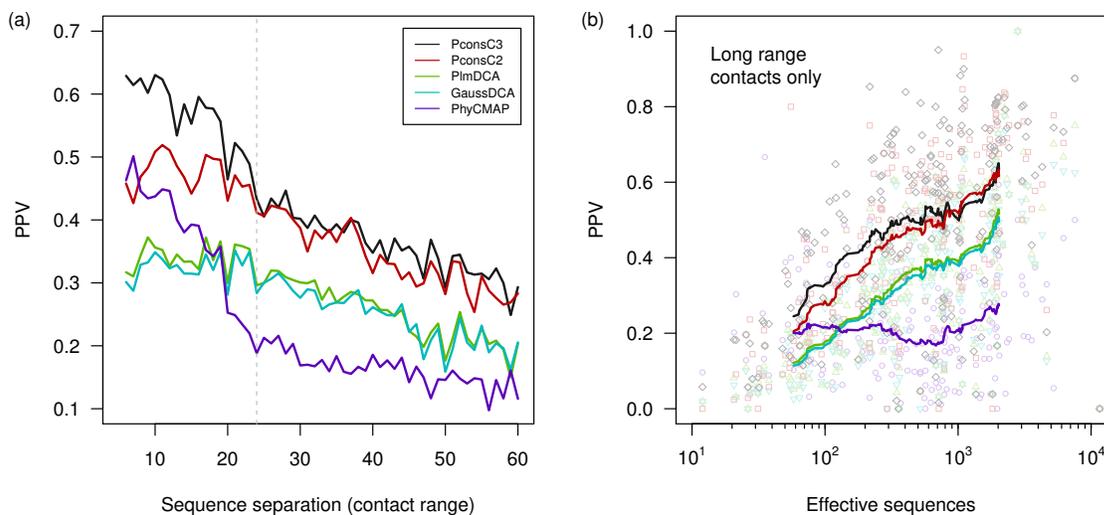
There are 16295 domains in Pfam 29.0 out of which 8562 do not have a significant match to PDB covering most of the domain length. This means 7733 Pfam domains have at least one representative in PDB and could thus be modeled by homology modelling. If we apply the measurements from Fig. 1 (a) there would be 1043 Pfam domains of unknown structure that could potentially be predicted by the best DCA method. This means these many domains have more than 1700 effective sequences in their alignment. Lowering the threshold of alignment size leads to an increase in the number



**Fig. 2.** Direct performance comparison between PconsC3 and other methods on the benchmark dataset. Proteins were assigned secondary structural classes based on their ECOD architecture assignment. Symbols represent the class of a protein.

of potential target domains. With PconsC2 2778 domains could be predicted. This number increases to 4646 when using a method such as PconsC3 that is able to accurately predict contacts from even smaller alignments.

**Performance by type of secondary structure.** Figure 2 shows a direct comparison between PconsC3 and other contact predictors. PconsC3 outperforms DCA methods on 207 proteins and PhyCMAP on 195 proteins independent on the type of sec-



**Fig. 3.** (a) PPV on the top  $N/2$  contacts at a specific sequence separation (number of residues between those participating in a contact). A minimum sequence separation of five residues was used to filter out local interactions of neighboring residues or helices. Long-range contacts have a separation of at least 24 residues (everything to the right of the dashed line). (b) Long range contact predictor performance in PPV against protein family size measured in effective sequences.

**Table 1. Average PPV of top  $N/2$  predicted contacts on the benchmark dataset for different secondary structural classes.**

	all	mainly- $\alpha$	mainly- $\beta$	$\alpha - \beta$
PconsC3	0.57	0.49	0.59	0.62
PconsC2	0.48	0.44	0.45	0.51
plmDCA	0.36	0.34	0.32	0.38
GaussDCA	0.34	0.33	0.31	0.36
PhyCMAP	0.32	0.23	0.34	0.36
counts	210	55	35	110

ondary structure. Compared to PconsC2, PconsC3 performs better in 166 out of 210 proteins of the benchmark dataset. Some of the largest improvements are made for  $\alpha - \beta$  (cross in Fig. 2) and mainly- $\beta$  proteins (triangle), whereas PconsC2 performs exceptionally well for one short  $\alpha$ -helical protein (PDB: 1ediA).

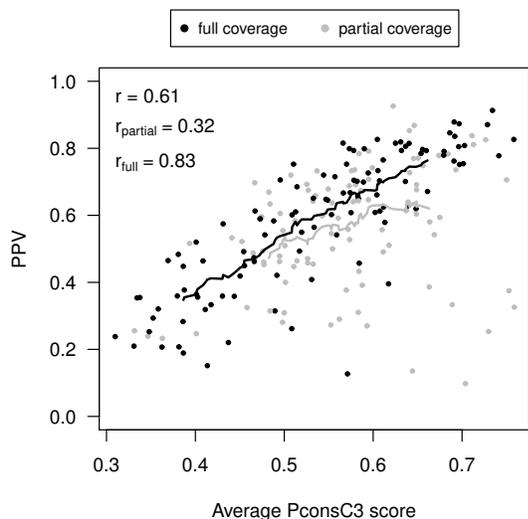
Table 1 shows the performance of contact predictors on different types of secondary structure. The first column lists performance on all proteins of the test set. Overall PconsC3 performs best for all classes independent of rank (Supplementary Fig. S2). On average it predicts more than 57% of  $N/2$  contacts correctly, compared to 48% for PconsC2, 36% for the best DCA method, and 32% for PhyCMAP. The improvement is largest on mainly- $\beta$  proteins with a 31% increase in PPV of PconsC3 over PconsC2 and 84% over plmDCA. This can be attributed mostly to PhyCMAP performing better than DCA in this particular class. In the  $\alpha - \beta$  class PhyCMAP is worse than DCA, while for mainly- $\alpha$  proteins both DCA methods clearly outperform PhyCMAP. Within the DCA methods plmDCA always performs better than GaussDCA. Although shown in Fig. 2, the class of few secondary structural elements was omitted in the table due to a small sample size (see Methods).

**Predicting long range contacts.** structure prediction [23]. There is a striking difference between PhyCMAP and DCA

based methods for long-range contacts. PhyCMAP predicts short to medium ranged contacts (with a sequence separation from 5 up to 23 residues) with higher quality than long-range contacts (Fig. 3a). For short-range contacts (up to 12 residues separation) PhyCMAP is actually on par with PconsC2 and significantly better than DCA methods, while it is significantly worse for long-range contacts. Although PconsC3 outperforms DCA methods independently of the sequence separation of contacting residues (Fig. 3a), it is clear that the increase in precision of PconsC3 in relation to PconsC2 and DCA methods is gradually larger for shorter contact ranges, suggesting that it benefits from the good performance of PhyCMAP in this range. On long-range contacts PconsC3 performs best for smaller protein families (Fig. 3b). This and the fact that the gap between PconsC2 and DCA methods decreases for long-range contacts in small families indicate the value of including a non-DCA method to PconsC3.

**Estimation of contact map quality.** The average PconsC3 contact scores can be used as a good indicator for contact map quality. Figure 4 (a) shows that the average contact score of the top ranked contacts has a Pearson correlation  $r$  of 0.61 against PPV. However, we noted the test dataset also includes alleged multi-domain proteins, i.e. proteins where most of the sequences in the alignments does not cover the entire domain, such as 2csmA and 2ejnA, Supplementary Fig. S3 (a and b). Most of the proteins with high average contact score and low PPV fall into that category (gray dots in the lower right region of Figure 4 (a)). This leads to the assumption that PconsC3 is overestimating the predictions in such cases. When ensuring proteins are mostly covered by at least half of all sequences (black dots)  $r$  increases to 0.83 ( $r_{covered}$ ) showing that the average PconsC3 score is an excellent estimator of the contact map quality in single domain proteins.

**Structure prediction.** The more accurate contact maps of PconsC3 improve structure prediction, confirming earlier ob-

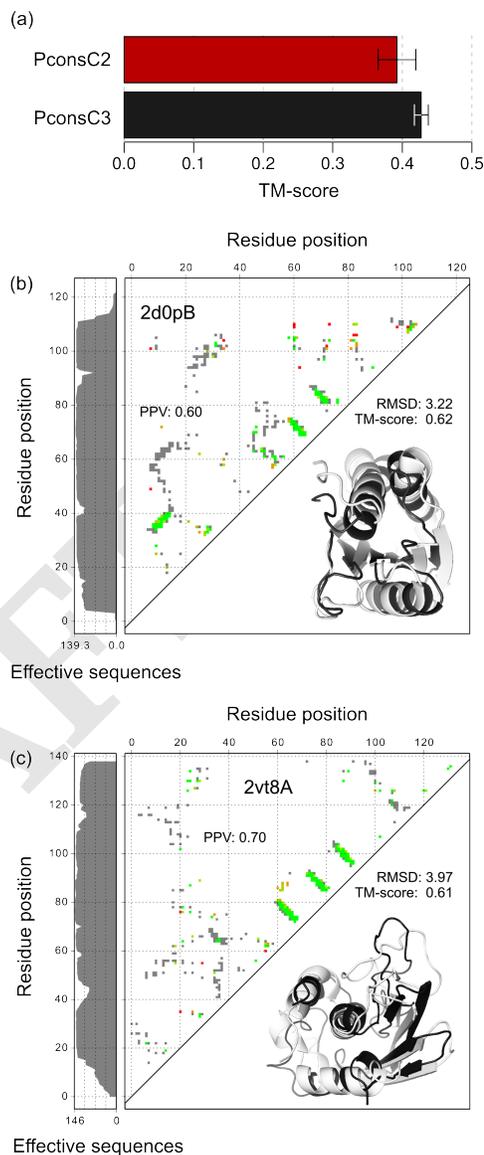


**Fig. 4.** PconsC3 score as an estimator for PPV. Pearson correlation coefficient is denoted  $r$  and  $r_{full}$  of all proteins and those with at least 80% of its residues covered by more than 50% of all sequences in the family alignment (length coverage), respectively. The dashed and solid red lines indicate a moving average with window size of 50 for all proteins and those with high length coverage, respectively.

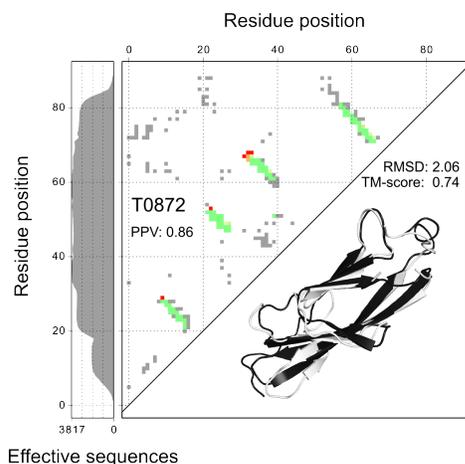
servations [22], Fig. 5 (a). The small improvement in average TM-score [24] when using PconsC3 contacts is significant (t-test p-value  $< 1.5 \cdot 10^{-2}$ ) and independent of the family size of the target protein. However, there are still many proteins with large families and supposedly good contact maps, for which PconsFold fails to converge properly (Supplementary Fig. S4). Using a more elaborate folding protocol [25] or generating more than 2000 Rosetta decoys might improve this situation. Anyhow, when using predicted contacts from PconsC3 the number of proteins with a TM-score of 0.5 or higher, meaning the fold has most likely been correctly identified, increases from 55 to 75.

The main advantage of PconsC3 over PconsC2 and DCA methods is that it can accurately predict the contacts for smaller protein families. The Diol dehydratase reactivase ATPase-like domain (PF08841) only contains 139 effective sequences but both the contact map and the model are in excellent agreement with the native structure (2d0pB), Fig. 5(b). The TM-score of the model is 0.61 while a model based on PconsC2 only has a TM-score of 0.40. The PI31 proteasome regulator N-terminal (PF11566) has 146 effective sequences and for this protein a TM-score of 0.61 is reached, Fig. 5(c).

**Blind prediction of T0872 in CASP12.** The submission phase for the twelfth Critical Assessment of Techniques for Protein Structure Prediction (CASP12) recently finished and the official evaluations are running as of writing this manuscript. Initial evaluation of CASP12 targets that can already be found in PDB revealed that the combination of PconsC3 and PconsFold successfully predicted the structure of Target T0872. Model 3 (Pcons-net\_TS3) has the highest TM-score of 0.74 followed by Pcons-net\_TS1. All five Pcons-net submissions rank among the top 10 models for this target. Figure 6 shows the predicted contact map as well as the structure of Model 3, both overlaid on the native contacts (gray) and structure



**Fig. 5.** Structure prediction. (a) TM-score of PconsFold when using PconsC2 (red) or PconsC3 (black) contact predictions. (b) Contact map for Diol dehydratase reactivase ATPase-like domain (Pfam: PF08841, PDB: 2d0pB) in the upper left triangle. Grey dots indicate native contacts in the PDB structure, green are correctly predicted contacts while yellow to red are false positive predictions. Left the sequence coverage measured in effective sequences. The lower right triangle shows the structure predicted with PconsFold using PconsC3 contacts (black) superimposed onto the native structure from PDB (light gray) (c) Contact map and structure for the PI31 proteasome regulator N-terminal (PF11566, 2vt8A).



**Fig. 6.** Blind prediction of target T0872 in CASP12. Predicted contact map on top of the native in the upper left triangle. The lower right triangle shows the structure predicted with PconsFold using PconsC3 contacts (black) superimposed onto the native structure from PDB (light gray).

(white), respectively. The high precision of the predicted contacts can be attributed to the large sequence coverage of 3817 effective sequences in the alignment. Furthermore, PconsFold was able to converge to a near native structure most likely due to the small size of the protein. For some larger targets this was not the case.

## Materials and Methods

**Datasets.** PconsC3 has been trained on a set of 180 protein families (supplementary Table S1). This training set comprises of 150 protein families from the original PSICOV dataset [26] plus 30 additional families with a small number of members from the test set as described in [17].

All evaluation has been made on a dataset of 210 (supplementary Table S2) proteins without any homology to any protein in the training set. All PDB IDs were matched against the ECOD [27] domain assignment from 2016-03-28. This set was obtained from the set used in the development of PconsC2 [17] and homology reduced such that no protein included in the test set shared an ECOD H-class with any of the proteins in the new training dataset. This homology reduction is much more stringent than using sequence information alone. The final list of proteins used as well as their ECOD H-class number and number of effective sequences are found in Tables S1 and S2. Alignments were created using HHblits [28] version 2.0.15 on the uniprot20 database bundled with HHSuite (date: 2016-02-26) with an e-value of 1. In order for HHblits to output and align all sequences the parameter `-all` has been used and `-maxfilt` and `-realign_max` were both set to 999999. These alignments were used as input for the DCA methods.

For the evaluation of Pfam domains the HHSuite database of Pfam 29.0 (date: 2016-05-03) was used to scan Uniprot at an e-value threshold of 1 using HHblits. The resulting set of alignments was then analyzed for effective number of sequences. For each domain the sequence that was highest ranked by HHblits has been defined as the domain representative. The length of a domain has been set to the length of its representative sequence. HHsearch version 3.0.0 was used to scan each family against the HHSuite database of protein data bank (PDB) sequences (date: 2016-03-02) to determine whether a given Pfam family has known or unknown structure. A hit has been considered significant if its E-value was below  $10^{-3}$  and if it covered at least 75% of the length of the family.

**Secondary structural classes.** To classify the dataset into the secondary structural classes mainly- $\alpha$ , mainly- $\beta$ , and  $\alpha - \beta$ , we used

the architecture assignment of ECOD for the PDB IDs of our benchmark set. ECOD uses a scheme with seven structural classes that we mapped into three in order to increase sample size and thus statistical significance of each class. The following mapping was applied:  $\alpha/\beta$ ,  $\alpha - \beta$  and  $\alpha + \beta$  to  $\alpha - \beta$ ;  $\alpha$  to mainly- $\alpha$ ;  $\beta$  and *extended* to mainly- $\beta$ . The secondary structural class *few* was omitted as it only contained 10 proteins. Supplemental Figure S3 shows a table analogous to 1, but with the original ECOD classification (including *few*).

**Contact prediction.** Julia implementations have been used for both plmDCA and GaussDCA, which are available on GitHub at <https://github.com/pagnani/plmDCA> and <https://github.com/carlobaldassi/GaussDCA.jl>, respectively. Both require Julia 0.3 or higher. PhyCMAP was obtained at <http://raptorx.uchicago.edu/download/>. Regularization strength of plmDCA was set to 0.02. GaussDCA and PhyCMAP were run with default parameters. The DCA methods were directly run on the alignments described above, whereas PhyCMAP runs its own workflow and thus uses its own alignment as described in [15]. PconsC2 was run as described before [17].

**PconsC3.** Figure S6 illustrates the workflow of PconsC3. Input features comprise contact predictions by plmDCA, GaussDCA as well as PhyCMAP, secondary structure prediction by PSIPRED 3.0 [29], and solvent accessibility prediction by NetSurfP 1.1 [30]. In PconsC3 PhyCMAP can be replaced by another contact predictor and we have successfully used CMapPro [16] with similar accuracy, data not shown.

Additionally, CD-HIT is run to generate statistics about the alignment (i.e. alignment depth at different sequence similarity cut-offs). The initial layer of PconsC3 takes these features as input and uses a random forest to predict a score for each possible contact. In contrast to previous work, PconsC3 applies pattern recognition already in the first layer. This results in an intermediate contact map. Every following layer uses all the initial features plus the output from the previous layer, given as a window of 11 by 11 residues around the current contact. Note that the pattern recognition method used in PconsC3 is analogous to convolutional layers of deep learning, as described in detail earlier [17].

The initial layer of PconsC3 (PconsC3-10) shows an increased precision over PconsC2 independent of the number of top-ranked predicted contacts used for evaluation (Fig. S2). The precision increases for each layer to saturate at the third layer. In contrast to PconsC2 the fourth and fifth layers does not increase the performance.

Each of the Random Forests comprising PconsC3 consists of 100 trees trained based on optimization of Gini impurity, with a constraint on node split with at least 100 samples per leaf. To reduce the memory footprint of training, as well as to prevent overfitting, starting from layer 1, we have disregarded a randomly chosen subset of 30% of the training samples, which appears to improve the generalizability of resulting statistical models.

At <https://github.com/mskwark/PconsC3/> instructions on how to setup and run PconsC3 locally are found. PconsC3 can also be used from a web-server at <http://c3.pcons.net/>, where predictions for  $\approx 14000$  of the sufficiently large Pfam domain families also can be found.

**Selecting top ranked contacts.** We analyzed the top ranked contacts using half of the number of observed contacts ( $N/2$ , dashed vertical line in Figure S2). This number roughly corresponds to the length of the protein ( $L$ ) (Supplementary Fig. S7), i.e. the same number of contacts used to analyze precision (PPV) earlier [17].

A native contact between two residues is present if their  $C_{\beta}$ -atoms is within  $8\text{\AA}$ . The contact score was used to rank predicted contacts and the top  $N/2$  contacts were used for evaluation. This allows for a fair comparison between the methods, while being easy to interpret, e.g. if a method has a PPV of 0.5 at  $N/2$  contacts, one can say that this method correctly predicts 25% of all observed contacts. Thereby, false and true negatives are implicitly taken into account. For this reason we decided to choose a cut-off based on  $N$  instead of the widely used cut-off based on the length of the input sequence  $L$ .

**Metrics.** Effective sequences is defined in analogy to [31] as:

$$B_{\text{eff}} = \sum_{b=1}^B \frac{1}{m_b} \quad [1]$$

where  $m_b$  is the number of sequences with at least 90% sequence identity  $m_b = |\{a \in \{1, \dots, B\} : \text{id}(\sigma^{(a)}\sigma^{(b)}) \geq 0.9\}|$ .

The quality of a predicted contact map is measured in positive predictive value (PPV), or precision:

$$PPV = \frac{TP}{TP + FP} \quad [2]$$

where TP is the number of predicted contacts that match a contact in the native structure (true positives) and FP the number of predicted contacts that don't (false positives).

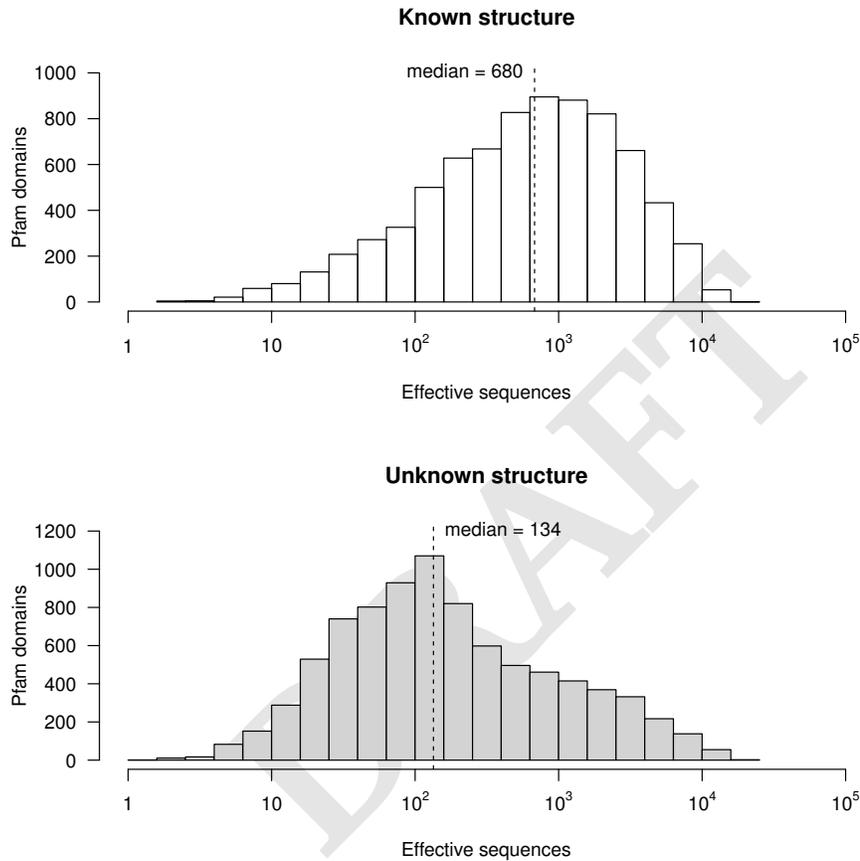
TM-score [24] is used to measure the similarity between predicted and native structure. To enable fair comparison with PconsC2, we used the same cut-off (top  $1.5 \cdot L$  contacts, where  $L$  denotes sequence length) as in [17] to select contacts for PconsFold. Preliminary observations clearly indicate that better performance can be obtained using another scheme, but we have not systematically evaluated this.

**ACKNOWLEDGMENTS.** This work was supported by grants from the Swedish Research Council (VR-NT 2012-5046) to A.E.) and Swedish e-Science Research Center (SeRC). Computational resources at the National Supercomputing Center were provided by SNIC. The website is maintained by the Bioinformatics Infrastructure for Life Sciences (BILS). The authors thank Nanjiang Shu for setting up the website. The authors declare that they have no competing financial interests.

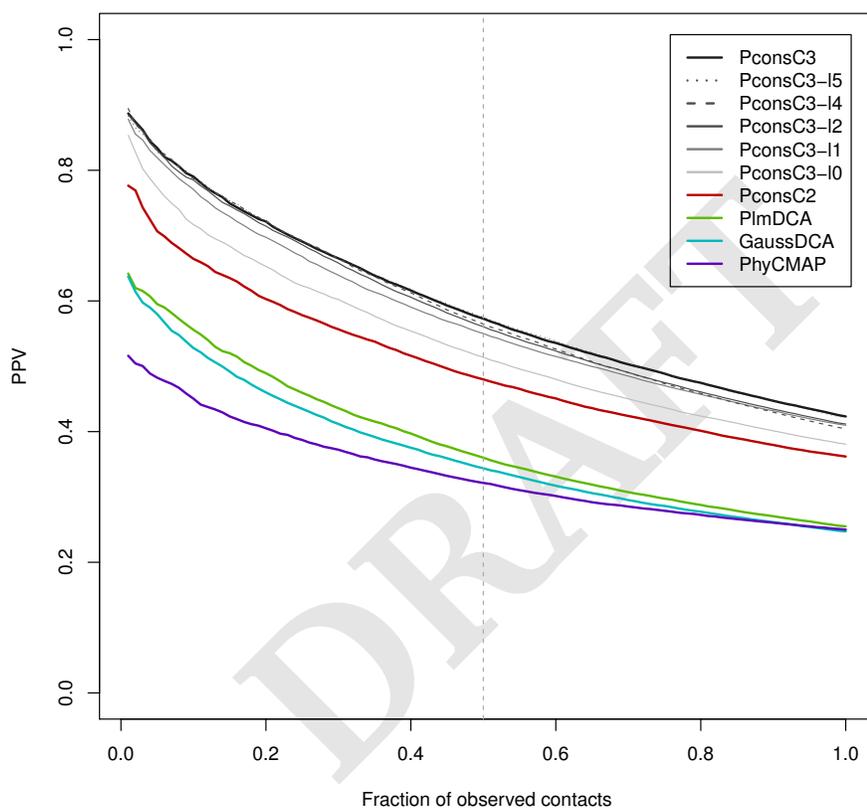
1. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* 106(1):67–72.
2. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology* 6(1):e1000633.
3. Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108(49):1293–301.
4. Marks DS et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS one* 6(12):e28766.
5. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nature biotechnology* 30(11):1072–1080.
6. Hopf TA et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
7. Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America* 109(24):1540–1547.
8. Hayat S, Sander C, Marks DS, Elofsson A (2015) All-atom 3D structure prediction of transmembrane  $\beta$ -barrel proteins from sequences. *Proceedings of the National Academy of Sciences of the United States of America* 112(17):5413–8.
9. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 3:e02030.
10. Hopf T et al. (2014) Sequence co-evolution gives 3d contacts and structures of protein complexes. *Elife* 3.
11. Aurell E (2016) The maximum entropy fallacy redux? *PLoS Comput Biol* 12(5):e1004777.
12. van Nimwegen E (2016) Inferring contacting residues within and between proteins: What do the probabilities mean? *PLoS Comput Biol* 12(5):e1004726.
13. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America* 110(39):15674–15679.
14. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309–17.
15. Wang Z, Xu J (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics (Oxford, England)* 29(13):i266–73.
16. Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. *Bioinformatics* 28(19):2449–2457.
17. Skwark MJ, Raimondi D, Michel M, Elofsson A (2014) Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology* 10(11).
18. Finn RD et al. (2014) Pfam: the protein families database. *Nucleic acids research* 42(1):222–230.
19. Berman HM et al. (2000) The Protein Data Bank. *Nucleic Acids Research* 28(1):235–242.
20. Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics* 276:341–356.
21. Baldassi C et al. (2014) Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 9(3):e92721.
22. Michel M et al. (2014) Pconsfold: improved contact predictions improve protein models. *Bioinformatics* 30(17):i482–8.
23. Grana O et al. (2005) Casp6 assessment of contact prediction. *Proteins: Structure, Function, and Bioinformatics* 61(S7):214–224.
24. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57(4):702–710.
25. Ovchinnikov S et al. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* 4:e09248.
26. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
27. Cheng H, Liao Y, Schaeffer RD, Grishin NV (2015) Manual classification strategies in the ECOD database. *Proteins* 83(7):1238–51.
28. Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 9(2):173–175.
29. Jones D (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292(2):195–202.
30. Petersen B, Petersen T, Andersen P, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 9:51.
31. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical review. E, Statistical, nonlinear, and soft matter physics* 87(1):012707.

## Supplementary Figures

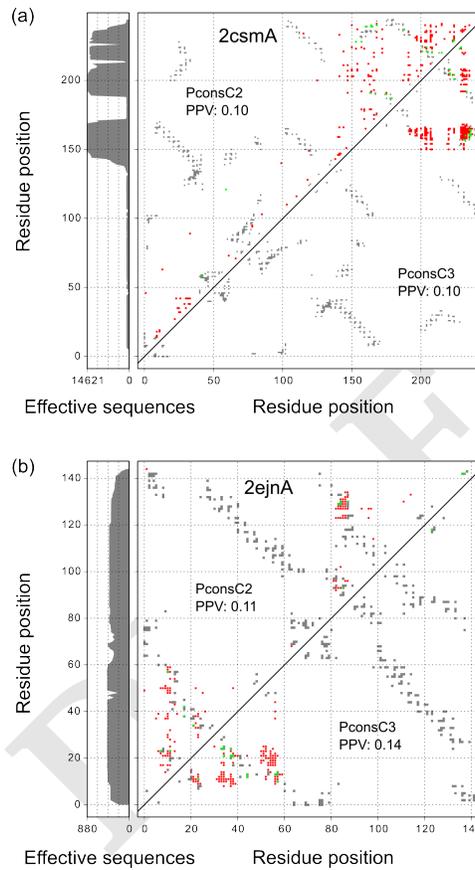
DRAFT



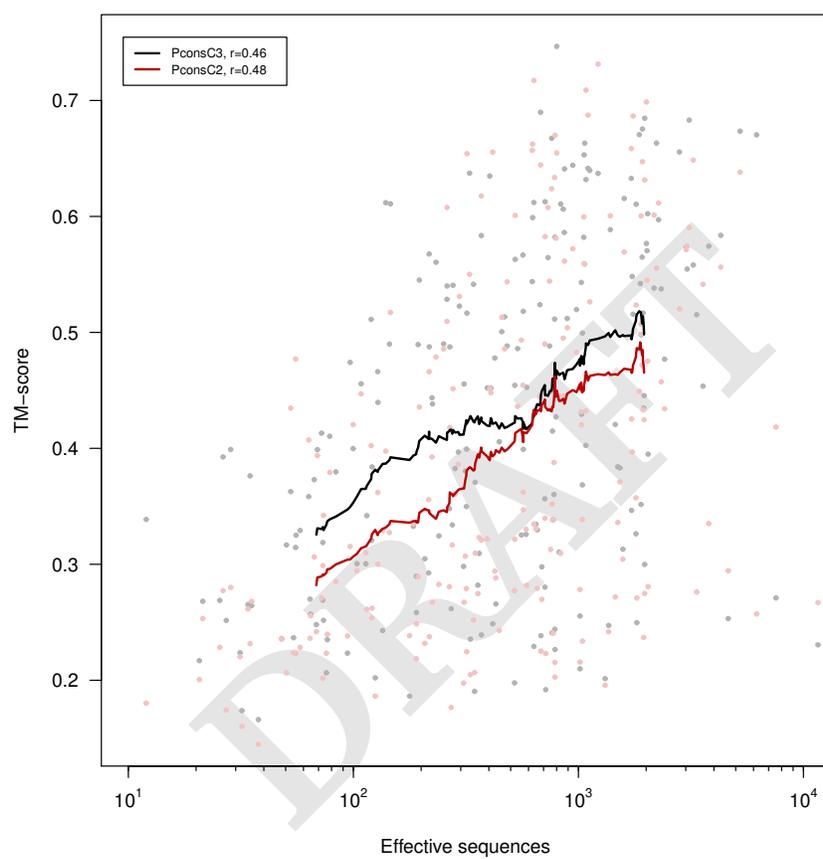
**Fig. S1.** Pfam family sizes measured in effective sequences. (a) Histogram over the size of all families with at least one hit in PDB that covers more than 80% and (b) of all families without such hit. (c) Cumulative counts of Pfam families over their size measured in effective sequences of the underlying alignment. There are 16295 protein families in Pfam 29.0 out of which 8562 do not have a significant hit to a structure in PDB, which would cover more than 75% of their length. The vertical line segments represent 100 and 1000 effective sequences. Horizontal line segments connect these to the corresponding number of Pfam families.



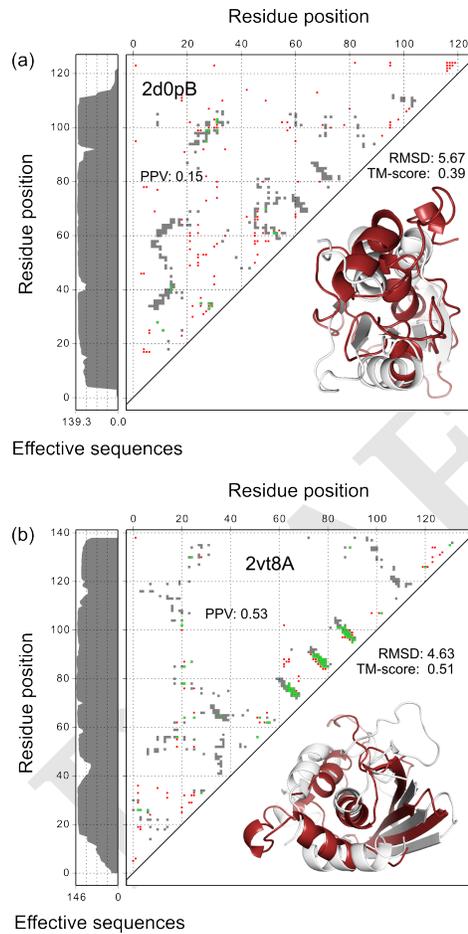
**Fig. S2.** Contact predictor performance on the independent test set measured in Positive Predictive Value (PPV or precision). Performance against the number of top ranked predicted contacts measured as a fraction of contacts observed in the native structure  $N$ . The dashed vertical line indicates the number of contacts used further on.



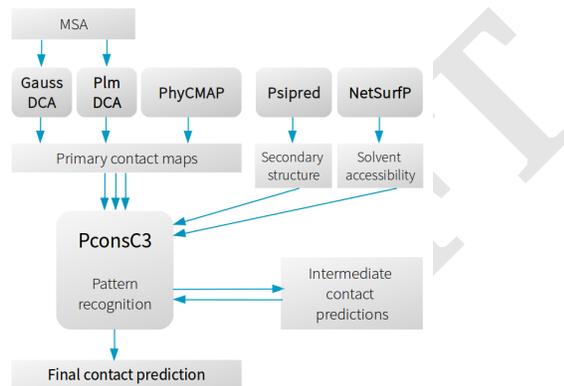
**Fig. S3.** Example proteins where most of the aligned sequences do not cover the entire length. The alignment coverage is indicated by the vertical panel on the left-hand side, where the width of the gray area represents the number of effective sequences at that position. (a) PDB: 2csmA is only covered in the terminal region (b) PDB: 2ejnA is mostly covered by only half of the sequences, thus consists of two separate parts.



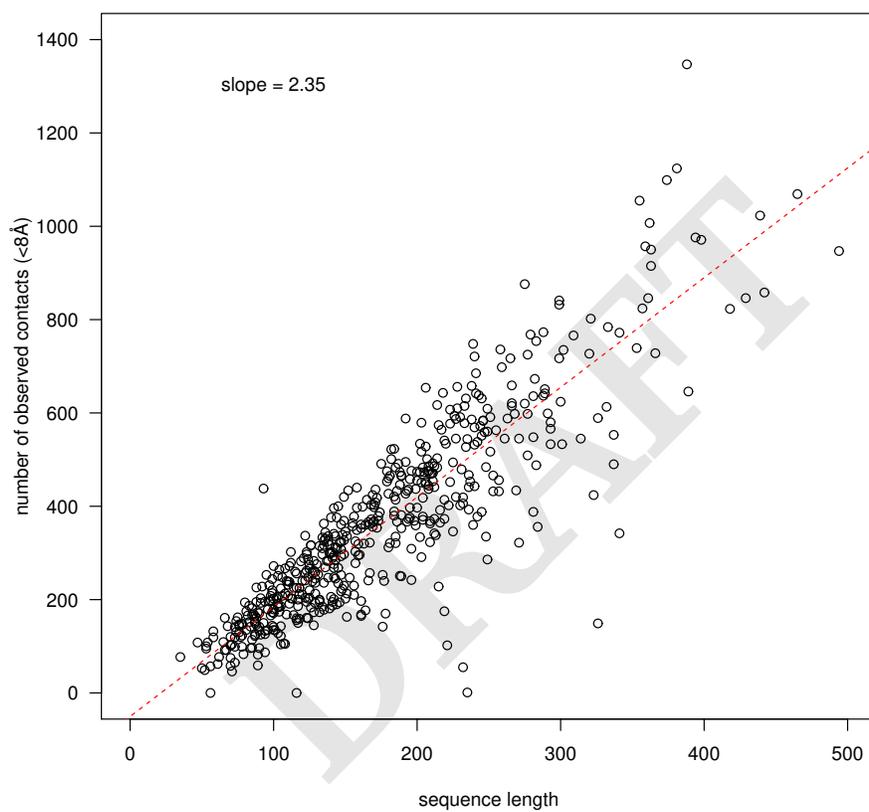
**Fig. S4.** TM-score of PconsFold against family size when using PconsC3 (black) or PconsC2 (red) contact predictions. Pearson correlation coefficient is denoted  $r$  and the lines indicate moving average with a window size of 50.



**Fig. S5.** Structure prediction. (a) Contact map for Diol dehydratase reactivase ATPase-like domain (Pfam: PF08841, PDB: 2d0pB) in the upper left triangle. The lower right triangle shows the structure predicted with PconsFold using PconsC3 contacts (black) superimposed onto the native structure from PDB (light gray) (b) Contact map and structure for the PI31 proteasome regulator N-terminal (PF11566, 2vt8A).



**Fig. S6.** PconsC3 workflow. GaussDCA and PlmDCA are combined with the non-DCA method PhyCMAP and additional secondary structure and solvent accessibility features. PconsC3 combines all features and iteratively predicts intermediate contact maps. In every iteration predictions from the previous layer are used as additional input providing a description of the contact pattern.



**Fig. S7.** Sequence length against the number of contacts observed in the native structure ( $N$ ) for all proteins in both training and test set. Native contacts are defined as  $C_{\beta}$ -atoms closer than  $8\text{\AA}$ . Counted are all native contacts with a sequence separation above 5 residues.

## Supplementary Tables

DRAFT

**Table S1. Training dataset**

PDB	H-class	$B_{eff}$	PDB	H-class	$B_{eff}$	PDB	H-class	$B_{eff}$
1A3AA	311.1	1726.20	1GUUA	101.1	2574.43	1P90A	2484.4	1058.83
1A6MA	106.1	548.31	1GZ2A	209.1	1965.15	1PCHA	322.1	931.28
1A70A	221.1	2131.21	1GZCA	10.1	486.72	1PKOA	11.1	2431.66
1AAPA	384.1	2533.09	1H0PA	75.1	2318.65	1Q67A	220.1	181.24
1ABAA	2485.1	1779.91	1H2EA	2111.21	2703.07	1QBEC	265.1	11.57
1AG6A	3156.1	1866.60	1H4XA	2496.1	3083.53	1QF9A	2004.1	1402.59
1AGQC	385.1	560.20	1H98A	205.1	3937.00	1QJPA	5084.1	2617.32
1AOEA	2111.5	2052.75	1HDOA	2003.1	1560.05	1QL0A	378.1	470.79
1ATLA	2498.1	1609.51	1HFCA	2498.1	3282.86	1R26A	2485.1	5975.54
1ATZA	2006.1	3776.10	1HH8A	109.4	10459.26	1ROAA	243.3	576.15
1AVSA	108.1	5753.10	1HLQA	374.1	108.53	1RW1A	2485.1	2142.18
1BDOA	325.1	2160.40	1HTWA	2004.1	3782.36	1RW7A	2007.1	2612.90
1BEBA	9.1	487.69	1HXNA	5.1	678.23	1RYBA	2011.5	1574.55
1BEHA	11.1	1059.91	1I1JA	4.1	4950.19	1SMXA	2.1	2655.35
1BKRA	193.1	373.66	1I1NA	2003.1	4807.76	1SVYA	224.1	726.59
1BRFA	375.1	662.92	1I4JA	218.2	1097.73	1T8KA	132.1	4159.78
1BSGA	4019.1	1344.14	1I58A	225.1	6851.76	1TIFA	221.2	2258.24
1BSGA	223.3	1344.14	1I5GA	2485.1	6276.14	1TQGA	601.3	1660.68
1C44A	268.1	1794.10	1I71A	380.1	619.65	1TQHA	2111.78	3066.46
1C52A	107.1	5049.41	1IHZA	2.6	1309.48	1TZVA	195.1	932.36
1C9OA	2.1	2237.31	1IIBA	2009.1	1094.72	1VFYA	376.1	1961.35
1CC8A	304.3	2589.60	1IM5A	2111.61	721.27	1VHUA	2111.27	1273.41
1CHDA	2111.46	1731.08	1IWDA	219.1	1357.38	1VJKA	221.1	2306.80
1CJWA	213.1	3763.54	1J3AA	2490.2	874.59	1VMBA	304.12	1052.12
1CKEA	2004.1	3060.75	1JBEA	2007.4	7297.65	1VP6A	10.12	5097.18
1CNT1	150.3	65.52	1JBKA	2004.1	838.09	1VZHB	11.1	984.69
1COMF	301.7	2454.76	1JFUA	2485.1	6465.70	1VZHB	375.1	984.69
1CTFA	308.1	695.64	1JFXA	2002.1	675.18	1WOHA	2484.1	1883.19
1CXYA	243.7	1466.56	1JKXA	2111.71	1451.61	1WHIA	3174.1	626.14
1CZNA	2007.2	2841.11	1JL1A	2484.1	4281.69	1WJXA	2.21	1107.79
1D0QA	375.1	1105.01	1J00A	328.4	491.88	1WKCA	2003.1	1544.73
1D1QA	2009.1	635.49	1J08A	4.1	3064.00	1XDZA	2003.1	4498.06
1D4OA	2003.1	670.45	1JOSA	327.1	1022.17	1XFFA	210.1	838.00
1DBXA	299.1	429.50	1JVWA	284.1	3418.06	1XKRA	866.1	371.41
1DEVK	73.1	140.51	1JWQA	2011.1	1670.13	1YWYA	220.2	24.52
1DIXA	280.1	327.30	1JYHA	886.1	1283.65	2ARCA	10.12	3209.89
1DLWA	106.1	617.75	1K6KA	148.1	1335.33	2AYGB	304.15	195.51
1DMGA	2111.59	1010.39	1K7CA	2007.5	3846.59	2BPA2	10.2	16.09
1DQGA	6.1	2673.29	1K7JA	297.1	969.61	2CUAA	3156.1	2523.73
1DSXA	226.1	978.15	1KIDA	2487.1	579.41	2FPNA	331.13	135.98
1EAZA	220.1	1650.07	1KQ6A	277.1	2218.74	2FQOA	309.1	184.86
1EJ0A	2003.1	1775.19	1KQRA	10.1	99.08	2HS1A	1.1	567.34
1EJ8A	11.1	1018.15	1KTGA	221.4	3856.80	2IVFC	11.1	390.09
1EK0A	2004.1	2405.08	1KU3A	101.1	4640.25	2MHRA	601.14	939.66
1F6BA	2004.1	2712.01	1KW4A	102.1	2361.74	2PHYA	223.1	9261.12
1FCYA	188.1	784.59	1LM4A	289.1	1699.33	2TPSA	2002.1	1447.57
1FK5A	185.1	577.92	1LO7A	222.1	1482.62	2VXNA	2002.1	2530.14
1FL0A	2.1	1034.68	1LPYA	235.1	908.37	3BORA	2004.1	4543.74
1FNAA	11.1	1652.63	1M4JA	224.1	317.28	3C5NA	844.1	159.49
1FQTA	66.1	1153.06	1M8AA	4.9	357.62	3DQGA	511.1	1707.22
1FQTA	4294.1	1153.06	1MK0A	821.1	1807.97	3HHWC	4069.1	57.58
1FVGA	304.3	1167.98	1MUGA	2111.69	362.56	3I6B	719.1	121.03
1FVKA	3009.1	4101.17	1NB9A	1.1	1813.25	3I6B	101.1	121.03
1FVKA	2485.1	4101.17	1NE2A	2003.1	3797.34	3I6B	3939.1	121.03
1FX2A	304.48	3084.82	1NFHB	328.1	173.63	3MOZB	2002.1	986.14
1G2RA	377.4	633.50	1NPSA	72.1	562.18	3NRHB	601.3	89.23
1G9OA	7.1	4435.82	1NRVA	214.1	1293.21	3PG6B	3169.1	227.15
1GBSA	235.1	2845.60	1NY1A	2002.3	3174.29	4KHBB	220.1	205.13
1GMIA	11.2	3279.27	1O1ZA	2002.1	1888.28	5MSFA	265.1	19.64
1GMXA	2009.1	4207.94	1O27A	842.1	356.44	5PTPA	1.1	3576.88

**Table S2. Test dataset**

PDB	H-class	$B_{eff}$	PDB	H-class	$B_{eff}$	PDB	H-class	$B_{eff}$	PDB	H-class	$B_{eff}$
1AHSC	10.5	34.00	1S3FB	2007.15	1884.35	2GVIA	298.2	764.12	3D2QD	906.1	1817.19
1C2YD	2111.74	2168.53	1S68A	206.1	1800.90	2GVIA	377.1	764.12	3DBYF	633.7	128.52
1C9YA	2111.2	1221.71	1SUDA	2499.1	2223.17	2H44A	131.1	3577.44	3DKXB	101.39	429.61
1CCTA	62.1	752.72	1SWXA	631.1	239.76	2HGHA	386.1	11605.96	3DKXB	304.55	429.61
1COZA	2005.1	1080.17	1SYHA	2111.17	2415.38	2HI7B	633.13	251.74	3EB6B	216.1	800.93
1DBRA	2111.73	2030.74	1TD4A	70.4	224.67	2HJJA	330.7	357.31	3EW1A	9.2	128.29
1DCHF	305.2	865.16	1TFKB	601.5	184.56	2HLOA	2494.1	108.21	3G74B	3068.1	190.29
1EDIA	632.2	55.36	1TJLD	192.1	794.31	2I9LI	3607.1	25.47	3GUVA	2111.66	2017.55
1EFDN	2111.15	1805.37	1TJLD	377.1	794.31	2IA9E	4180.1	393.40	3GYVA	4096.1	275.15
1F46B	331.7	232.82	1UWZB	2492.1	1970.16	2II9B	10.7	318.61	3GZFC	3084.1	21.41
1F68A	633.1	1717.59	1VCRA	5077.1	633.76	2J1KQ	5092.1	54.19	3H8DB	3091.1	519.89
1FHIA	312.1	2333.52	1VJNA	247.1	4292.73	2J3WA	223.2	223.52	3H90A	5082.1	1958.14
1FJRB	4.12	177.71	1VQZA	244.3	1528.59	2J8WB	601.2	319.49	3H90A	327.7	1958.14
1FS0G	2111.9	855.74	1VQZA	314.1	1528.59	2JOVA	4328.1	328.17	3HPGL	611.3	73.67
1G61A	232.1	196.34	1W8AA	207.1	6190.19	2JYNA	6060.1	114.75	3HTYJ	9.23	833.89
1GJJA	130.1	656.04	1W9GB	4326.1	103.72	2KYSA	632.1	26.35	3I9OB	2007.15	125.06
1GLGA	2111.16	2004.75	1WD5A	2111.73	1863.72	2KZSA	611.1	47.96	3IQZF	2111.24	4633.67
1GPSA	387.1	314.45	1WIGA	377.1	1968.97	2MOMB	3640.1	68.44	3K43B	63.1	709.16
1H68A	5001.1	568.60	1WPVB	849.1	66.97	2NQ2A	5065.1	1748.96	3K8RB	719.2	456.28
1I95E	212.1	783.50	1X0PJ	304.11	469.63	2NR9A	5081.1	1059.50	3KZLA	2111.45	277.09
1I95E	330.1	783.50	1X48A	330.1	3012.74	2OF5H	110.1	867.09	3LW5L	5064.1	75.87
1I97T	604.9	787.23	1X8HA	247.1	1597.12	2OGFD	283.3	96.68	3M71A	109.37	370.08
1IMBB	4018.1	1385.68	1X91A	633.4	625.66	2OHCA	242.2	331.23	3MEZA	5.3	1575.83
1IMBB	2111.88	1385.68	1XBAA	206.1	5230.81	2OHCA	2008.2	331.23	3N1GA	307.1	1041.02
1IMXA	367.1	207.66	1XQFA	5049.1	623.56	2OJ5C	5092.1	715.72	3NJSA	292.1	193.91
1IR1S	302.2	73.15	1XS6A	70.2	687.67	2ONKC	5080.1	2270.61	3O7JA	4110.1	482.11
1IS9A	109.2	282.98	1Y4HD	9.6	12.00	2OPIA	281.1	678.62	3OFEB	3122.1	63.28
1JGPR	616.1	732.71	1Y60C	212.1	52.63	2PAVP	223.2	260.82	3OQIA	2005.1	73.13
1JHOL	5037.1	134.95	1Y66F	2486.1	1123.12	2PLSF	217.2	1103.38	3P45J	2111.76	778.97
1K6LH	4.6	1016.07	1YHOO	2490.3	891.82	2Q7RA	5038.2	821.17	3PC7B	2111.68	1936.04
1K6LH	5002.1	1016.07	1YQFF	897.1	216.01	2QQDE	303.1	69.18	3PJZA	5054.1	944.04
1KNVB	2008.1	20.67	1YWSA	375.3	1079.05	2QYFD	859.1	55.61	3QE7A	3226.1	370.60
1KNYB	601.7	2040.19	1Z7ME	2111.17	1897.10	2RDOL	2490.3	585.50	3QNQA	3227.1	569.01
1KNYB	316.1	2040.19	1ZD7B	69.1	1495.32	2RMRA	509.1	566.75	3RBYB	9.14	113.56
1KQPA	2005.1	1056.26	1ZJ0A	2111.6	1949.71	2RTBB	9.2	83.57	3T3TB	241.2	408.23
1LDIA	5048.1	1899.20	1ZWYC	2111.12	403.64	2VGRA	867.1	94.10	3UD2B	12.5	272.08
1LQKB	211.1	3040.40	2A84A	283.1	3238.54	2VT8A	241.15	146.08	3UWSA	2111.76	217.06
1M12A	198.1	766.88	2A84A	2005.1	3238.54	2WNKA	3156.3	76.57	3UYUB	207.1	267.03
1MB6A	387.1	35.30	2A9KB	237.1	589.37	2WNYA	882.1	119.93	3V3LB	4357.1	536.71
1MFRP	150.1	744.09	2AMCA	314.1	1733.67	2XVTF	3720.1	64.58	3VHHB	9.2	78.49
1MR7A	208.1	3352.63	2AV5D	304.57	295.77	2Y9PB	2111.117	50.40	3VX6A	3351.1	190.61
1N2ZB	2111.15	1925.17	2B9NX	304.109	1066.80	2YADB	3417.1	72.14	3ZNUG	304.4	523.53
1N5BA	241.1	444.92	2BWEL	103.1	2810.14	2YZOA	2111.103	260.46	3ZUXA	3236.1	1456.61
1N60C	244.3	3107.13	2C2OA	2111.76	1918.39	2ZITD	237.1	786.86	4A5ZB	300.1	3792.17
1N60C	217.1	3107.13	2CB6A	237.1	308.72	3A1JB	227.1	317.70	4AI3A	5095.1	1016.49
1NQLB	389.1	7548.37	2CCCA	872.1	310.24	3ANZW	11.13	37.83	4ARDB	170.1	57.99
1OAGA	136.1	415.92	2CDMC	304.55	521.63	3AXGI	231.1	425.06	4AU0B	2002.2	144.62
1OTFF	315.1	710.96	2CJRB	808.1	31.34	3B2UB	207.1	647.54	4DLHB	3445.1	268.16
1P3HE	236.1	1042.04	2CSMA	164.1	681.45	3B71B	601.16	64.17	4E1YB	109.2	1032.43
1PCFA	295.1	292.73	2D0PB	2111.13	139.29	3B7AA	108.2	633.86	4E6FA	331.3	121.05
1PDFE	3019.1	27.23	2D2CN	5069.1	557.87	3BLAB	519.1	1314.23	4F0DA	609.1	772.37
1PDFE	877.1	27.23	2DIOC	9.13	362.20	3BLAB	312.1	1314.23	4F0DA	237.1	772.37
1PS1A	141.1	760.75	2E2AB	604.1	674.11	3BP9B	170.1	120.60	4HBRC	809.2	980.47
1RD9D	2.2	1362.36	2EJNA	173.1	28.49	3CPWT	377.1	709.72	4IOSH	5092.1	217.32
1RH7C	929.1	345.56	2F0RA	216.1	1722.59	3CVZC	3237.1	31.96	4IZJC	212.1	785.61
1RL9A	199.1	347.74	2FEEB	5053.1	1029.91	3CVZC	302.3	31.96	4IZJC	3010.1	785.61
1RL9A	321.1	347.74	2FJCO	150.1	341.80	3CXJC	241.1	34.73	4J32B	3368.1	99.71

**Table S3. Average PPV of top  $N/2$  predicted contacts on the independent test set for the original ECOD secondary structure classes.**

method	all	$\alpha/\beta$	$\alpha+\beta$	$\alpha$	$\beta$	extended	few	mixed
PconsC3	0.57	0.62	0.61	0.49	0.59	0.46	0.54	0.77
PconsC2	0.48	0.55	0.49	0.44	0.46	0.39	0.44	0.78
PlmDCA	0.36	0.43	0.35	0.34	0.32	0.36	0.37	0.61
GaussDCA	0.34	0.40	0.33	0.33	0.31	0.36	0.36	0.60
PhyCMAP	0.32	0.34	0.37	0.23	0.35	0.00	0.39	0.45
counts	210	40	69	55	34	1	10	1