Protein structure prediction using metagenome sequence data

Alex Chu
CS 371 - Prof. Ron Dror
January 23, 2018

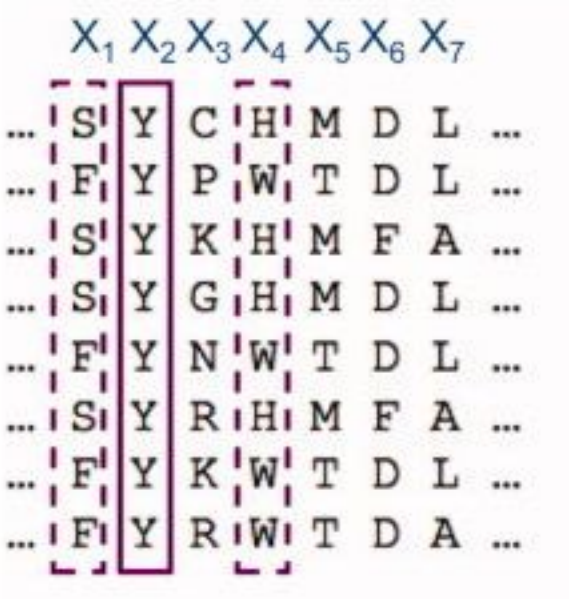
14,849 Pfam protein families

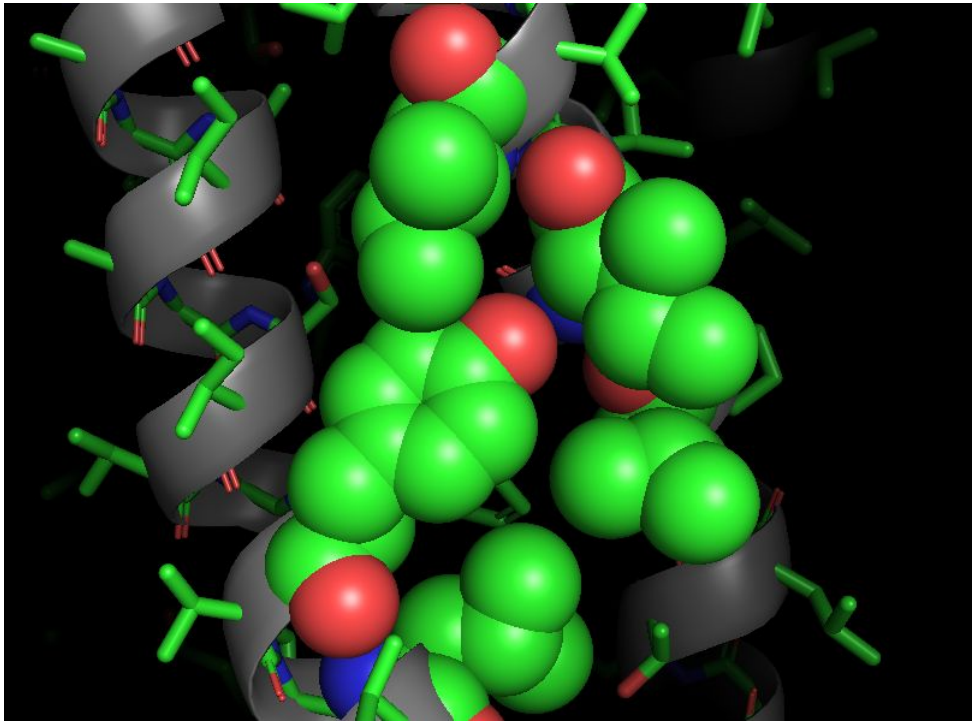
4,752 contain at least one experimentally solved 3D structure

4,886 can be modeled to some extent using comparative homology modeling

5,211 with no known structure and no structurally characterized homologs...

Ab initio/de novo structure prediction is not very good… so use evolutionary couplings





Balakrishnan et al, *Proteins* 2011

Approach

Used to calculate a simple covariance matrix, but too many false positives. (Positions that covary, but are not structurally linked.)

GREMLIN is one technique that mitigates this error (learns a probabilistic graphical model from a multiple sequence alignment)

Approach

Combine GREMLIN with existing de novo prediction software from Rosetta

Large-scale determination of previously unsolved protein structures using evolutionary information

Sergey Ovchinnikov[1], Lisa Kinch[2], Hahnbeom Park[1], Yuxing Liao[3], Jimin Pei[2], David E Kim[1], Hetunandan Kamisetty[4], Nick V Grishin[2,3], David Baker[1,5]*

... but still limited by the amount of sequence data available.

Use metagenomics data!

~2 billion partial and full-length proteins from ~5000 metagenomes from the Integrated Microbial Genomes database

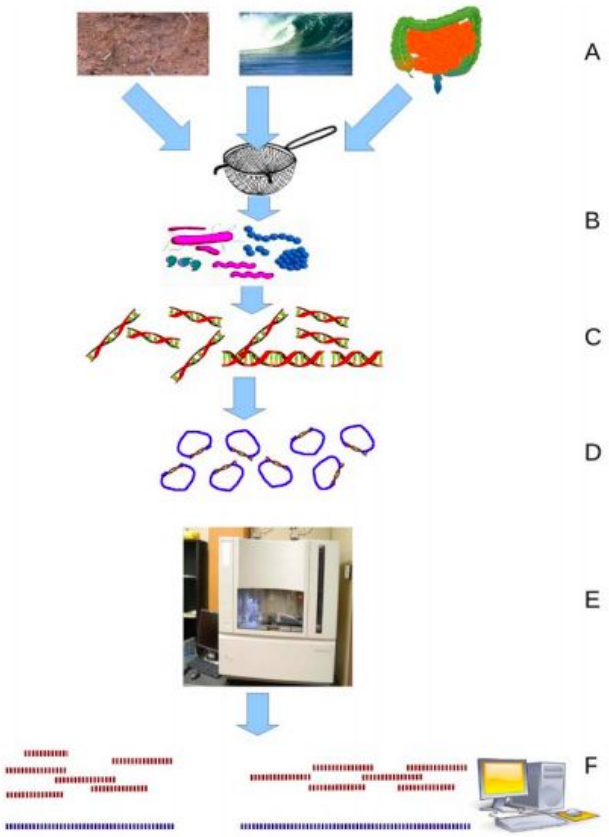
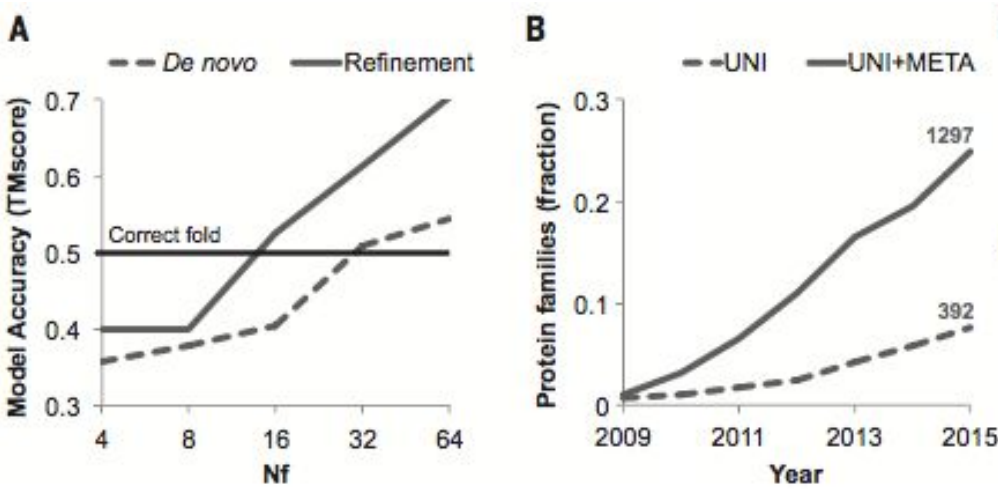


Figure 1. Environmental Shotgun Sequencing (ESS). (A) Sampling from habitat; (B) filtering particles, typically by size; (C) DNA extraction and lysis; (D) cloning and library; (E) sequence the clones; (F) sequence assembly.

Wooley et al, *PLOS Comput Biol* 2010

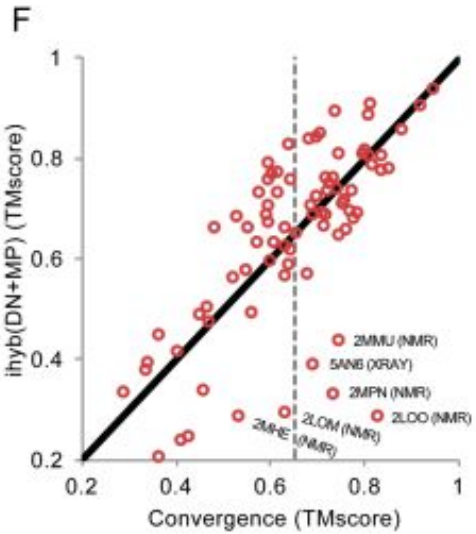
How useful was this extra data?

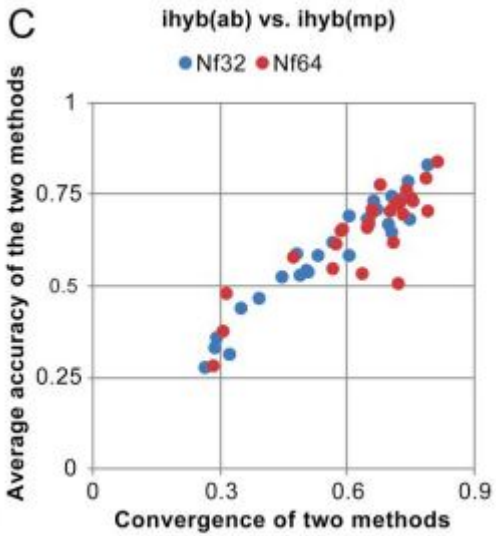


Nf: a metric that describes how amenable a protein family is to this method, by relating the length of the protein, the number of sequences in the family, and the diversity of the sequences

How do we assess the quality of a predicted structure?

It turns out structure prediction convergence is a good measure for the quality of the prediction



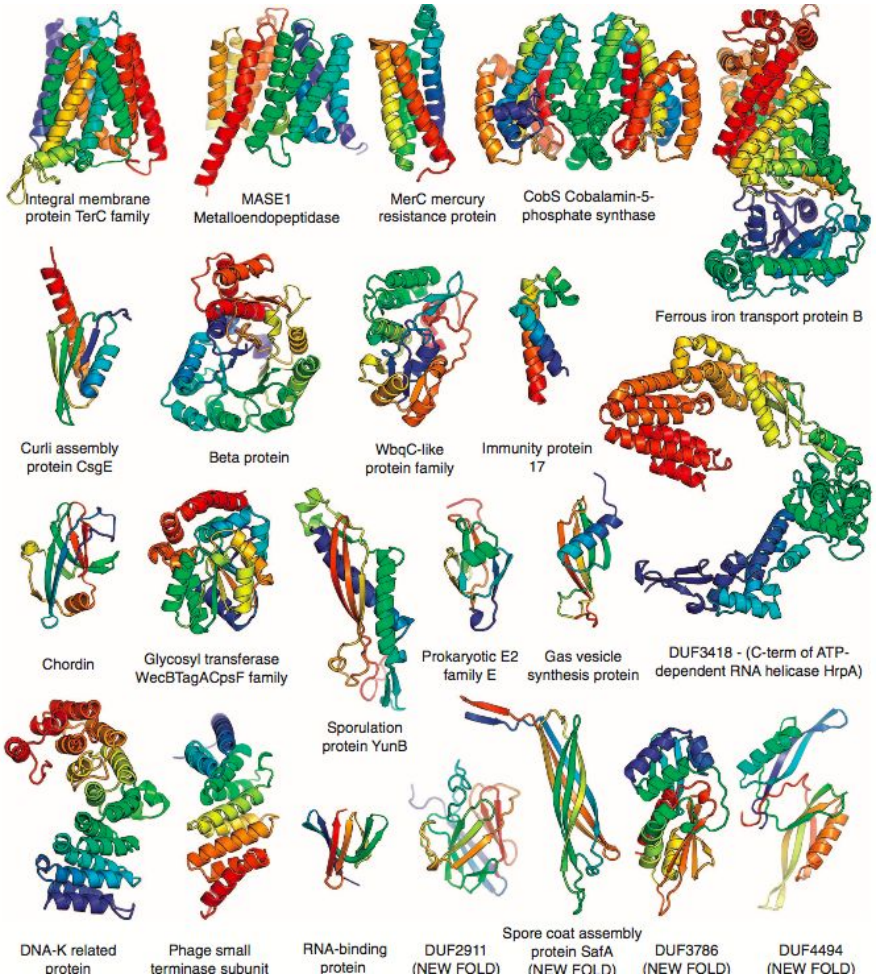


Findings

Metagenomic data allowed prediction 33% of unmodeled Pfams (compared with 16% without metagenome data)

612 new Pfams predicted

137 of these are novel folds



Strengths

Leveraged advances over the last ~10 years in high-throughput sequencing (especially in metagenomics), and in evolutionary coupling analysis.

This methods has generated one of the largest advances ever in structural genomics (quality predictions generated for >600 new families and >100 novel folds discovered), with promises of more as more sequence data becomes available.

Limitations

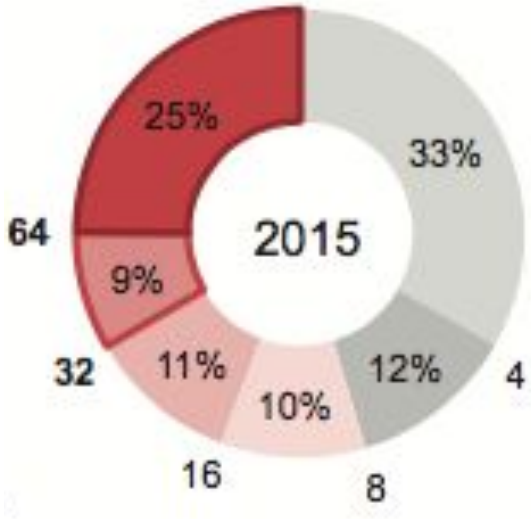
The algorithm doesn’t explicitly model the presence of membrane bilayers or endogenous ligands.

How generalizable is this method? Even for families with Nf > 64, prediction calculations converge just over half of the time.

Bacterial genomes and proteins are highly overrepresented in metagenomics studies.

Limitations

We still can’t predict structure for over half of the families with unknown structure. (But is this even a fair criticism?)



Potential Next Steps

Can this be used at all to improve current homology modeling methods?

Can we improve the method to predict more structures with less sequences (i.e. at lower Nf values)?

Can we predict or incorporate functional sites into the predictions?

Questions?

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, Jinbo Xu

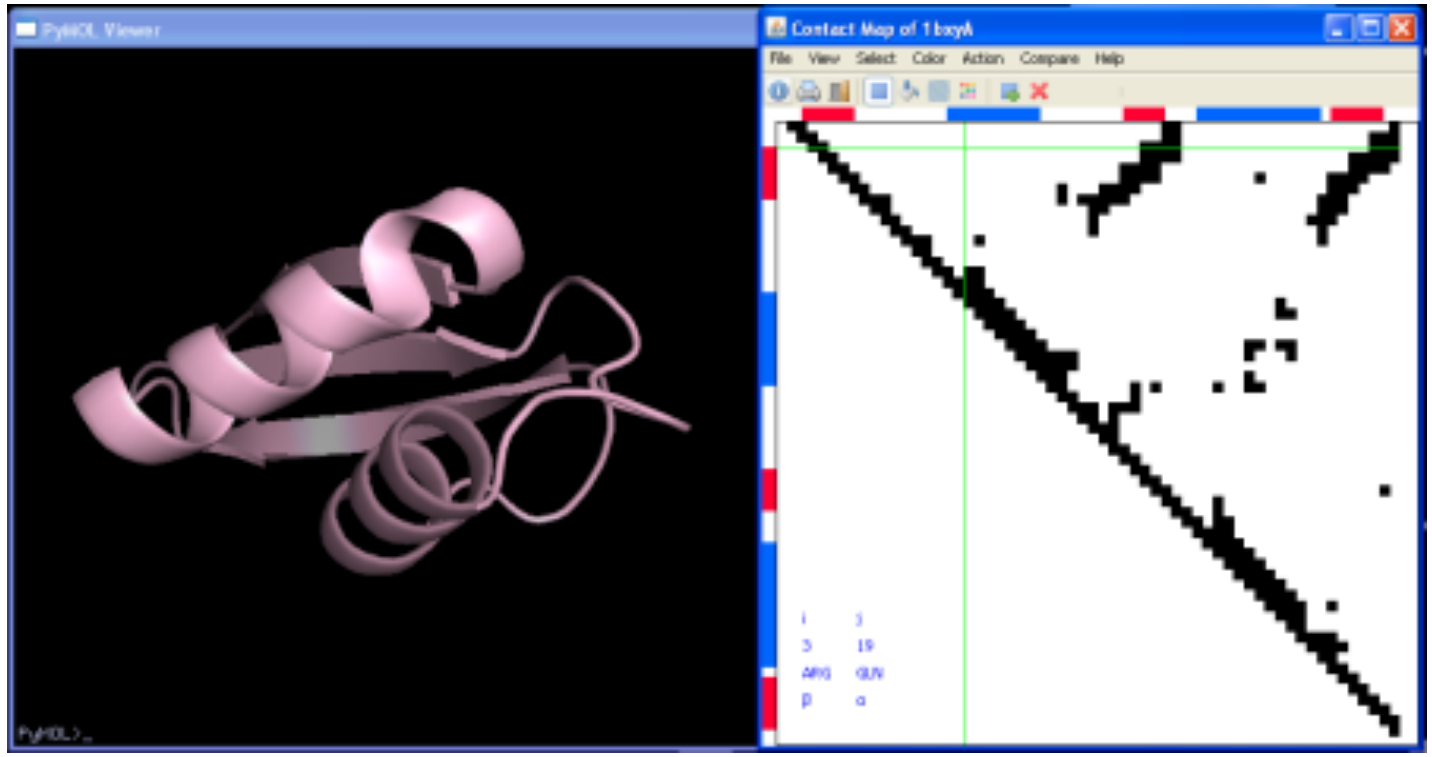
Ankit Baghel

CS371

01/23/18

Creating Protein Contact Maps

Protein Contact Maps



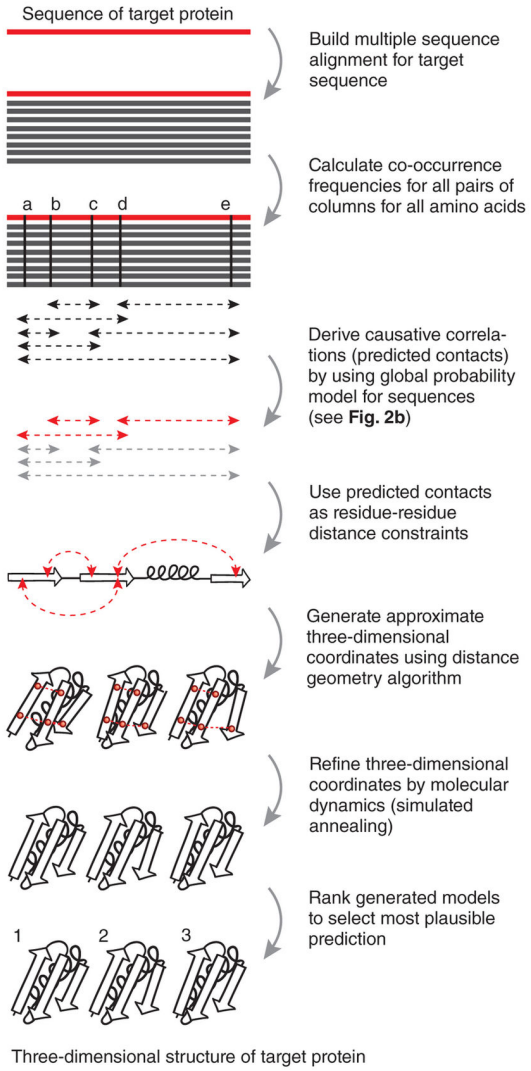
http://www.bioinformatics.org/cmview/screenshots.html

Existing Contact Prediction Methods

- Evolutionary Coupling Analysis (ECA)
  - PSICOV
  - plmDCA
  - Gremlin
  - **CCMpred**
- Supervised Machine Learning
  - **MetaPSICOV**
  - SVMSEQ
  - CMAPpro
  - **PconsC2**
  - PhyCMAP
  - CoinDCA-NN
  - CMAPpro

**Not an exhaustive list.

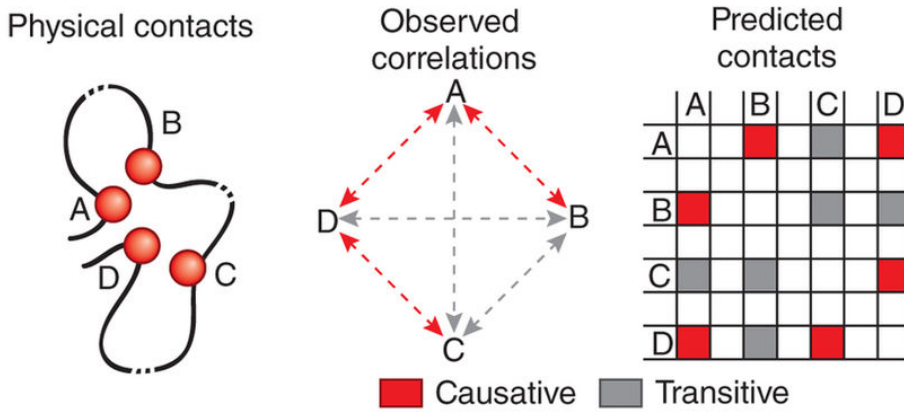
ECA Driven Contact Prediction Uses Correlations Between Residues



Protein structure prediction from sequence variation

Debora S Marks, Thomas A Hopf & Chris Sander

Nature Biotechnology **30**, 1072–1080 (2012)
doi:10.1038/nbt.2419
Download Citation
Protein structure predictions Proteomics
Received: 28 August 2012
Accepted: 15 October 2012
Published online: 08 November 2012



Supervised Machine Learning Incorporates More Context

OPEN ACCESS Freely available online

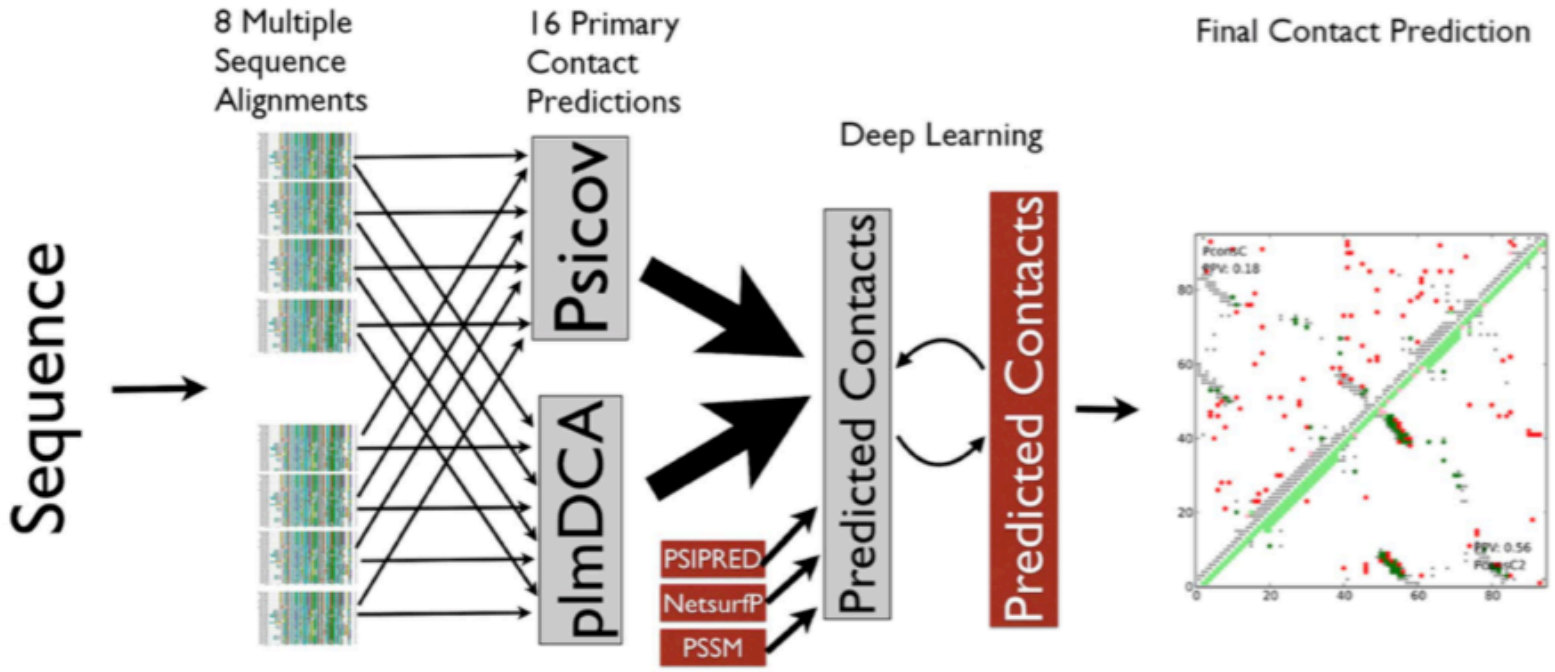
PLOS | COMPUTATIONAL BIOLOGY

Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns

Marcin J. Skwark[1,2,3], Daniele Raimondi[1,2,4], Mirco Michel[1,2], Arne Elofsson[1,2]*

1 Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, 2 Science for Life Laboratory, Stockholm University, Solna, Sweden, 3 Department of Information and Computer Science, Aalto University, Aalto, Finland, 4 Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, La Plaine Campus, Triomflaan, Brussels, Belgium

PconsC2 Contact Prediction



~~Creating Protein Contact Maps~~

Deep Residual Learning

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

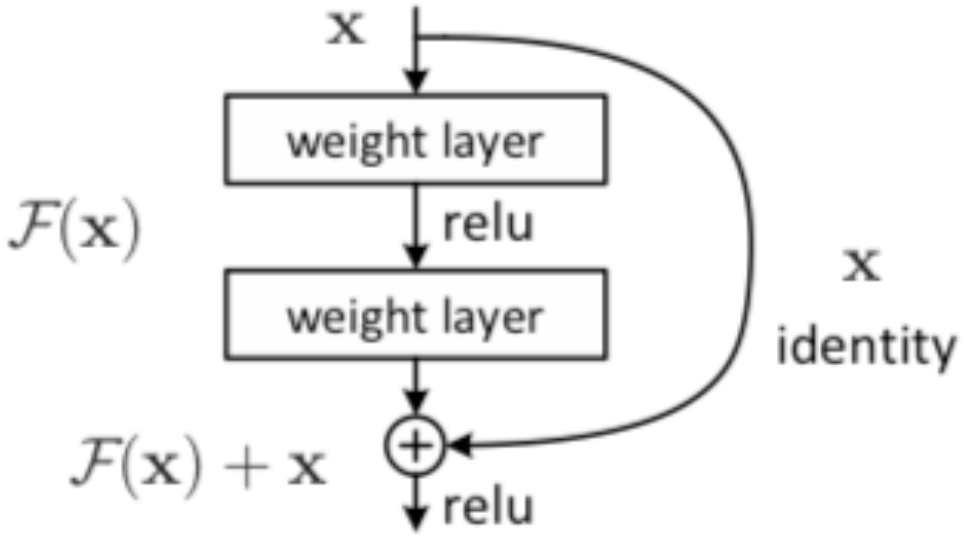
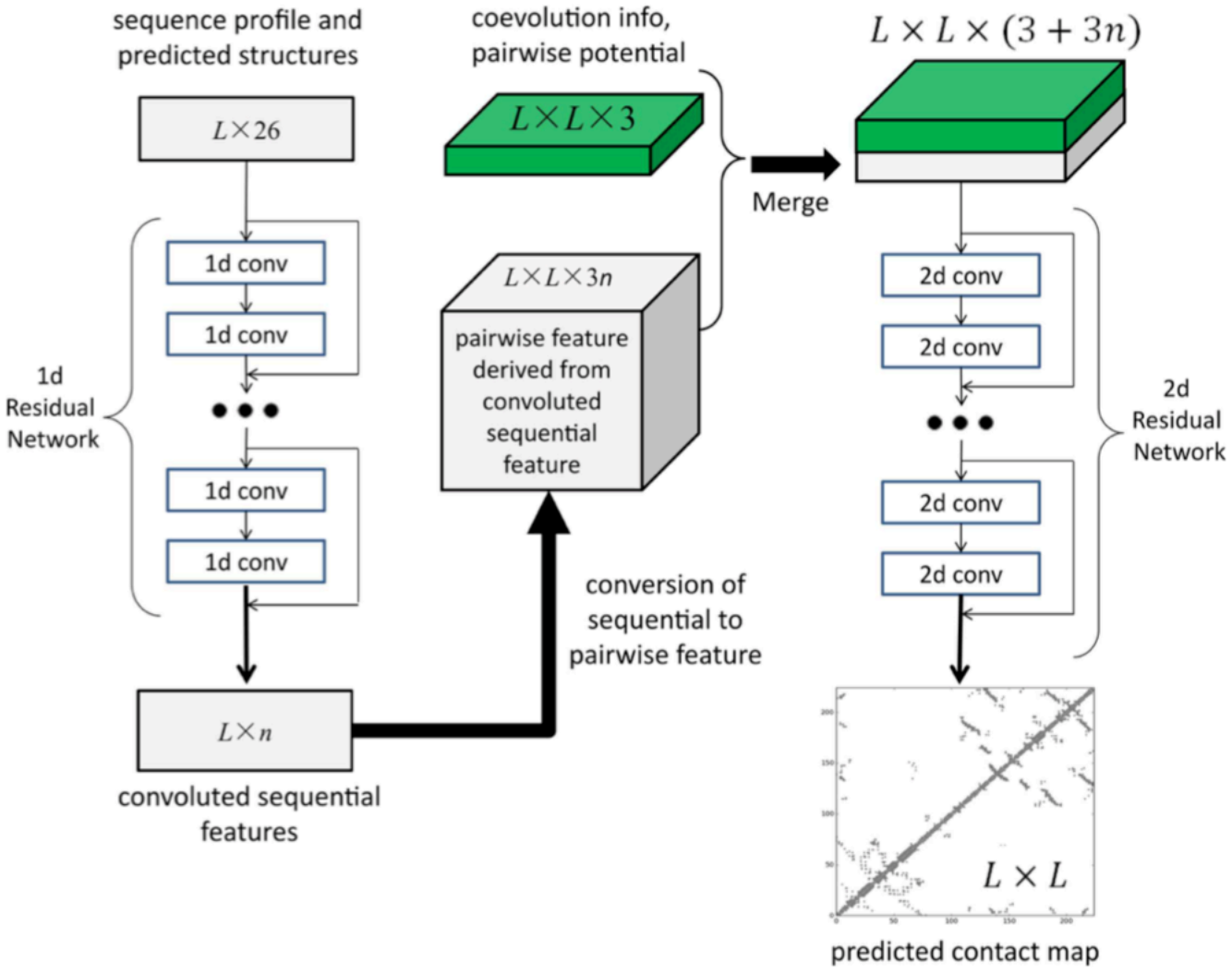
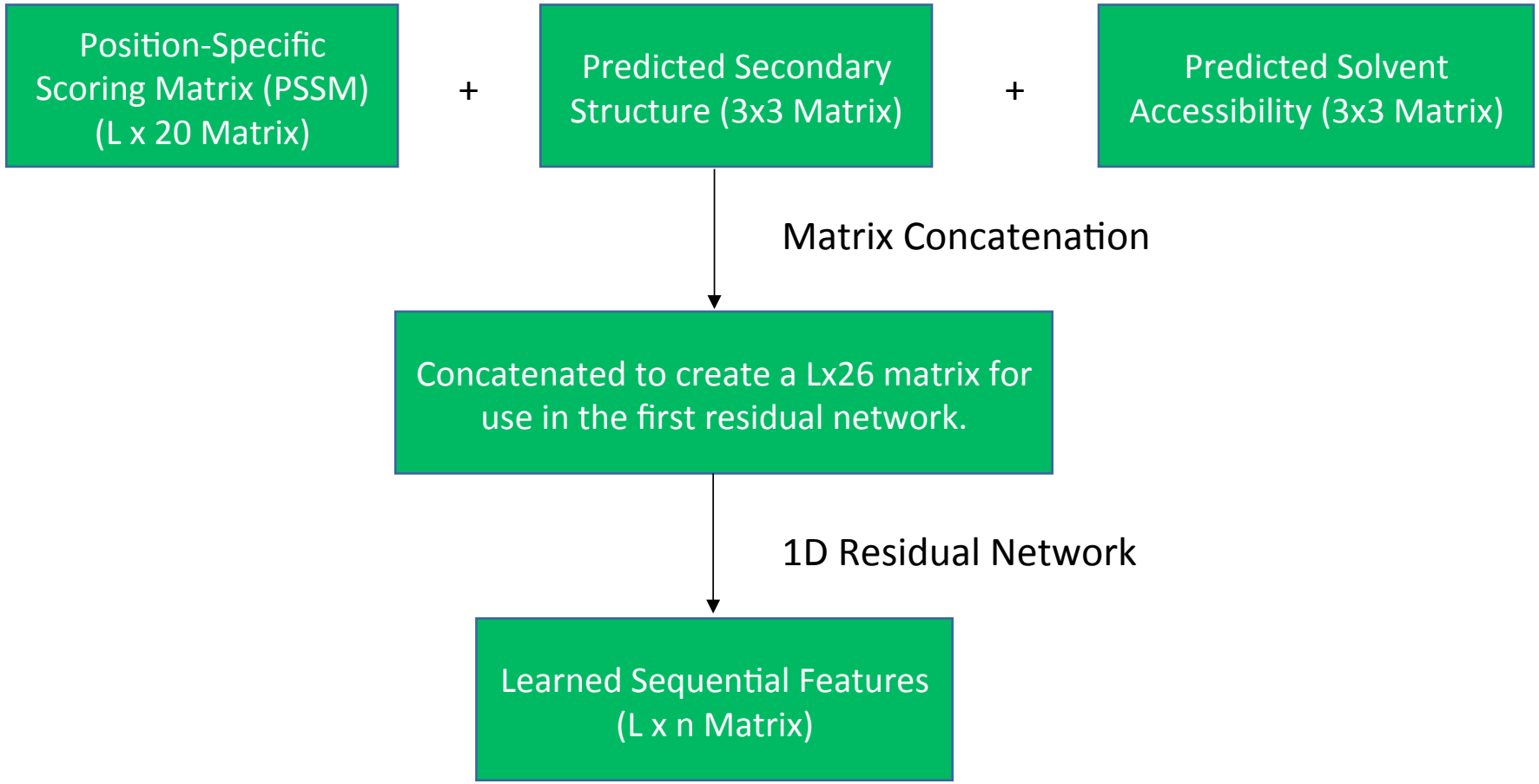


Figure 2. Residual learning: a building block.

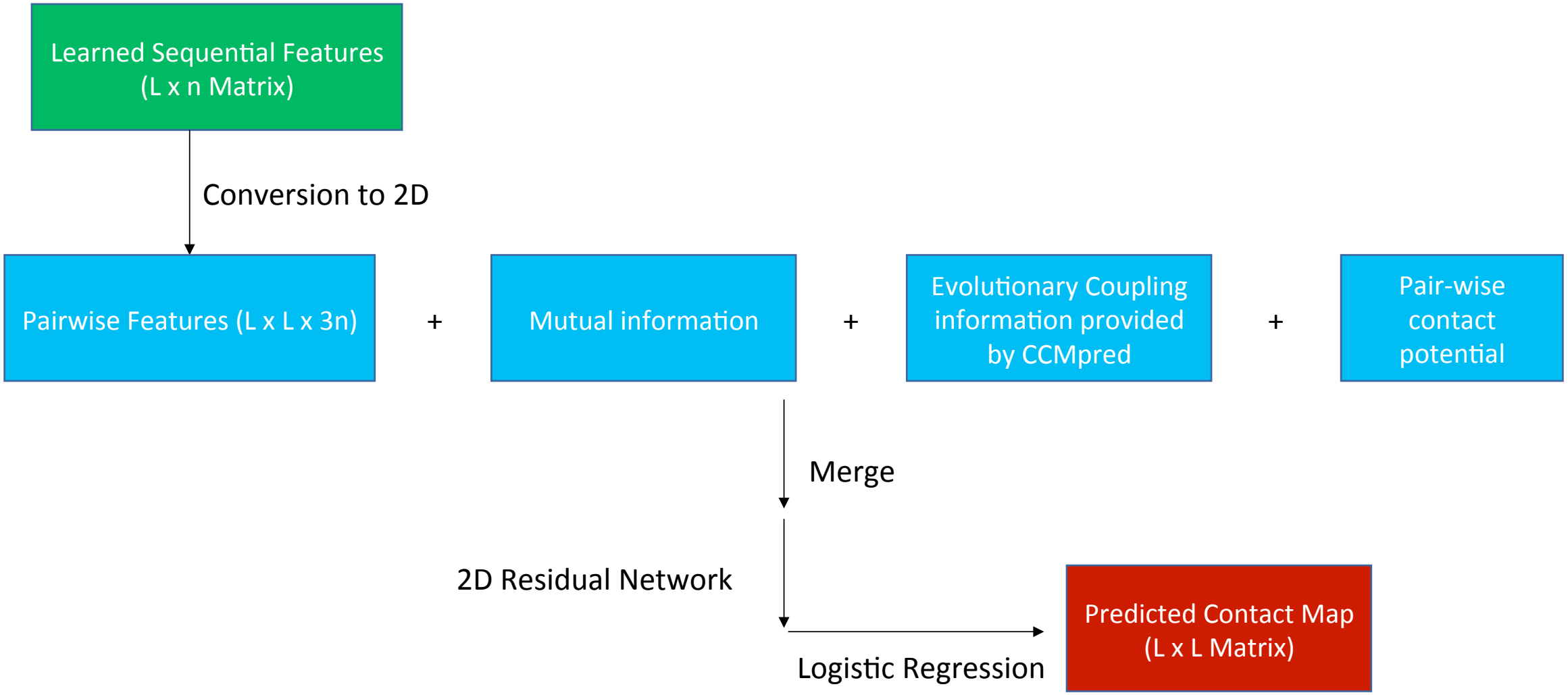
Deep Residual Learning for RaptorX Contact Prediction



Protein Features for the First Residual Network



Protein Features for Second Residual Network



~~Deep Residual Learning~~

Accuracy of Predicted Contact Maps

Training Set

- Subset of PDB25
- All proteins have less than 25% sequence identity with any other protein
- 6767 proteins
- Contains only ~100 membrane proteins

Test Set

- 150 Pfam families
- 105 CASP11 test proteins
- 76 hard CAMEO test proteins from 2015
- 398 membrane proteins
  - 400 residues at most
  - At most 40% sequence identity

Contact Prediction via Deep Residual Learning Has High Accuracy

Short sequence distance between two residues is in the range [6,11]
Medium sequence distance between two residues is in the range [12,23]
Long sequence distance between two residues is in the range ≥ 24

Table 1. Contact prediction accuracy on the 150 Pfam families.

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.50	0.40	0.26	0.17	0.64	0.52	0.34	0.22	0.74	0.68	0.53	0.39
PSICOV	0.58	0.43	0.26	0.17	0.65	0.51	0.32	0.20	0.77	0.70	0.52	0.37
CCMpred	0.65	0.50	0.29	0.19	0.73	0.60	0.37	0.23	0.82	0.76	0.62	0.45
plmDCA	0.66	0.50	0.29	0.19	0.72	0.60	0.36	0.22	0.81	0.76	0.61	0.44
Gremlin	0.66	0.51	0.30	0.19	0.74	0.60	0.37	0.23	0.82	0.76	0.63	0.46
MetaPSICOV	0.82	0.70	0.45	0.27	0.83	0.73	0.52	0.33	0.92	0.87	0.74	0.58
Our method	0.93	0.81	0.51	0.30	0.93	0.86	0.62	0.38	0.98	0.96	0.89	0.74

Accuracy is defined as the percent of the top L/k predicted contacts that correspond to native contacts where L is the length of the protein.

RaptorX Contact Prediction (bottom row) has higher accuracy than the compared methods for all sequence distance ranges.

Contact Prediction via Deep Residual Learning Has High Accuracy

Short sequence distance between two residues is in the range [6,11]
Medium sequence distance between two residues is in the range [12,23]
Long sequence distance between two residues is in the range ≥ 24

Table 2. Contact prediction accuracy on 105 CASP11 test proteins.

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.25	0.21	0.15	0.12	0.33	0.27	0.19	0.13	0.37	0.33	0.25	0.19
PSICOV	0.29	0.23	0.15	0.12	0.34	0.27	0.18	0.13	0.38	0.33	0.25	0.19
CCMpred	0.35	0.28	0.17	0.12	0.40	0.32	0.21	0.14	0.43	0.39	0.31	0.23
plmDCA	0.32	0.26	0.17	0.12	0.39	0.31	0.21	0.14	0.42	0.38	0.30	0.23
Gremlin	0.35	0.27	0.17	0.12	0.40	0.31	0.21	0.14	0.44	0.40	0.31	0.23
MetaPSICOV	0.69	0.58	0.39	0.25	0.69	0.59	0.42	0.28	0.60	0.54	0.45	0.35
Our method	0.82	0.70	0.46	0.28	0.85	0.76	0.55	0.35	0.81	0.77	0.68	0.55

Accuracy is defined as the percent of the top L/k predicted contacts that correspond to native contacts where L is the length of the protein.

RaptorX Contact Prediction (bottom row) has higher accuracy than the compared methods for all sequence distance ranges.

Contact Prediction via Deep Residual Learning Has High Accuracy

Short sequence distance between two residues is in the range [6,11]
Medium sequence distance between two residues is in the range [12,23]
Long sequence distance between two residues is in the range ≥ 24

Table 3. Contact prediction accuracy on 76 past CAMEO hard targets.

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.17	0.13	0.11	0.09	0.23	0.19	0.13	0.10	0.25	0.22	0.17	0.13
PSICOV	0.20	0.15	0.11	0.08	0.24	0.19	0.13	0.09	0.25	0.23	0.18	0.13
CCMpred	0.22	0.16	0.11	0.09	0.27	0.22	0.14	0.10	0.30	0.26	0.20	0.15
plmDCA	0.23	0.18	0.12	0.09	0.27	0.22	0.14	0.10	030	0.26	0.20	0.15
Gremlin	0.21	0.17	0.11	0.08	0.27	0.22	0.14	0.10	0.31	0.26	0.20	0.15
MetaPSICOV	0.56	0.47	0.31	0.20	0.53	0.45	0.32	0.22	0.47	0.42	0.33	0.25
Our method	0.67	0.57	0.37	0.23	0.69	0.61	0.42	0.28	0.69	0.65	0.55	0.42

Accuracy is defined as the percent of the top L/k predicted contacts that correspond to native contacts where L is the length of the protein.

RaptorX Contact Prediction (bottom row) has higher accuracy than the compared methods for all sequence distance ranges.

Contact Prediction via Deep Residual Learning Has High Accuracy

Short sequence distance between two residues is in the range [6,11]
Medium sequence distance between two residues is in the range [12,23]
Long sequence distance between two residues is in the range ≥ 24

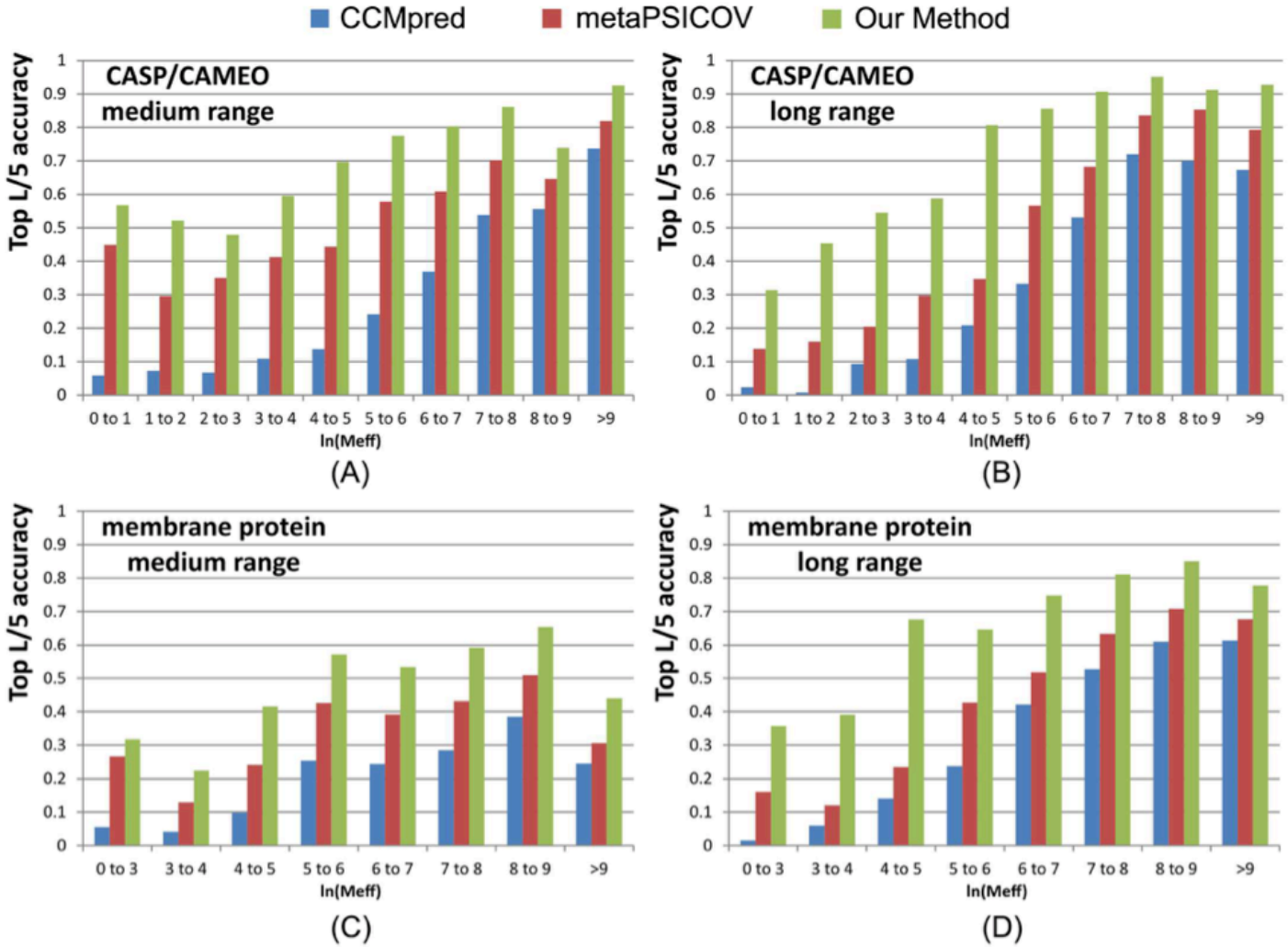
Table 4. Contact prediction accuracy on 398 membrane proteins.

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.16	0.13	0.09	0.07	0.28	0.22	0.13	0.09	0.44	0.37	0.26	0.18
PSICOV	0.22	0.16	0.10	0.07	0.29	0.21	0.13	0.09	0.42	0.34	0.23	0.16
CCMpred	0.27	0.19	0.11	0.08	0.36	0.26	0.15	0.10	0.52	0.45	0.31	0.21
plmDCA	0.26	0.18	0.11	0.08	0.35	0.25	0.14	0.09	0.51	0.42	0.29	0.20
Gremlin	0.27	0.19	0.11	0.07	0.37	0.26	0.15	0.10	0.52	0.45	0.32	0.21
MetaPSICOV	0.45	0.35	0.22	0.14	0.49	0.40	0.27	0.18	0.61	0.55	0.42	0.30
Our method	0.60	0.46	0.27	0.16	0.66	0.53	0.33	0.22	0.78	0.73	0.62	0.47

Accuracy is defined as the percent of the top L/k predicted contacts that correspond to native contacts where L is the length of the protein.

RaptorX Contact Prediction (bottom row) has higher accuracy than the compared methods for all sequence distance ranges.

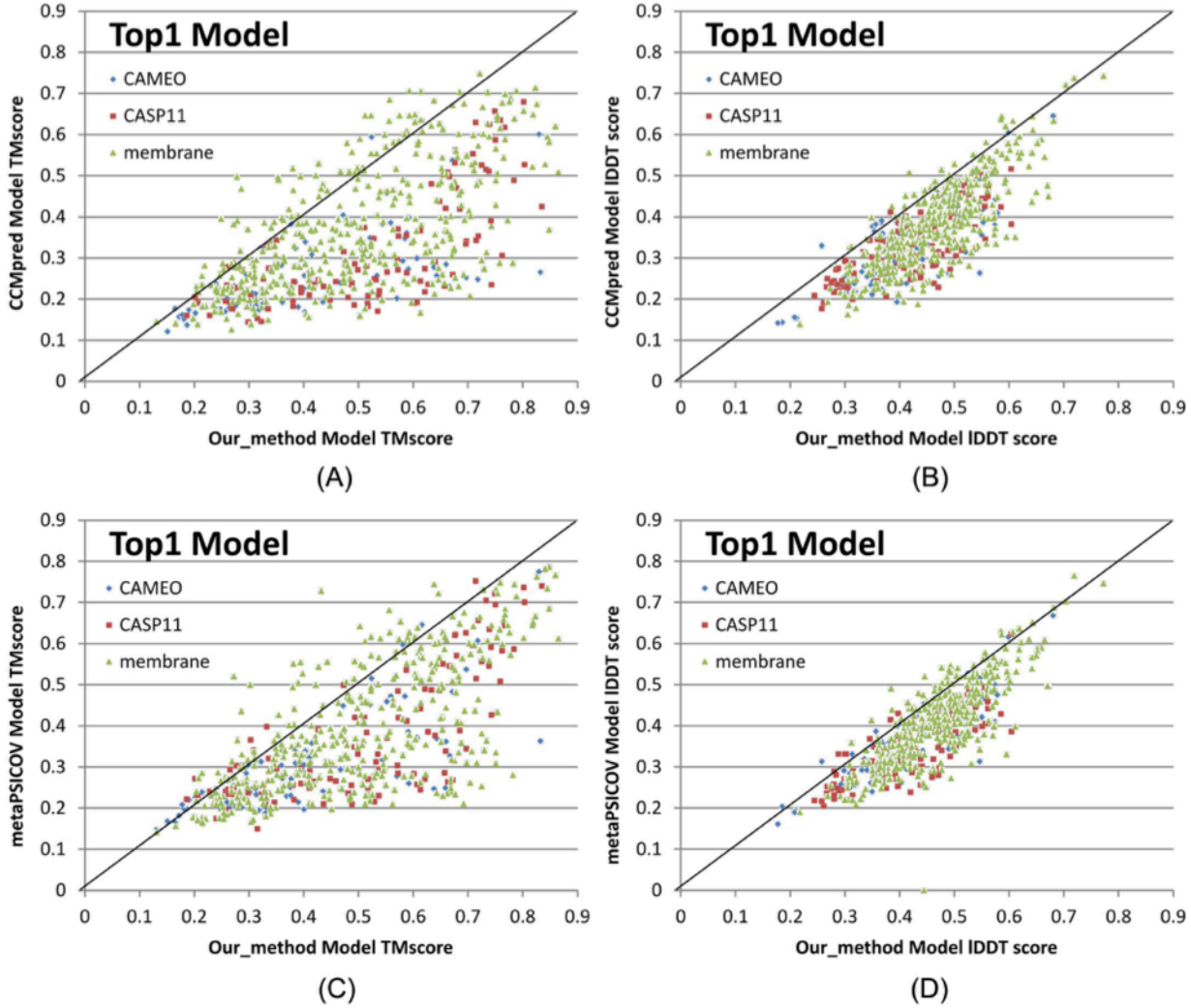
Increased Performance Due to Deep Residual Learning is Independent of Available Homologous Information



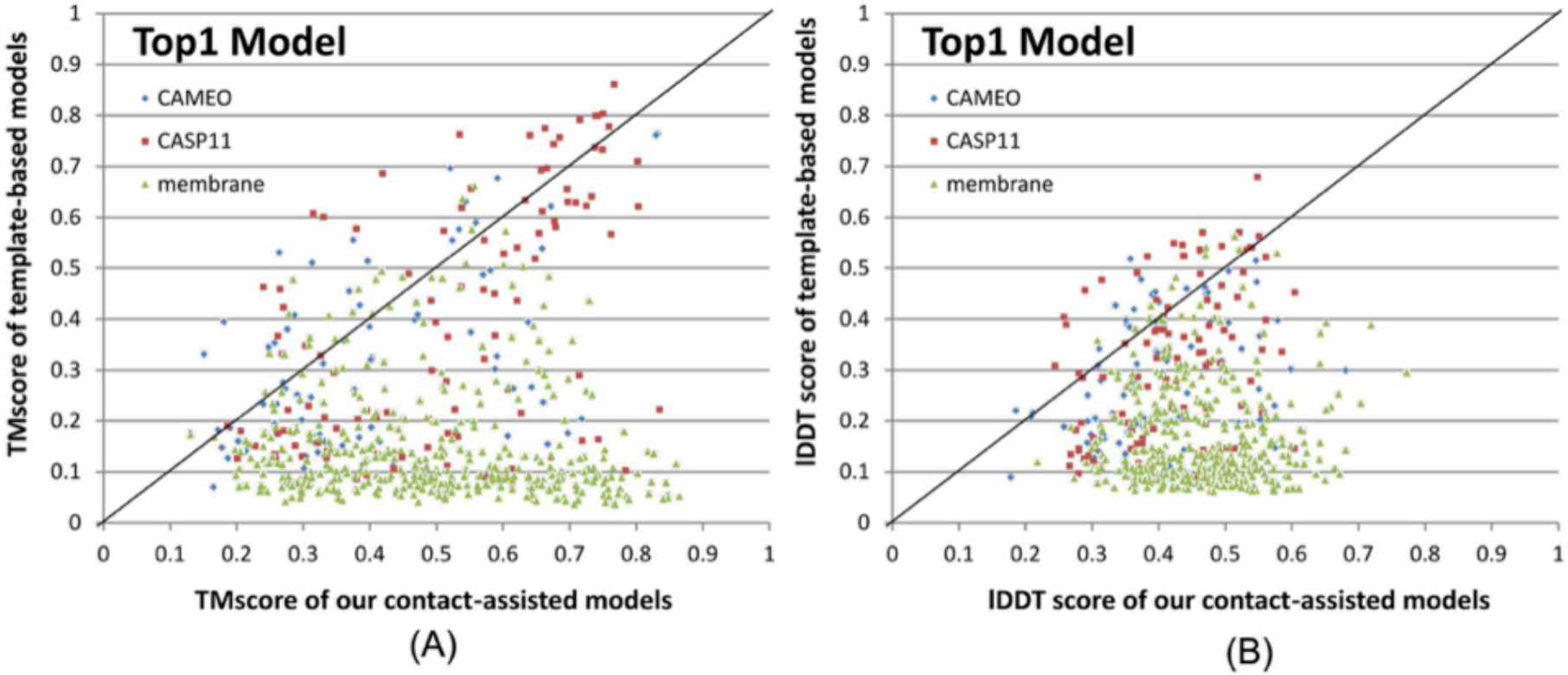
~~Accuracy of Predicted Contact Maps~~

Contact-Assisted Protein Folding Results

Contact-Assisted Protein Folding Benefits from Deep Residual Learning



Learned Features Are Not Template-Based



~~Contact-Assisted Protein Folding Results~~

CAMEO Blind Tests

CAMEO Blind Tests of Contact Prediction

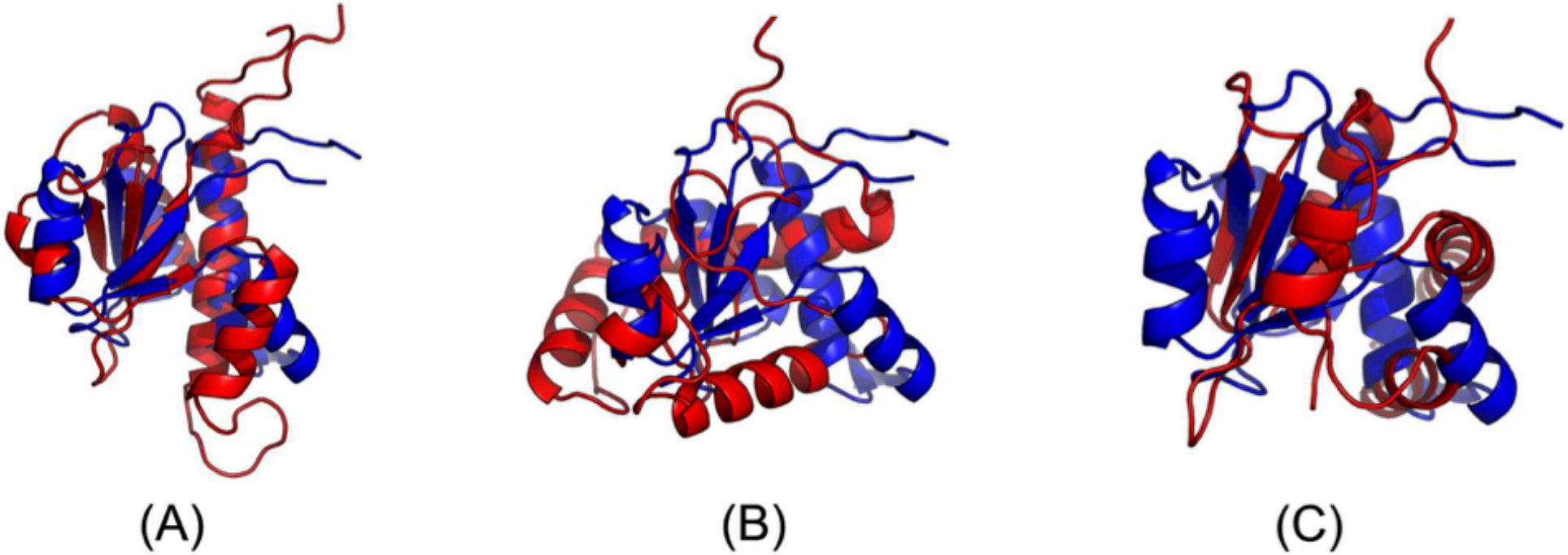


Fig 10. Superimposition between the predicted models (red) and the native structure (blue) for the CAMEO test protein (PDB ID 5dcj and chain A). The models are built by CNS from the contacts predicted by **(A)** our method, **(B)** CCMpred, and **(C)** MetaPSICOV.

CAMEO Blind Tests of Contact Prediction

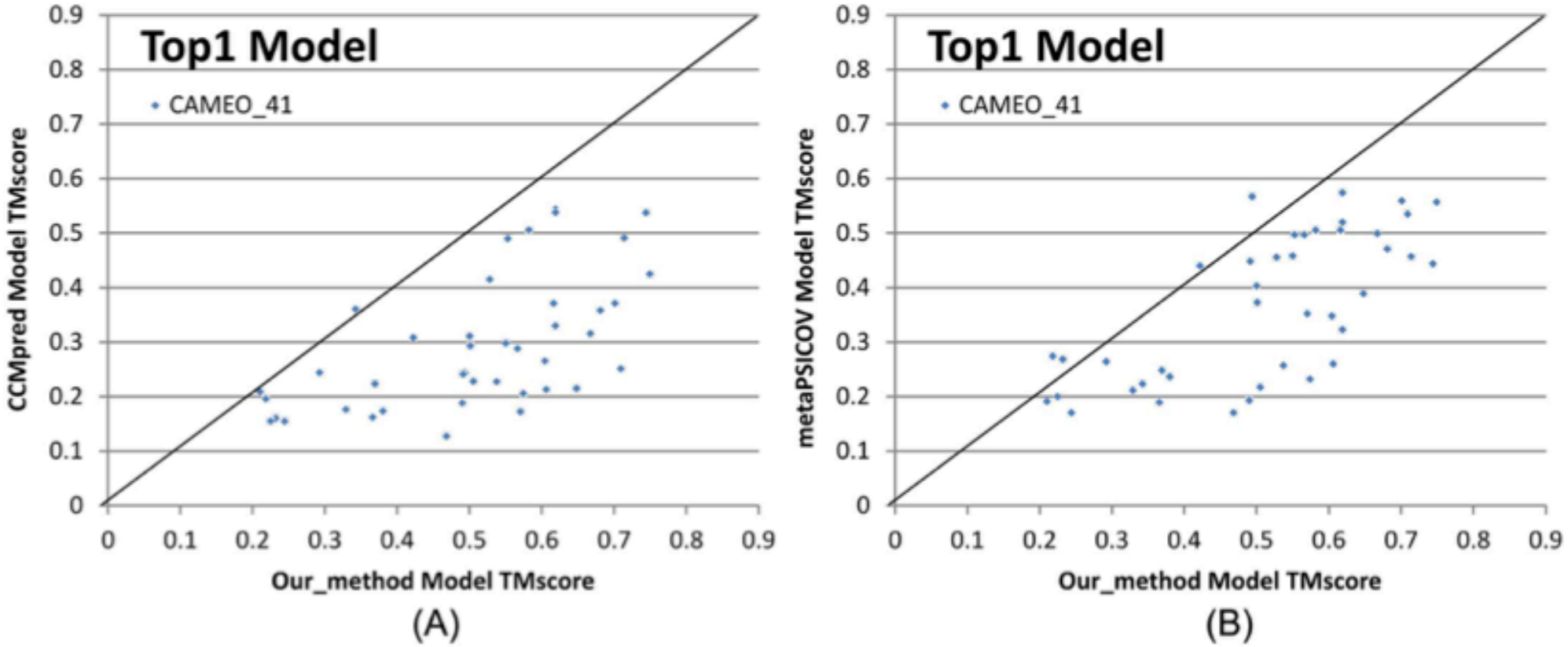


Fig 5. Quality comparison (measured by TMscore) of contact-assisted models generated by our server, CCMpred and MetaPSICOV on the 41 CAMEO hard targets. (A) our server (X-axis) vs. CCMpred and (B) our server (X-axis) vs. MetaPSICOV.

~~CAMEO Blind Tests~~

Strengths/Weaknesses

Strengths

- Very thorough in comparing its predictions against different types of proteins and prediction approaches.
- Uses a non-redundant training set.
- Considers all residue pairs for contact simultaneously.
- Blind testing through CAMEO.
- Performs surprisingly well on membrane proteins.

Weaknesses

- Use of extensive hidden layers makes learned features difficult to describe.
- Does not quantify its false-positive rate.
- Is not as unique an approach as implied (see PConsC2).
- Does not compare its method of contact map prediction to that of PConsC2.
- Tested membrane proteins were constrained

Supervised Machine Learning Incorporates More Context

OPEN ACCESS Freely available online

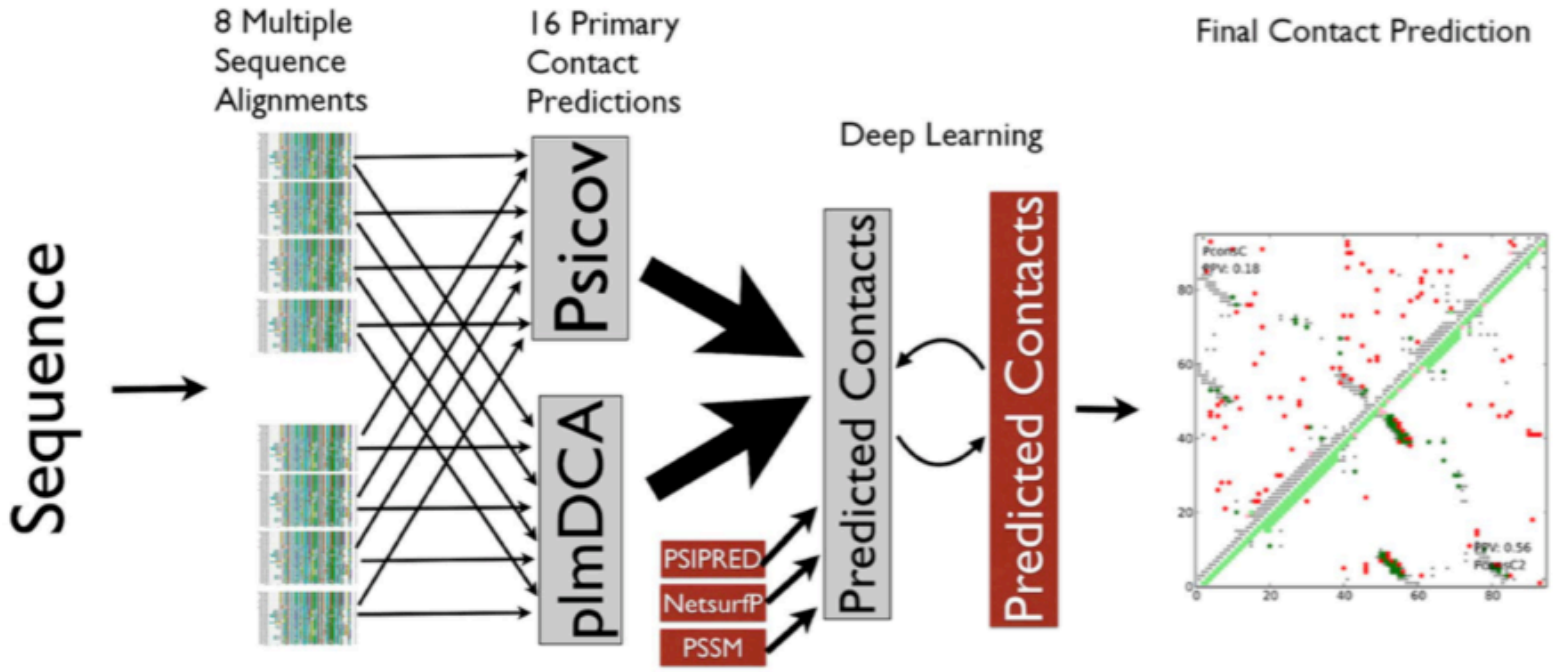
PLOS | COMPUTATIONAL BIOLOGY

Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns

Marcin J. Skwark[1,2,3], Daniele Raimondi[1,2,4], Mirco Michel[1,2], Arne Elofsson[1,2]*

1 Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, 2 Science for Life Laboratory, Stockholm University, Solna, Sweden, 3 Department of Information and Computer Science, Aalto University, Aalto, Finland, 4 Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, La Plaine Campus, Triomflaan, Brussels, Belgium

PconsC2 Contact Prediction



Test Set

- 150 Pfam families
- 105 CASP11 test proteins
- 76 hard CAMEO test proteins from 2015
- 398 membrane proteins
  - 400 residues at most
  - At most 40% sequence identity