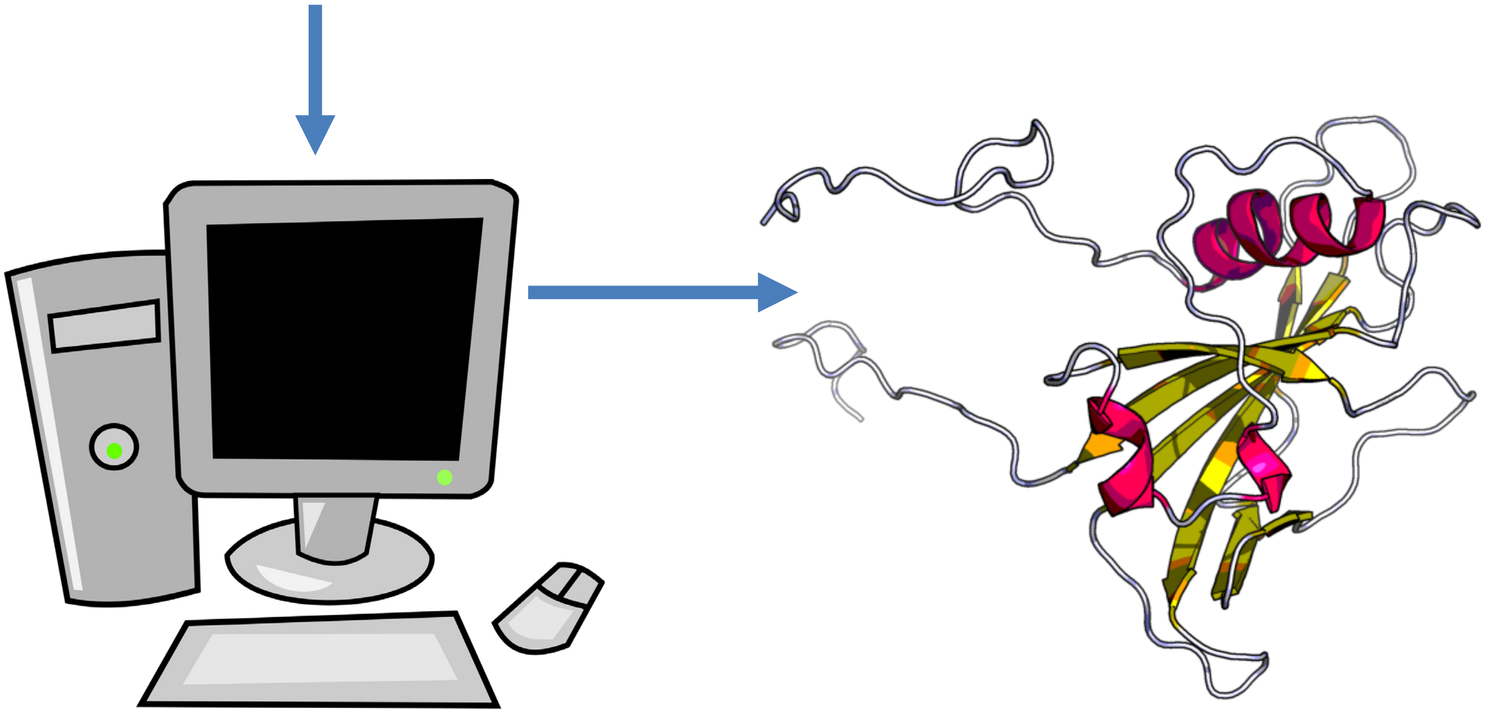
Protein structure prediction by deep learning

Jinbo Xu

Toyota Technological Institute at Chicago

Protein Structure Prediction

MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK



Has been studied several decades

Methods for Protein Structure Prediction

Template-based modeling

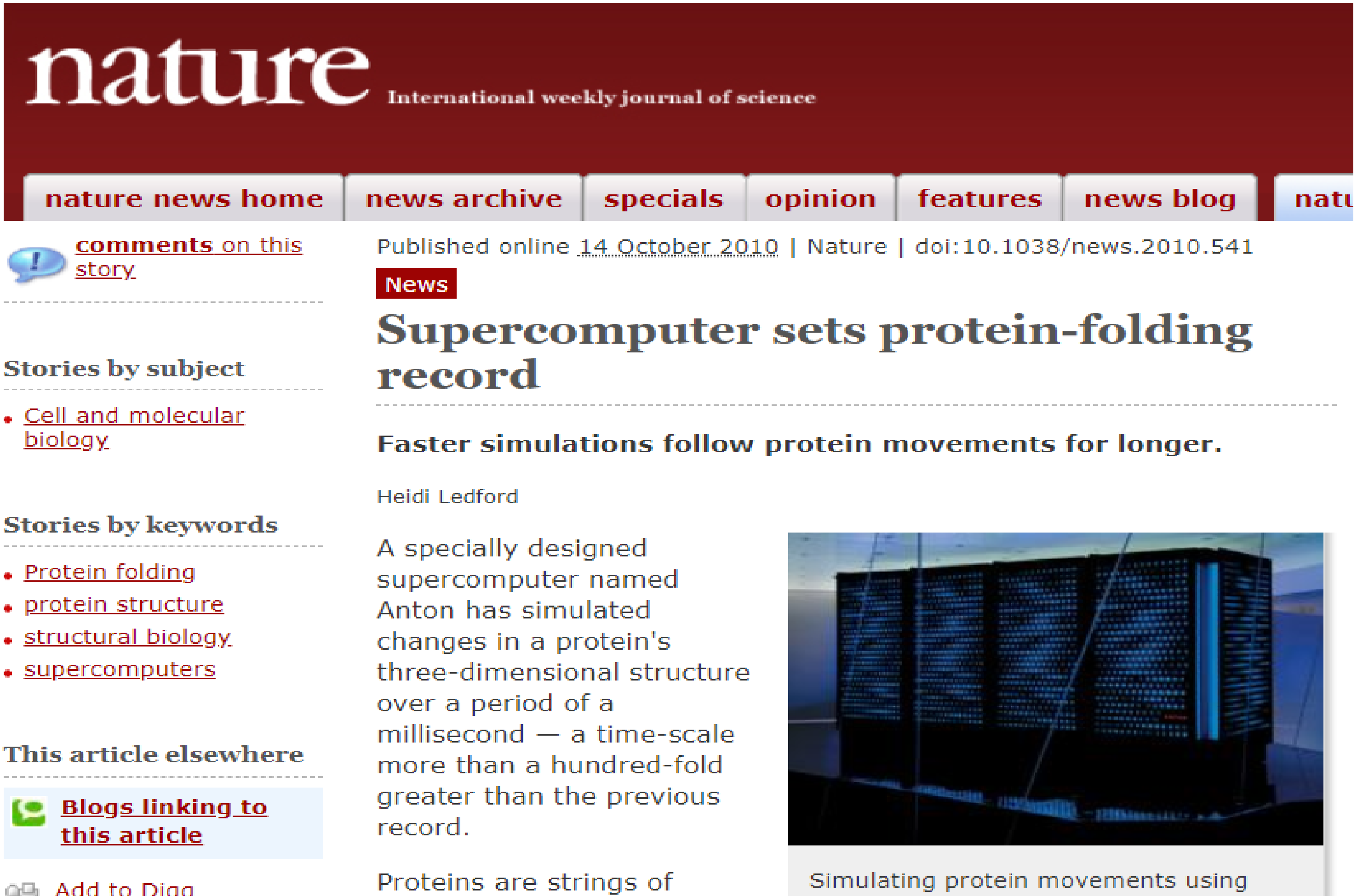
- Use a solved structure as template to model a protein under prediction
- WAS the 1st choice for protein modeling
- Not effective for some proteins, e.g., membrane proteins

Template-free modeling

- Previously used when no templates in PDB
- Now used unless very good templates in PDB

State of the Art Until 2015

- A lot of computing power needed
- Success rate is low even for small proteins



Published online 14 October 2010 | Nature | doi:10.1038/news.2010.541

News

Supercomputer sets protein-folding record

Faster simulations follow protein movements for longer.

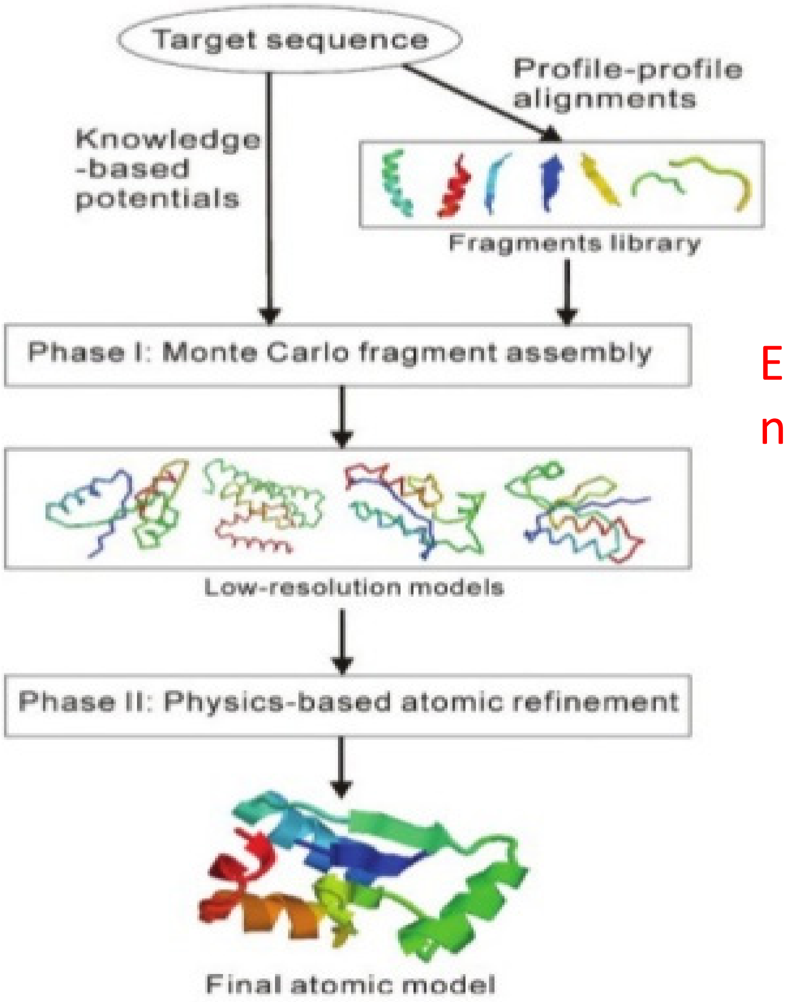
Heidi Ledford

A specially designed supercomputer named Anton has simulated changes in a protein's three-dimensional structure over a period of a millisecond — a time-scale more than a hundred-fold greater than the previous record.

Proteins are strings of

Simulating protein movements using

Old method: fragment assembly

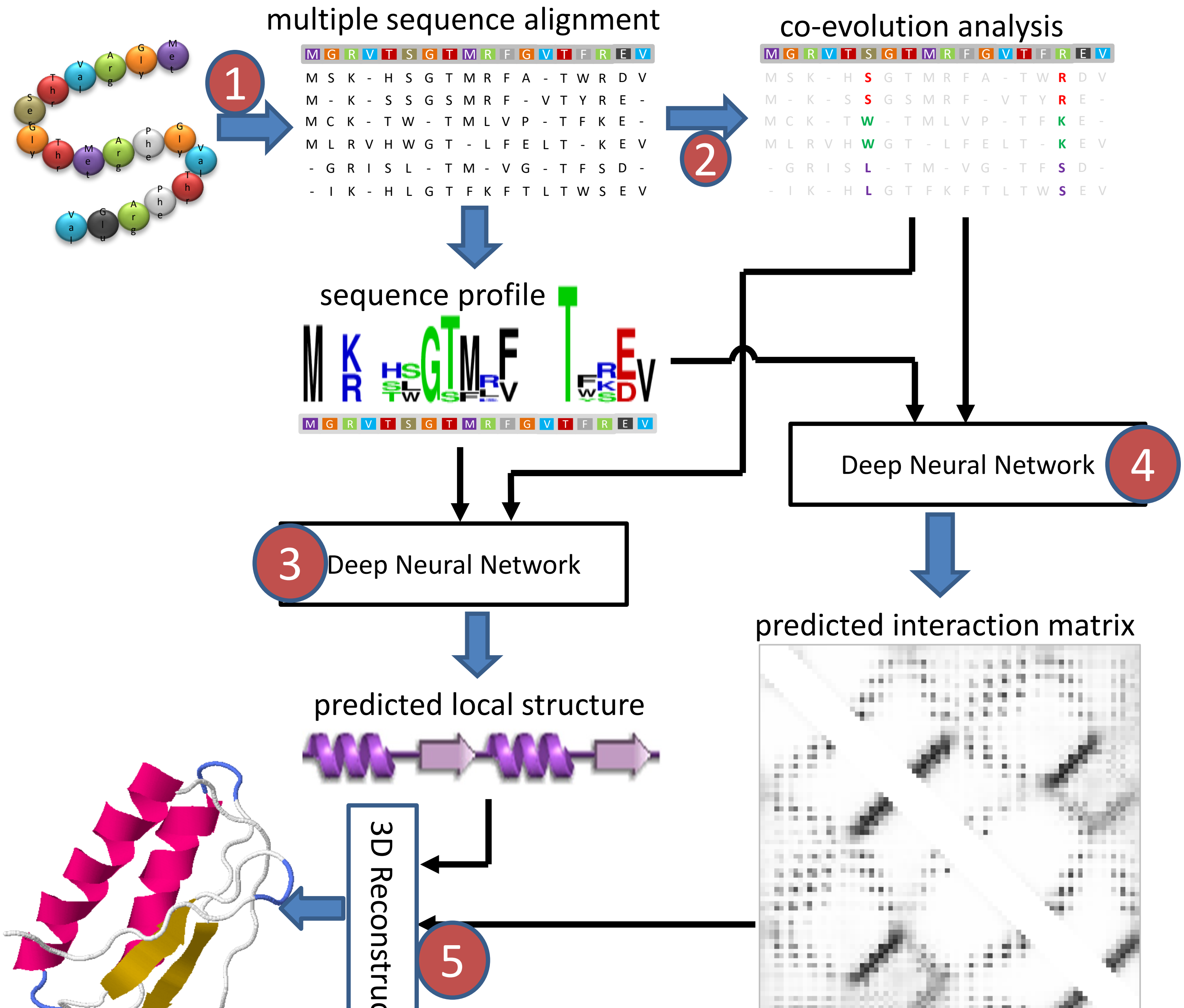

Target sequence
Profile-profile alignments
Knowledge -based potentials
Fragments library
Phase I: Monte Carlo fragment assembly
Low-resolution models
Phase II: Physics-based atomic refinement
Final atomic model

Extensive conformation sampling needed even for a small protein

Fig.: Flowchart of the ROSETTA protocol

https://www.slideshare.net/ag1805x/ab-initio-protein-structure-prediction

New Strategy



multiple sequence alignment

co-evolution analysis

sequence profile

Deep Neural Network

Deep Neural Network

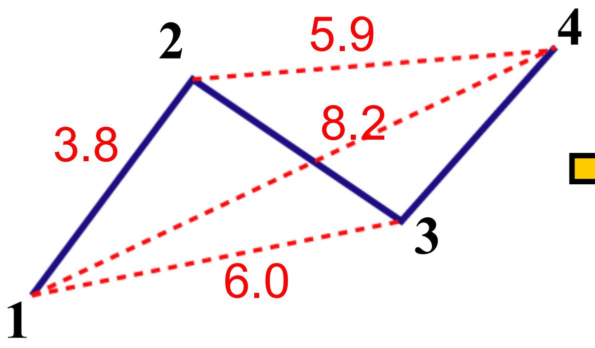
predicted interaction matrix

predicted local structure

3D Reconstruc

1
2
3
4
5
M G R V T S G T M R F G V T F R E V
M S K - H S G T M R F A - T W R D V
M - K - S S G S M R F - V T Y R E -
M C K - T W - T M L V P - T F K E -
M L R V H W G T - L F E L T - K E V
- G R I S L - T M - V G - T F S D -
- I K - H L G T F K F T L T W S E V
M G R V T S G T M R F G V T F R E V
M S K - H S G T M R F A - T W R D V
M - K - S S G S M R F - V T Y R E -
M C K - T W - T M L V P - T F K E -
M L R V H W G T - L F E L T - K E V
- G R I S L - T M - V G - T F S D -
- I K - H L G T F K F T L T W S E V
M G R V T S G T M R F G V T F R E V
Met Gly Arg Val Thr Ser Gly Thr Met Arg Phe Gly Val Thr Phe Arg Glu Val

Protein Distance & Contact Matrix



$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 3.8 & 6.0 & 8.2 \\ 2 & 3.8 & 0 & 3.8 & 5.9 \\ 3 & 6.0 & 3.8 & 0 & 3.8 \\ 4 & 8.2 & 5.9 & 3.8 & 0 \end{array}$$

Distance matrix

Discretize

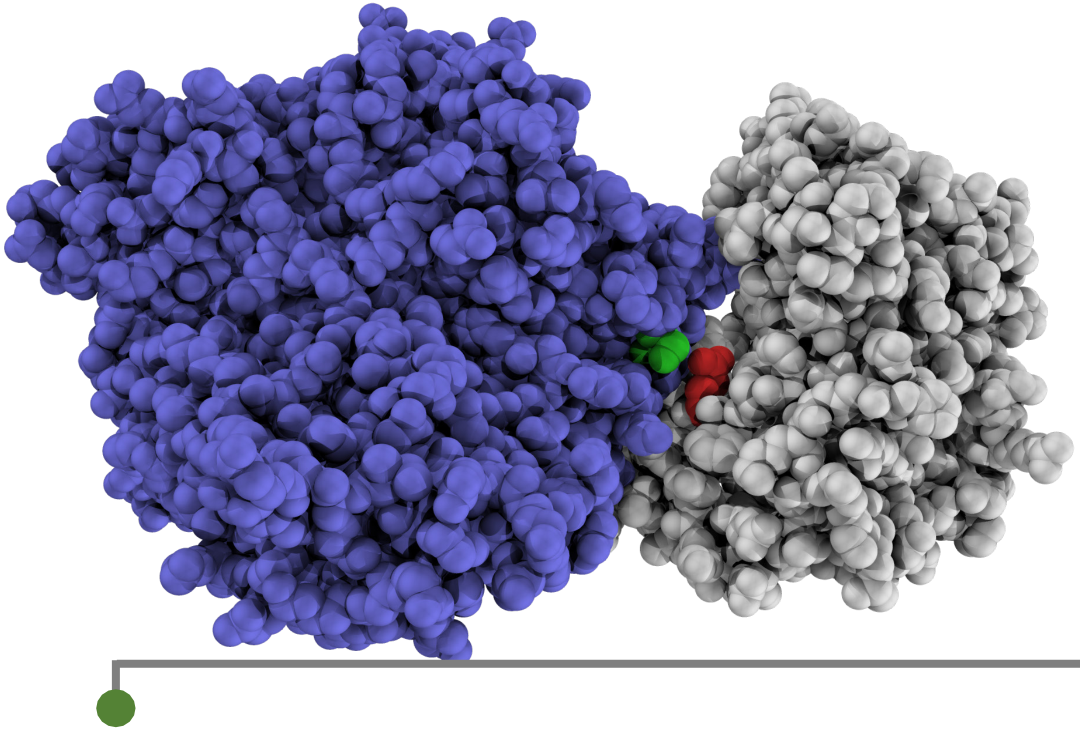
$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 1 & 1 & 0 & 1 \\ 4 & 0 & 1 & 1 & 0 \end{array}$$

Contact matrix

Ideas that work

- Global statistical method for co-evolution analysis (in ~2008)
- Deep convolutional residual neural network (ResNet) for prediction of residue interaction, i.e., contact/distance prediction (in ~2016)
- Transformer-like network for residue-residue interaction prediction (in ~2020)

Amino acids in direct physical contact tend to covary or "coevolve" across related proteins



For example, a mutation that causes one amino acid to get bigger is more likely to preserve protein structure and function (and thus survive) if another amino acid gets smaller to make space

...GANPMHGRDQ**S**GAVASLTSVA...
...GANPMHGRDQ**E**GAVASLTSVA...
...GANPMHGRDE**K**GAVASLTSVG...
...GANPMHGRDS**H**GWLASCLSVA...
...GANPMNGRDV**K**GFVAAGASVA...
...GANPMHGRDR**D**GAVASLTSVA...
...GANPMHGRDQ**V**GAVASLTSVA...
...GANPMHGRDQ**E**GAVASLTSVA...

...VEDLMK**E**VVTYRHFMNASGG...
...VEALMA**R**VLSYRHFMNASGG...
...VATVMK**Q**VMTYRHYLRATGG...
...VARAMR**E**IGKYAQVLKISRG...
...VPELMQ**D**LTSYRHFMNASGG...
...ADHVLR**R**LSDFVPALLPLGG...
...FERART**A**LEAYAAPLRAMGG...
...VPEVMK**K**VMSYRHYLKATGG...

Co-evolution Detection

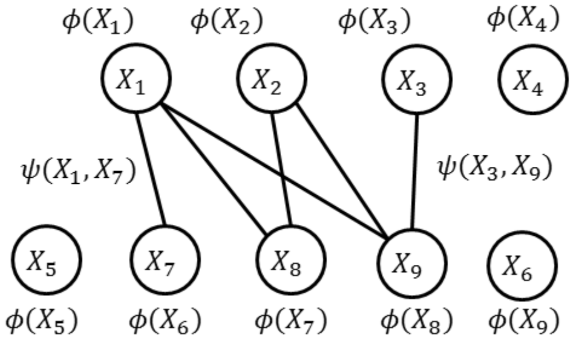
- Mutual Information does NOT work well
  - If amino acid A in contact with B and B in contact with C, then A and C probably is correlated even if they are NOT in direct contact
  - Not any two correlated amino acids are in contact
- Global statistical methods
  - Direct coupling analysis (DCA): find a set of direct contacts that may explain all observed correlation patterns

Co-evolution Analysis:

Model MSA by Graphical Model

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
G	R	K	A	Y	S	A	B	V
G	R	K	-	Y	S	A	B	A
F	L	V	-	L	Y	I	V	A
K	L	V	-	L	Y	I	V	A

MSA



MRF

The generating probability of a sequence S:

$$P(S) = \frac{1}{Z} \prod_i \phi(X_i) \prod_{(i,k)} \psi(X_i, X_k)$$

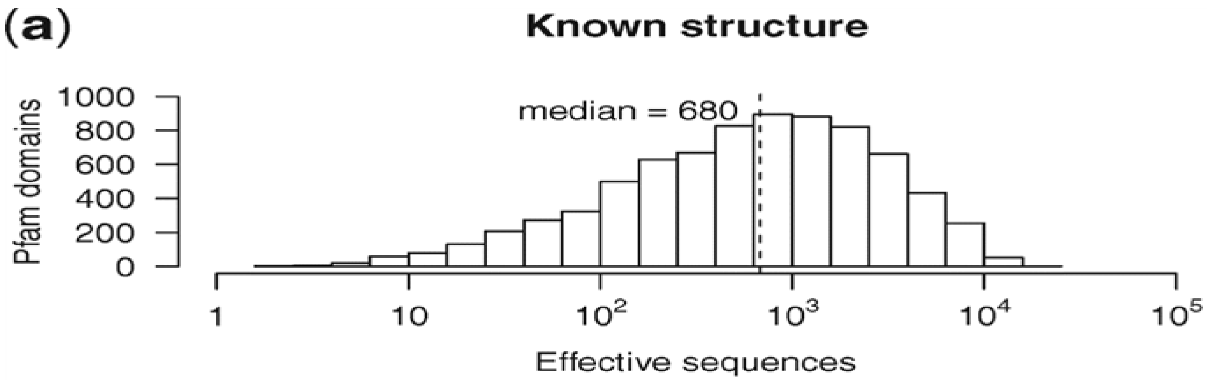
1. ψ encodes residue correlation relationship
2. Infer ϕ, ψ by maximum- or pseudo-likelihood
3. A special case is Gaussian Graphical Model

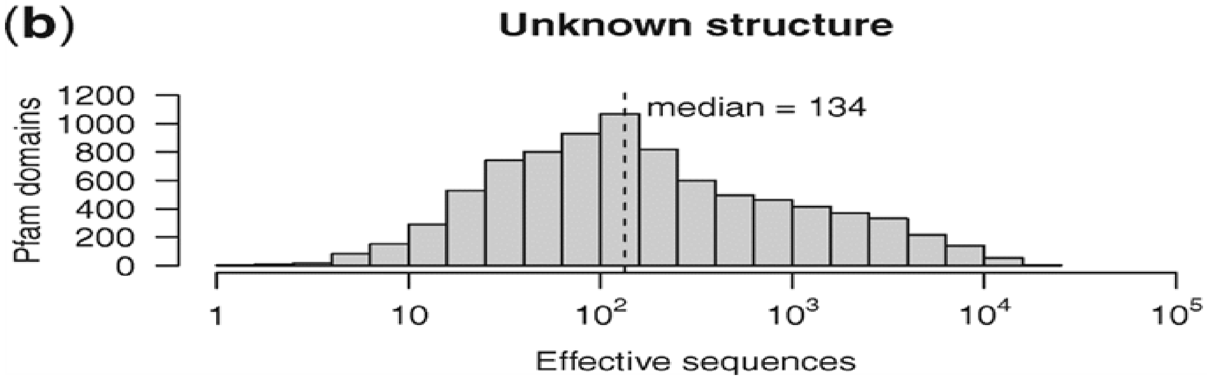
References for Global Methods

- Martin Weigt et al. *Identification of direct residue contacts in protein–protein interaction by message passing*. PNAS, 2008
- Burger Lukas and van Nimwegen Erik. *Disentangling direct from indirect co-evolution of residues in protein alignments*. *PLoS Comput Biol*, 6(1):e1000633, 2010. Bayesian network model
- Christopher Langmead group. *Learning generative models for protein fold families*. PROTEINS 2010. (pseudo-likelihood)
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. *Protein 3D Structure Computed from Evolutionary Sequence Variation*. PLoS ONE, 2011. (maximum-entropy)
- Jones group. *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments*. Bioinformatics, 2012. (maximum-likelihood)

Co-evolution Method Limit

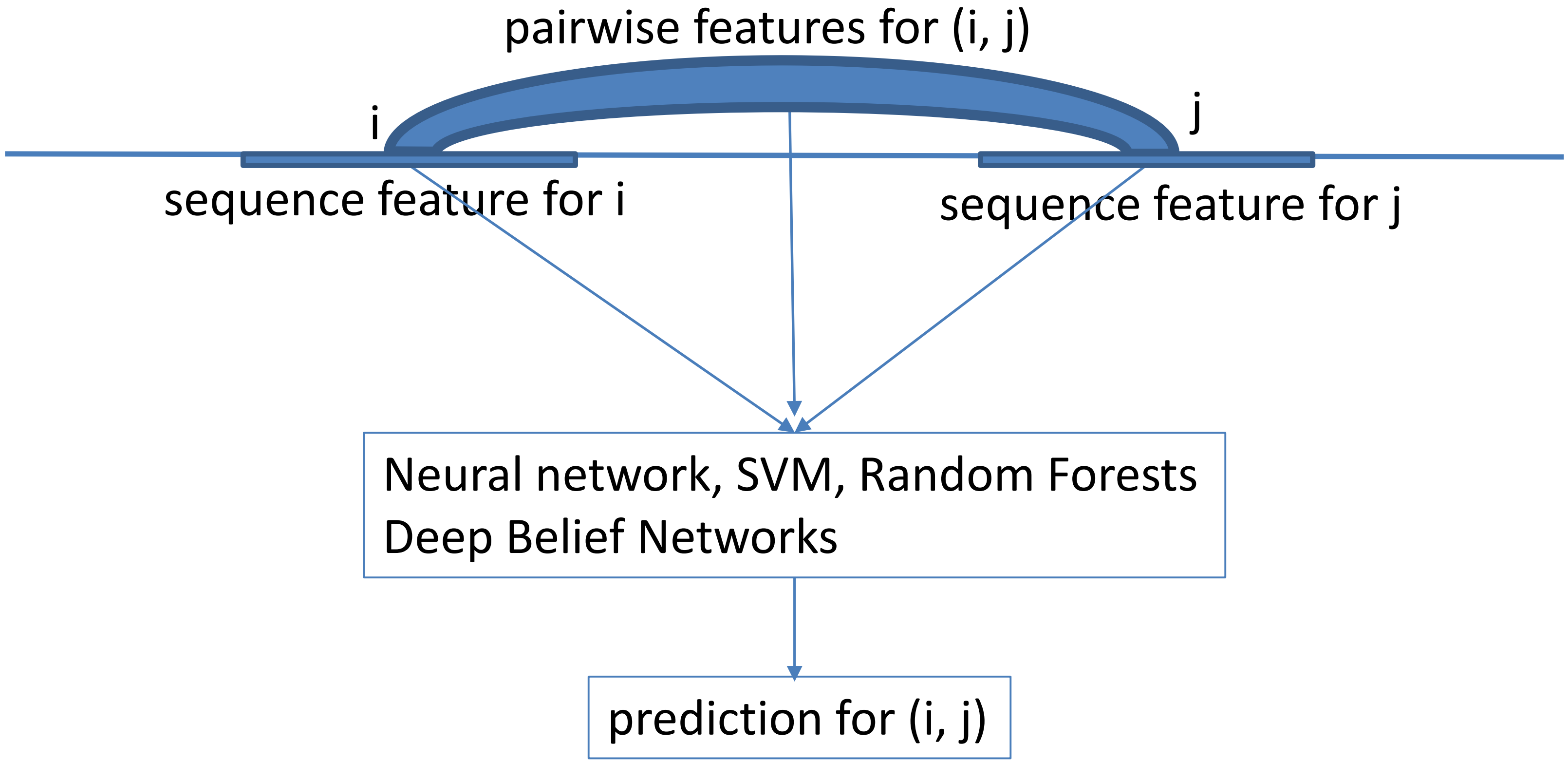
Needs a large number of non-redundant sequence homologs





Picture taken from Bioinformatics, Elofsson 2017

Previous Supervised Learning Methods



pairwise features for (i, j)

sequence feature for i

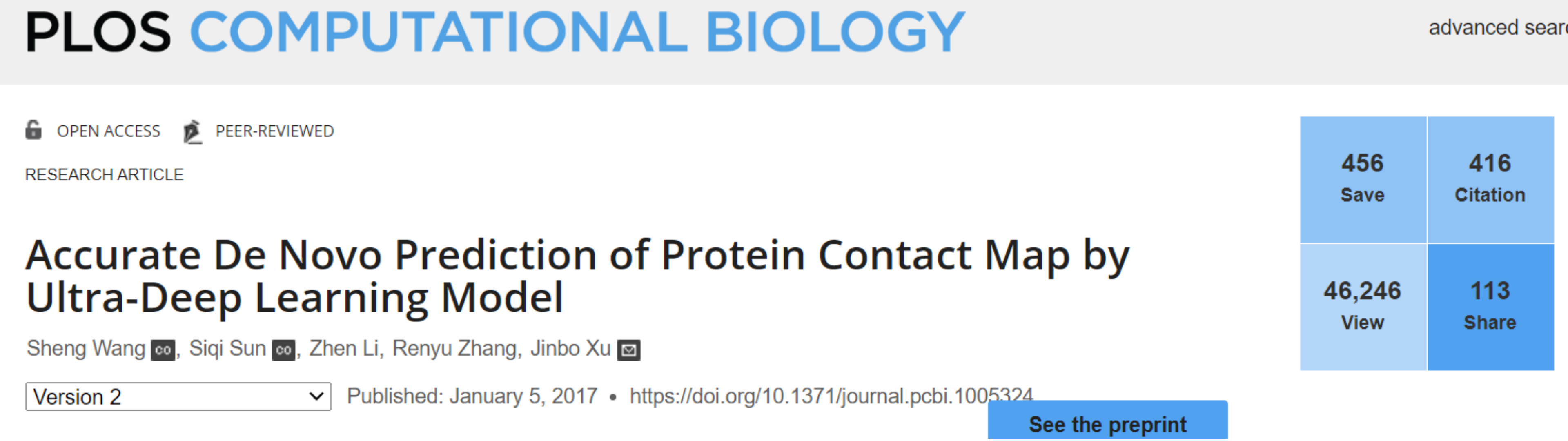
sequence feature for j

Neural network, SVM, Random Forests
Deep Belief Networks

prediction for (i, j)

Key issue: ignore the impact of all other residues

Fully Deep Convolutional Residual Neural Network



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

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

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

Version 2

Published: January 5, 2017 • https://doi.org/10.1371/journal.pcbi.1005324

456 Save

416 Citation

46,246 View

113 Share

First time show contact prediction improved by DL

1. Predicted contacts may fold hard targets in CAMEO
2. Work for membrane proteins (Cell Systems, 2017)
3. Work for complex contact prediction (NAR, 2018)

2018 PLoS CB Research Prize in Breakthrough/Innovation

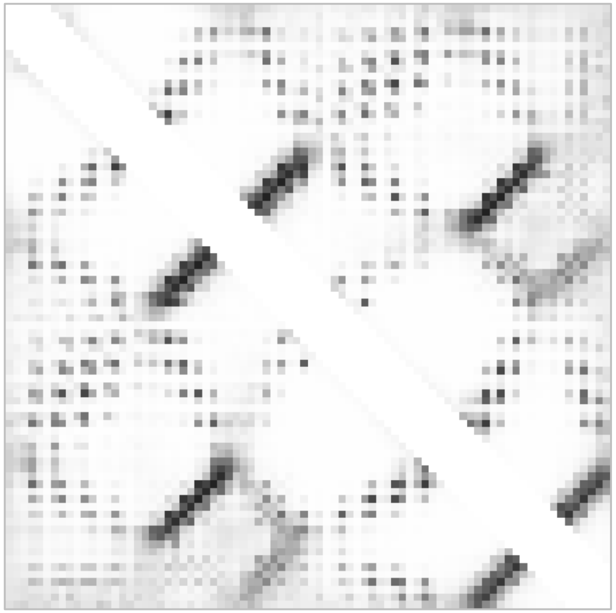
Key Idea for Protein Contact Prediction

1. Model atomic interaction map as an image, each atom pair as a pixel

2. Predict all contacts of a protein simultaneously

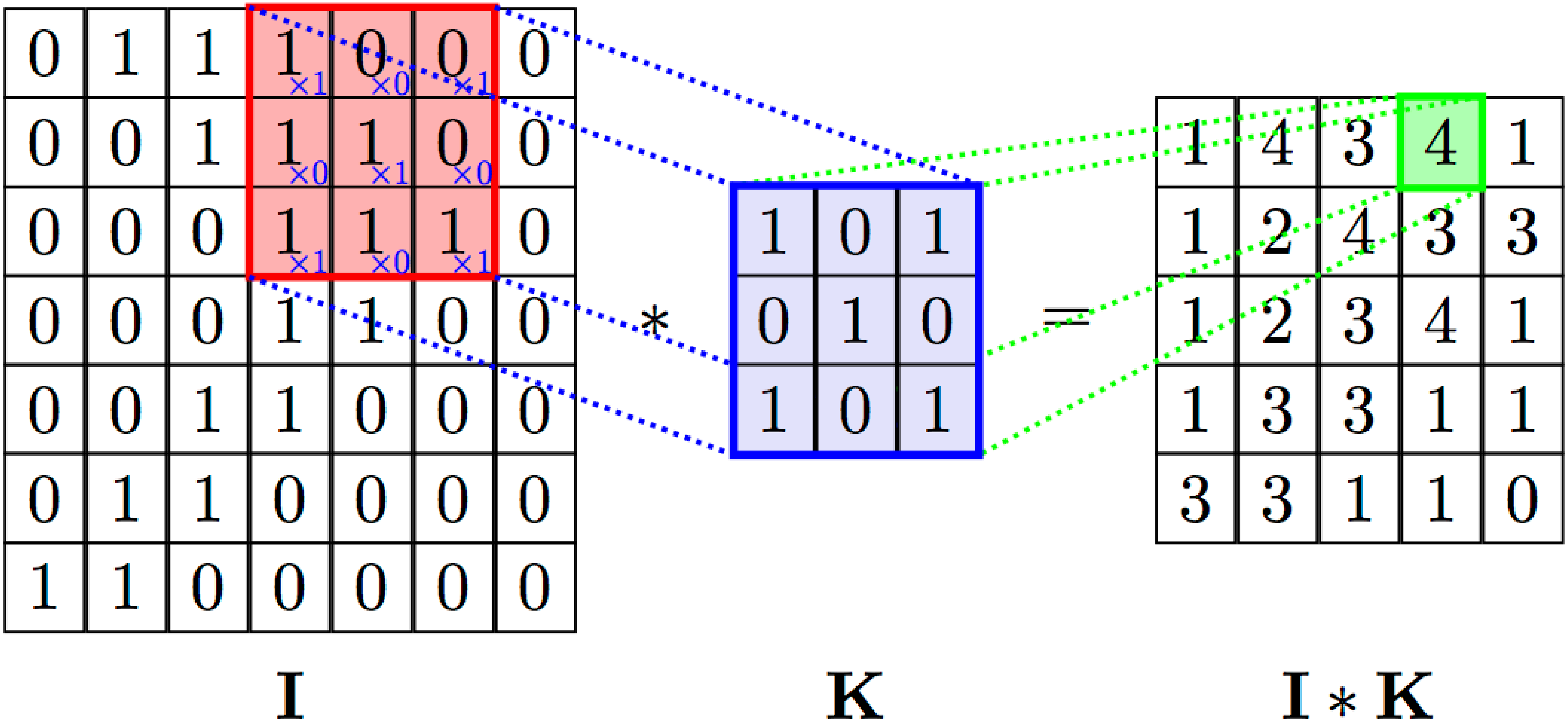
3. Formulate the problem as an image semantic segmentation problem

4. Apply very deep convolution neural network



What’s Convolution?

Pattern Detection & Information Collection



0	1	1	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	1	1	0
0	0	0	1	1	0	0
0	0	1	1	0	0	0
0	1	1	0	0	0	0
1	1	0	0	0	0	0

$\mathbf{I}$

1	0	1
0	1	0
1	0	1

$\mathbf{K}$

1	4	3	4	1
1	2	4	3	3
1	2	3	4	1
1	3	3	1	1
3	3	1	1	0

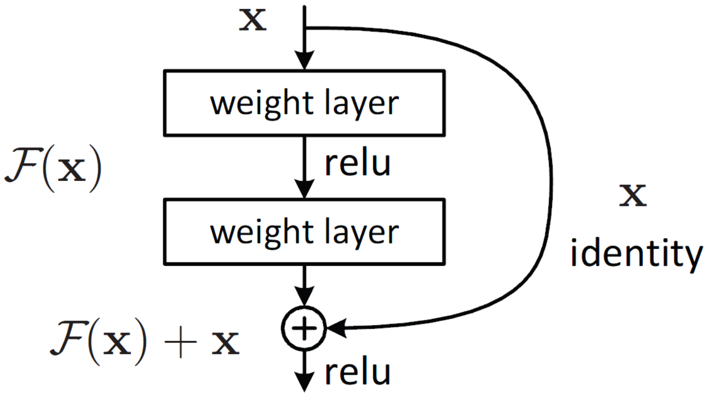
$\mathbf{I} * \mathbf{K}$

What’s Residual ?

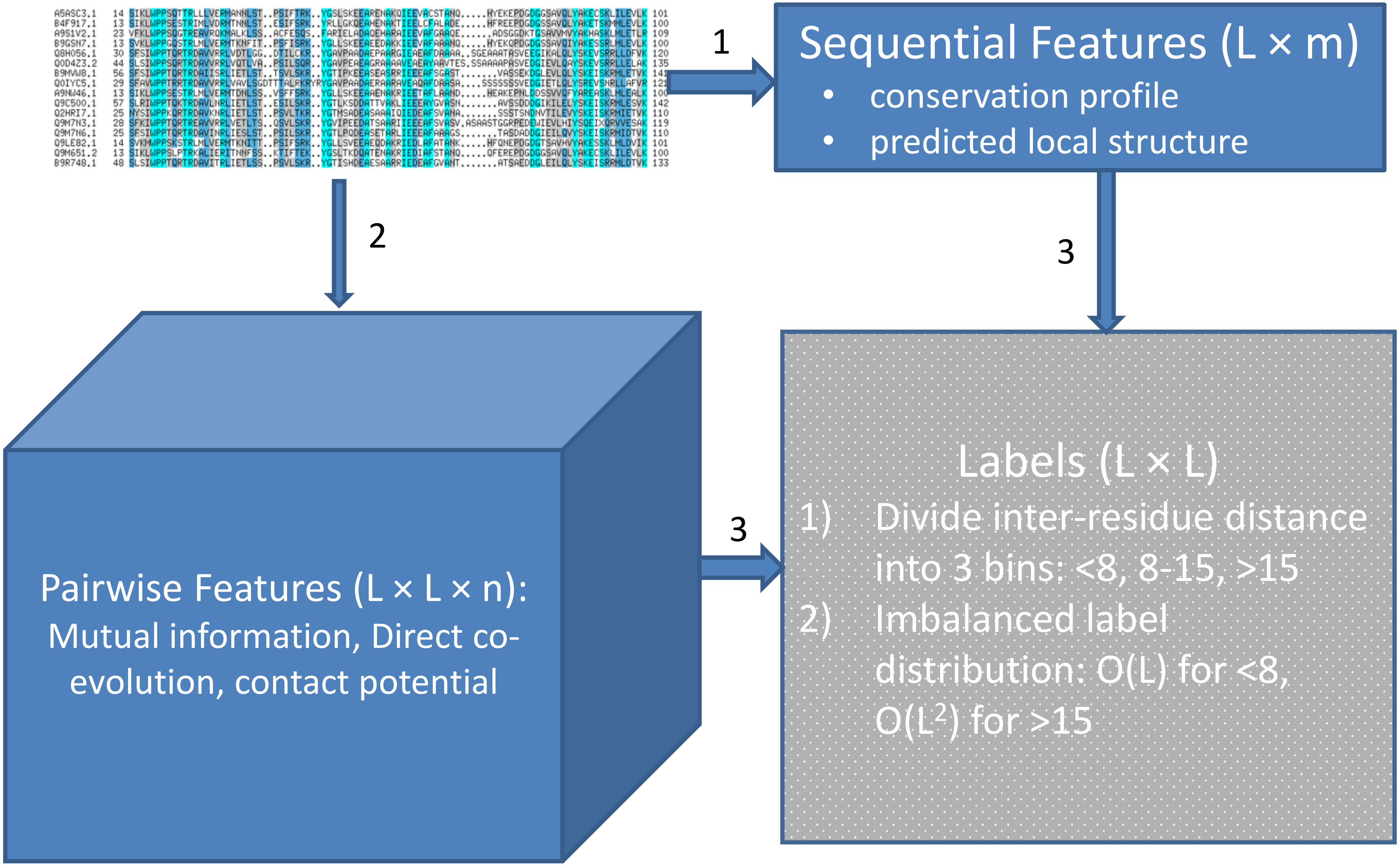
Convolutional network is old, but residual network invented in 2015

- To predict y from x, first predict y-x from x
- Add x and predicted y-x to get y
- y-x is the residual
- Equivalent to add shortcut between x and y

- This makes it easy to stack many convolutional layers together to form a very deep network



Protein Features and Labels



```
A5ASC3.1  14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEARENAKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK 101
B4F917.1  13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKTIEELCFALADE.....HFREEPDGDGSSAVQLYAKETSKMMLEVLK 100
A9S1V2.1  23 VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQEHARAIEEVAFGAAQE......ADSGGDKTGSAVVMVYAKHASKLMLETLR 109
B9GSN7.1  13 SVKLWPPGQSTRLMLVERMTKNFIT..PSFISRK..YGLLSKEEAEEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLMLEVLK 100
Q8H056.1  30 SFSIWPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPAADAEPAARGIEAEAFDAAAA..SGEAAATASVEEGIKALQLYSKEVSRRLLDFVK 120
Q0D4Z3.2  44 SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAEAGRAAAAVEAEAYAAVTES.SSAAAAPASVEDGIEVLQAYSKEVSRRLLELAK 135
B9MVW8.1  56 SFSIWPPTQRTRDAIISRLIETLST..TSVLSKR..YGTIPKEEASEASRRIEEEAFSGAST.......VASSEKDGLEVLQLYSKEISKRMLETVK 141
Q0IYC5.1  29 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALRKRYRYGAVPAADAERAARAVEAQAFDAASA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1  13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C500.1  57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1  25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTMSADEASAAAIQIEDEAFSVANA.......SSSTSNDNVTILEVYSKEISKRMIETVK 110
Q9M7N3.1  28 SFKIWPPTQRTREAVVRRLVETLTS..QSVLSKR..YGVIPEEDATSAARIIEEEAFSVASV.ASAASTGGRPEDEWIEVLHIYSQEIXQRVVESAK 119
Q9M7N6.1  25 SFSIWPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQDEASETARLIEEEAFAAAGS.......TASDADDGIEILQVYSKEISKRMIDTVK 110
Q9LE82.1  14 SVKMWPPSKSTRLMLVERMTKNITT..PSIFSRK..YGLLSVEEAEQDAKRIEDLAFATANK.....HFQNEPDGDGTSAVHVYAKESSKLMLDVIK 101
Q9M651.2  13 SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGSLTKDQATENAKRIEDIAFSTANQ.....QFEREPDGDGGSAVQLYAKECSKLILEVLK 100
B9R748.1  48 SLSIWPPTQRTRDAVITRLIETLSS..PSVLSKR..YGTISHDEAESAARRIEDEAFGVANT.......ATSAEDDGLEILQLYSKEISRRMLDTVK 133
```

Sequential Features (L × m)

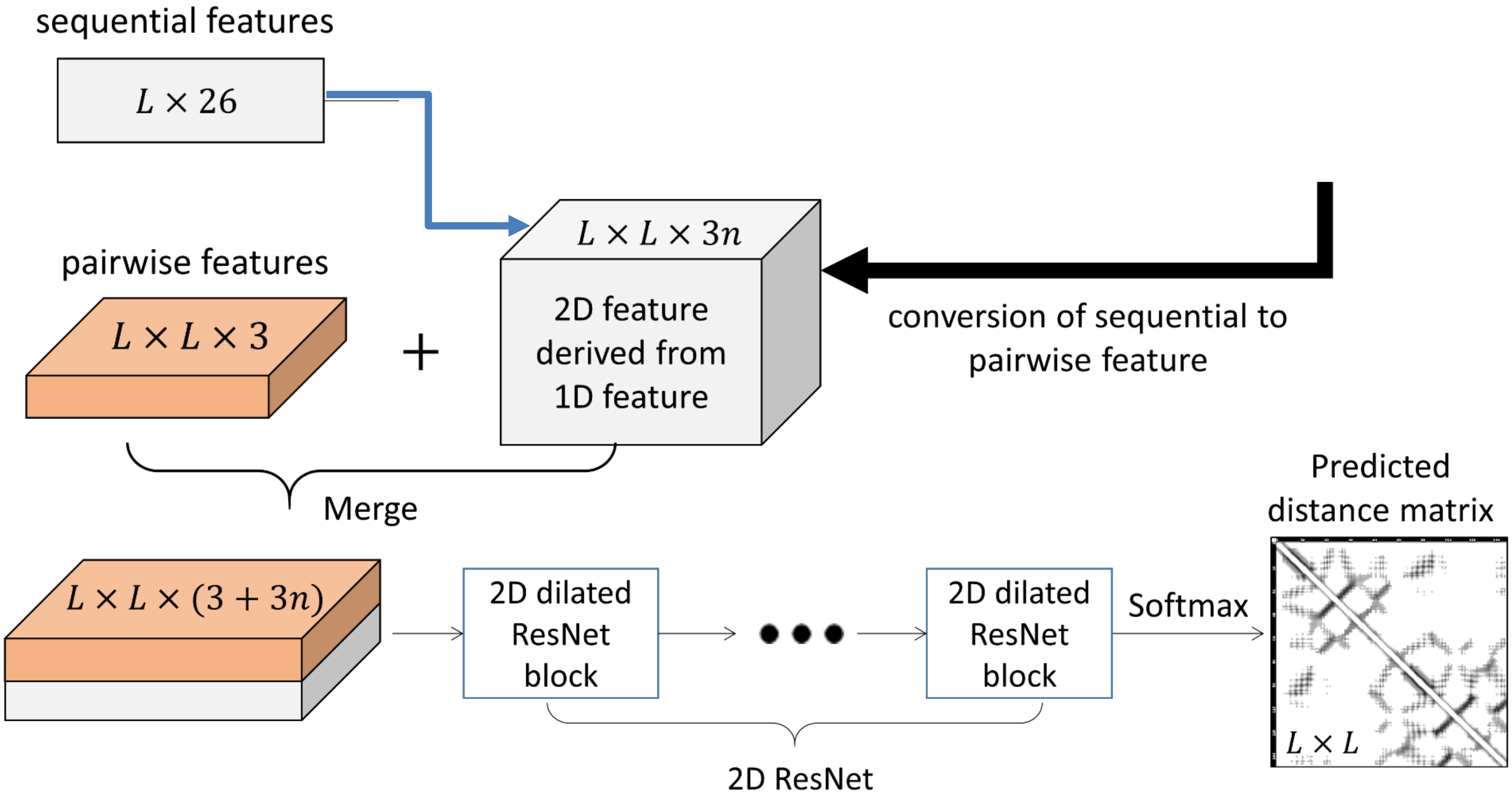
- conservation profile
- predicted local structure

Pairwise Features (L × L × n):
Mutual information, Direct co-evolution, contact potential

Labels (L × L)

1) Divide inter-residue distance into 3 bins: <8, 8-15, >15
2) Imbalanced label distribution: O(L) for <8, O(L^2) for >15

Deep ResNet for Contact/Distance Prediction



sequential features

$L \times 26$

pairwise features

$L \times L \times 3$

+

$L \times L \times 3n$

2D feature derived from 1D feature

conversion of sequential to pairwise feature

Merge

$L \times L \times (3 + 3n)$

2D dilated ResNet block

2D dilated ResNet block

2D ResNet

Softmax

Predicted distance matrix

$L \times L$

CASP12 Ranking of Contact Prediction

(Schaarschmidt et al, PROTEINS 2017)

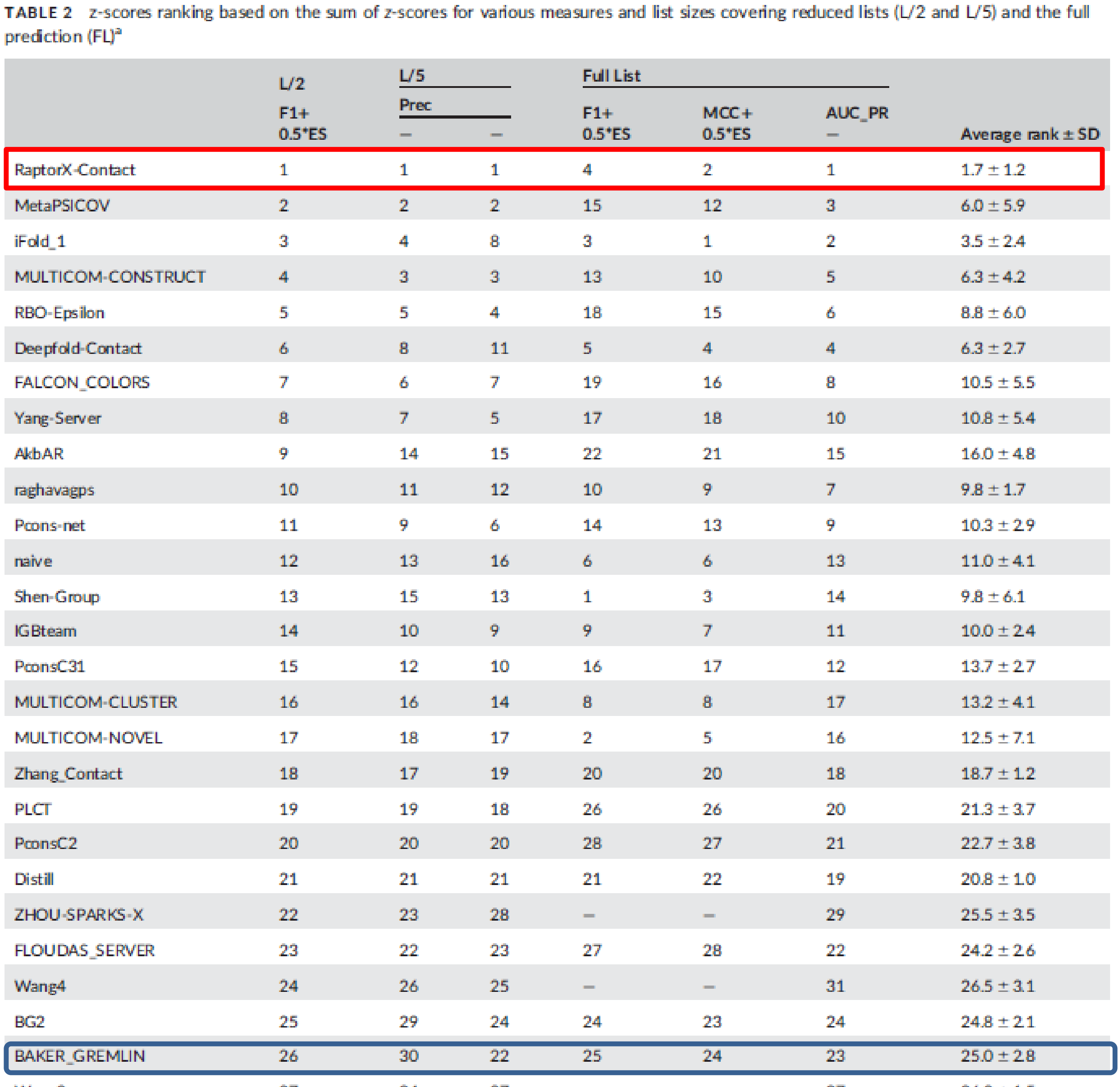


TABLE 2 z-scores ranking based on the sum of z-scores for various measures and list sizes covering reduced lists (L/2 and L/5) and the full prediction (FL)[a]

	L/2	L/5		Full List			
	F1+ 0.5*ES	Prec		F1+ 0.5*ES	MCC+ 0.5*ES	AUC_PR	
		—	—			—	Average rank ± SD
RaptorX-Contact	1	1	1	4	2	1	1.7 ± 1.2
MetaPSICOV	2	2	2	15	12	3	6.0 ± 5.9
iFold_1	3	4	8	3	1	2	3.5 ± 2.4
MULTICOM-CONSTRUCT	4	3	3	13	10	5	6.3 ± 4.2
RBO-Epsilon	5	5	4	18	15	6	8.8 ± 6.0
Deepfold-Contact	6	8	11	5	4	4	6.3 ± 2.7
FALCON_COLORS	7	6	7	19	16	8	10.5 ± 5.5
Yang-Server	8	7	5	17	18	10	10.8 ± 5.4
AkbAR	9	14	15	22	21	15	16.0 ± 4.8
raghavagps	10	11	12	10	9	7	9.8 ± 1.7
Pcons-net	11	9	6	14	13	9	10.3 ± 2.9
naive	12	13	16	6	6	13	11.0 ± 4.1
Shen-Group	13	15	13	1	3	14	9.8 ± 6.1
IGBteam	14	10	9	9	7	11	10.0 ± 2.4
PconsC31	15	12	10	16	17	12	13.7 ± 2.7
MULTICOM-CLUSTER	16	16	14	8	8	17	13.2 ± 4.1
MULTICOM-NOVEL	17	18	17	2	5	16	12.5 ± 7.1
Zhang_Contact	18	17	19	20	20	18	18.7 ± 1.2
PLCT	19	19	18	26	26	20	21.3 ± 3.7
PconsC2	20	20	20	28	27	21	22.7 ± 3.8
Distill	21	21	21	21	22	19	20.8 ± 1.0
ZHOU-SPARKS-X	22	23	28	—	—	29	25.5 ± 3.5
FLOUDAS_SERVER	23	22	23	27	28	22	24.2 ± 2.6
Wang4	24	26	25	—	—	31	26.5 ± 3.1
BG2	25	29	24	24	23	24	24.8 ± 2.1
BAKER_GREMLIN	26	30	22	25	24	23	25.0 ± 2.8

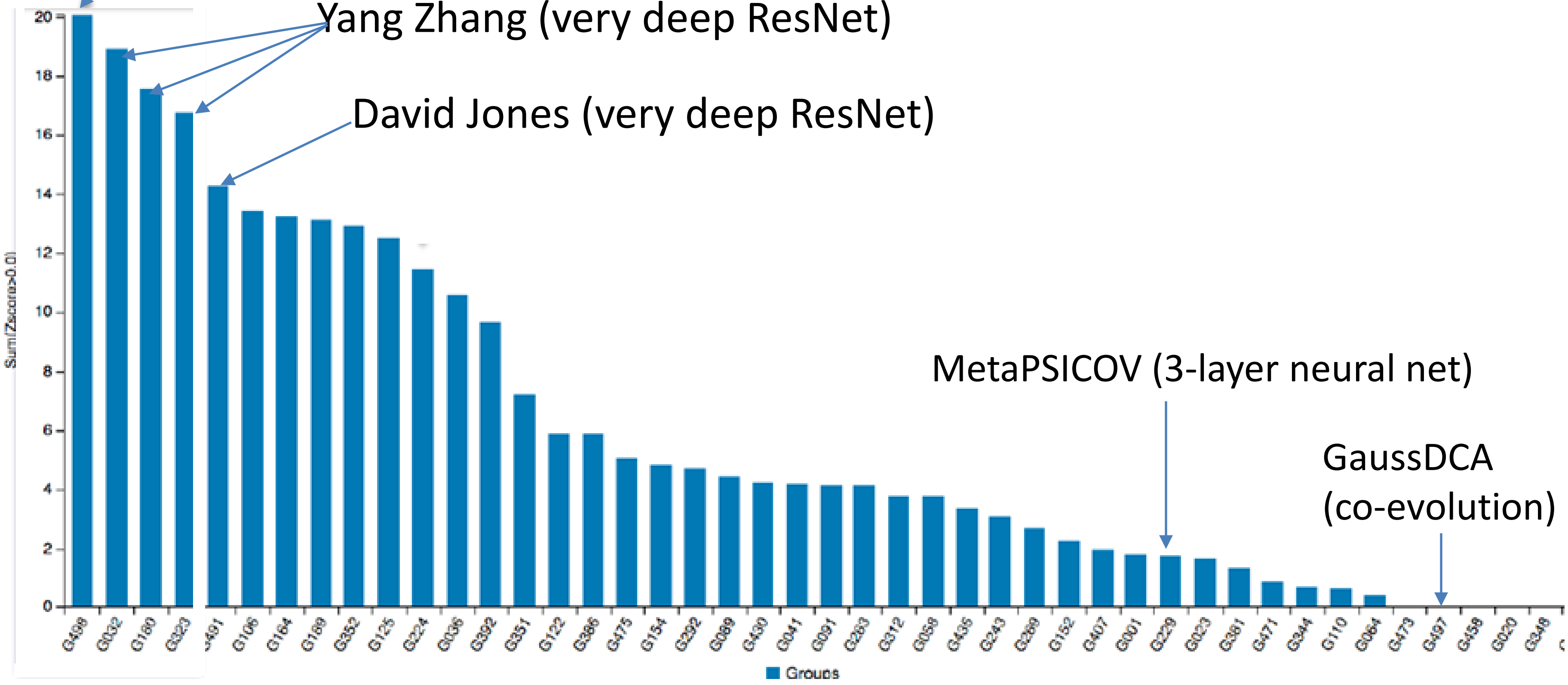
Very Deep ResNet

Deep CNN by Peng et al

Co-evolution

CASP13 Ranking of Contact Prediction

(Courtesy of Dr. Andras Fiser)



RaptorX (our server, very deep ResNet)

Yang Zhang (very deep ResNet)

David Jones (very deep ResNet)

MetaPSICOV (3-layer neural net)

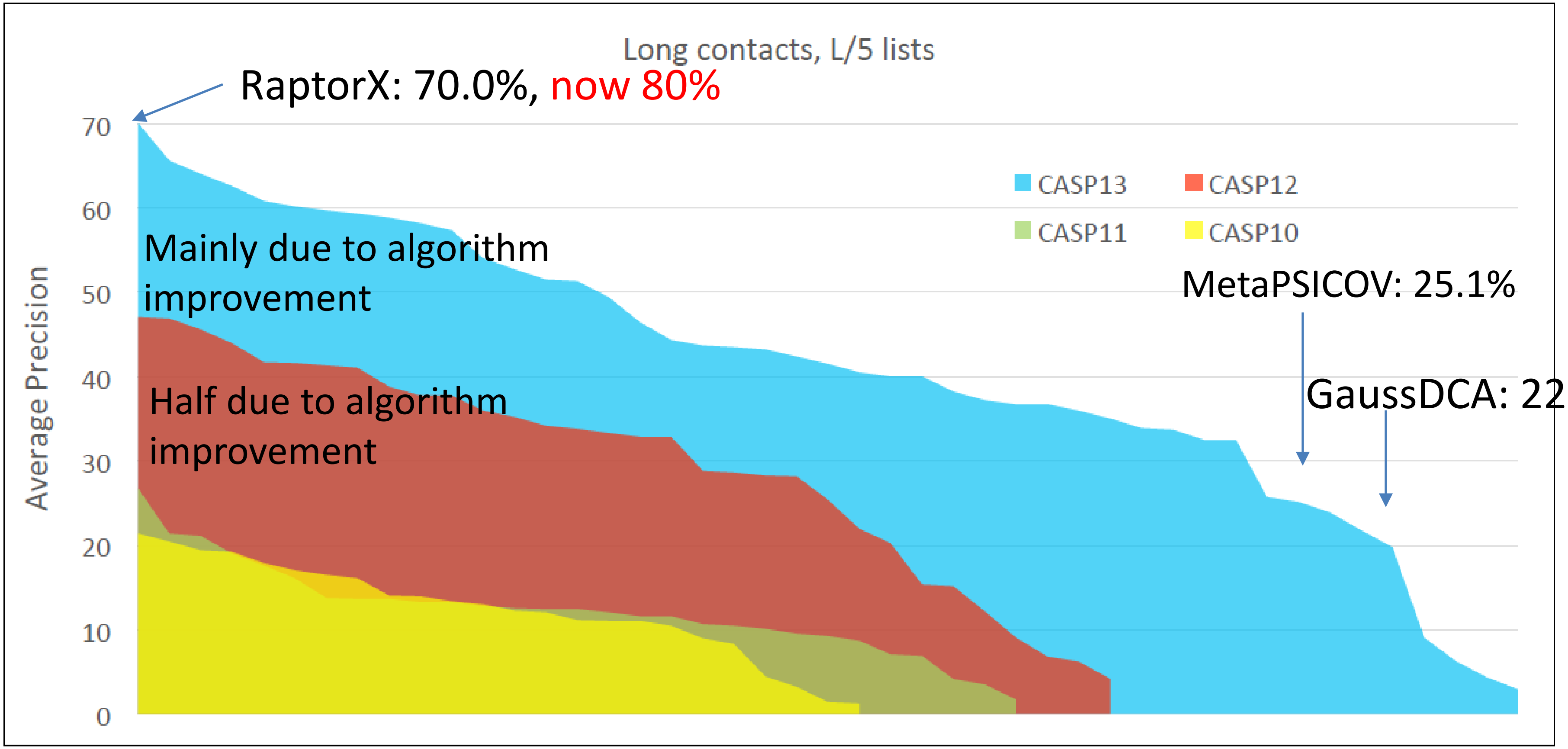
GaussDCA (co-evolution)

Contact accuracy on 31 CASP13 FM Targets

	Top L/5	Top L/2	Top L
F1 of long-range contact prediction			
AlphaFold	22.7	36.9	41.9
RaptorX	23.3	36.2	41.1
Zhang	21.2	34.1	39.2
RaptorX (post-CASP13)	**27.7**	**45.1**	**52.1**
Precision of long–range contact prediction			
RaptorX	70.0	58.0	45.0
trRosetta (post-CASP13)	78.5	66.9	51.9
RaptorX (post-CASP13)	**80.8**	**69.0**	**58.1**

CASP Progress on Contact Prediction

Courtesy of Dr. Andras Fiser



RaptorX: 70.0%, now 80%

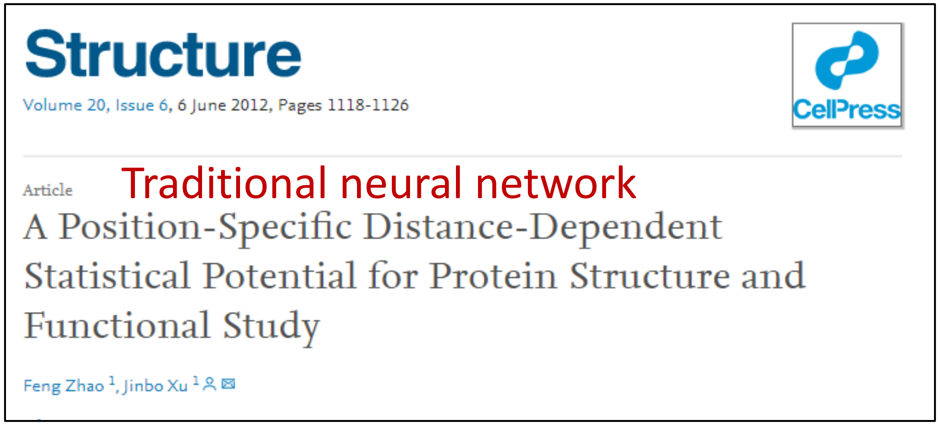
Mainly due to algorithm improvement

Half due to algorithm improvement

MetaPSICOV: 25.1%

GaussDCA: 22%

CASP10: 23 groups, 15 non-redundant
CASP11: 28 groups, 22 non-redundant
CASP12: 31 groups, 24 non-redundant
CASP13: 44 groups, 34 non-redundant



Structure

Volume 20, Issue 6, 6 June 2012, Pages 1118-1126

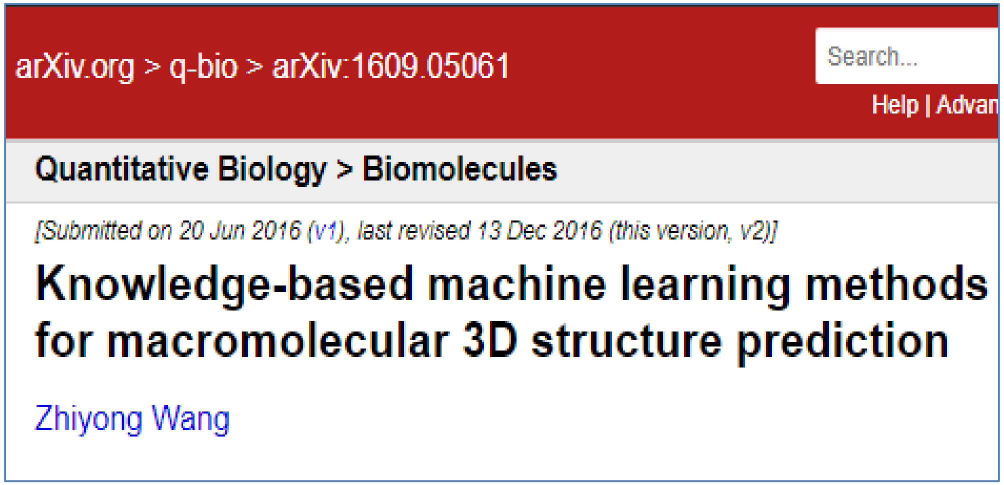
CellPress

Article

Traditional neural network

A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study

Feng Zhao [1], Jinbo Xu [1]



arXiv.org > q-bio > arXiv:1609.05061

Search...

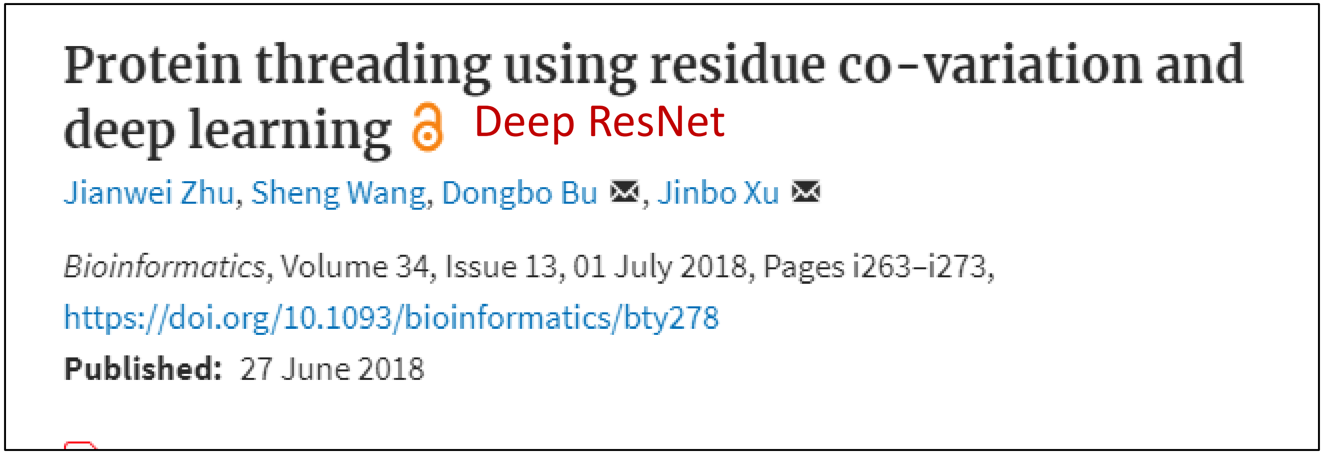
Help | Advan

Quantitative Biology > Biomolecules

[Submitted on 20 Jun 2016 (v1), last revised 13 Dec 2016 (this version, v2)]

Knowledge-based machine learning methods for macromolecular 3D structure prediction

Zhiyong Wang



Protein threading using residue co-variation and deep learning

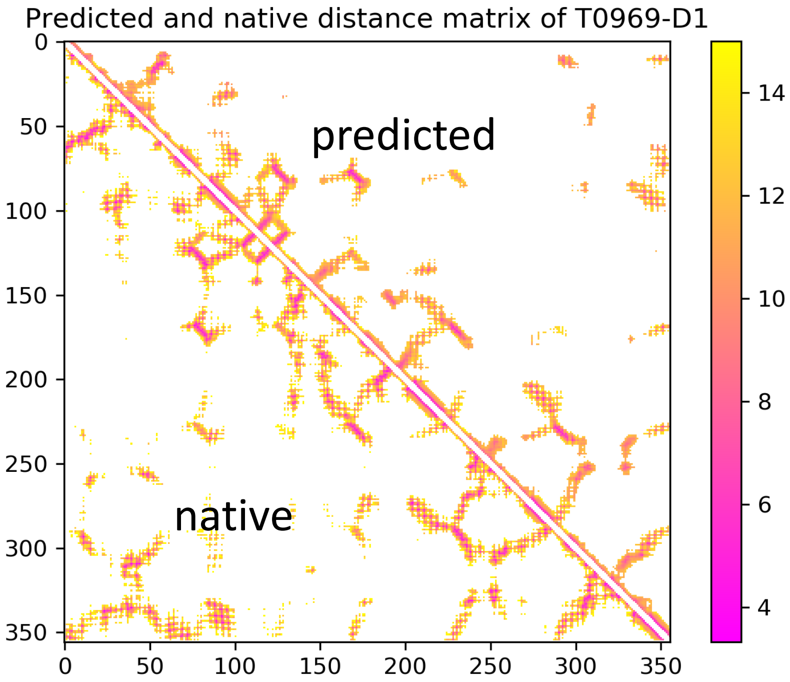
Deep ResNet

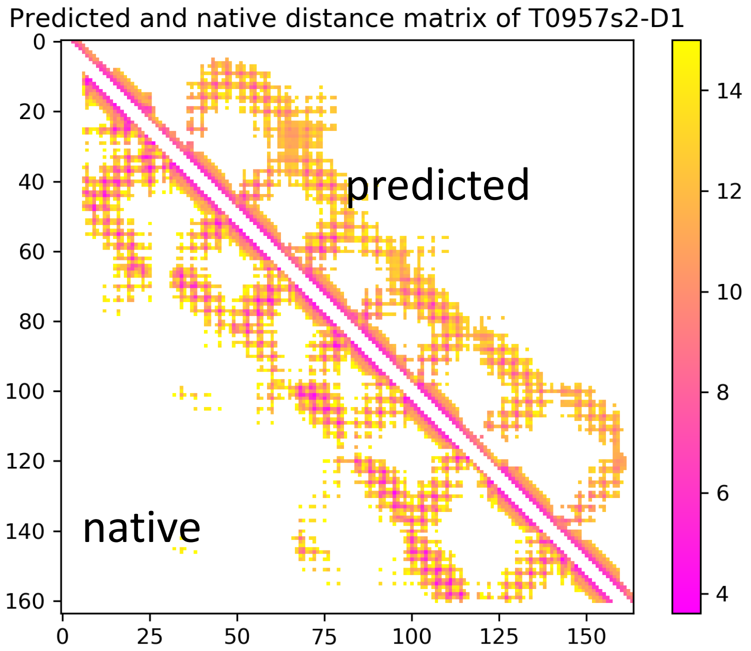
Jianwei Zhu, Sheng Wang, Dongbo Bu, Jinbo Xu

Bioinformatics, Volume 34, Issue 13, 01 July 2018, Pages i263–i273, https://doi.org/10.1093/bioinformatics/bty278

Published: 27 June 2018

CASP13 Distance Prediction Examples





Reference:

Jinbo Xu.

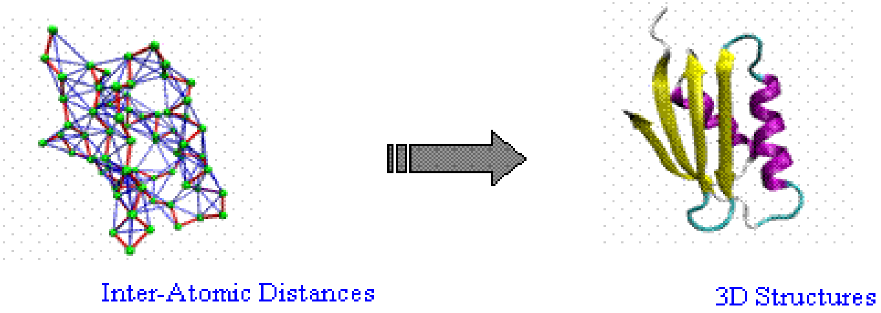
Distance-based protein folding powered by deep learning.

PNAS 2019.

Build 3D models from distance (1)

Distance geometry (embedding):

- e.g., CNS: a program used to build 3D structures from experimental data
- Best suitable for experimental data which usually has small error



Taken from http://orion.math.iastate.edu/pidd/systemdescription.htm

by solving a so-called distance geometry problem.

Build 3D models from distance (2)

Energy minimization:

- Convert predicted distance probability to statistical potential
- Minimize potential by conformation sampling and/or gradient descent

Potential=-log P(predicted distance)/background probability of distance

Reference:
F Zhao and J Xu. A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study, STRUCTURE 2012.

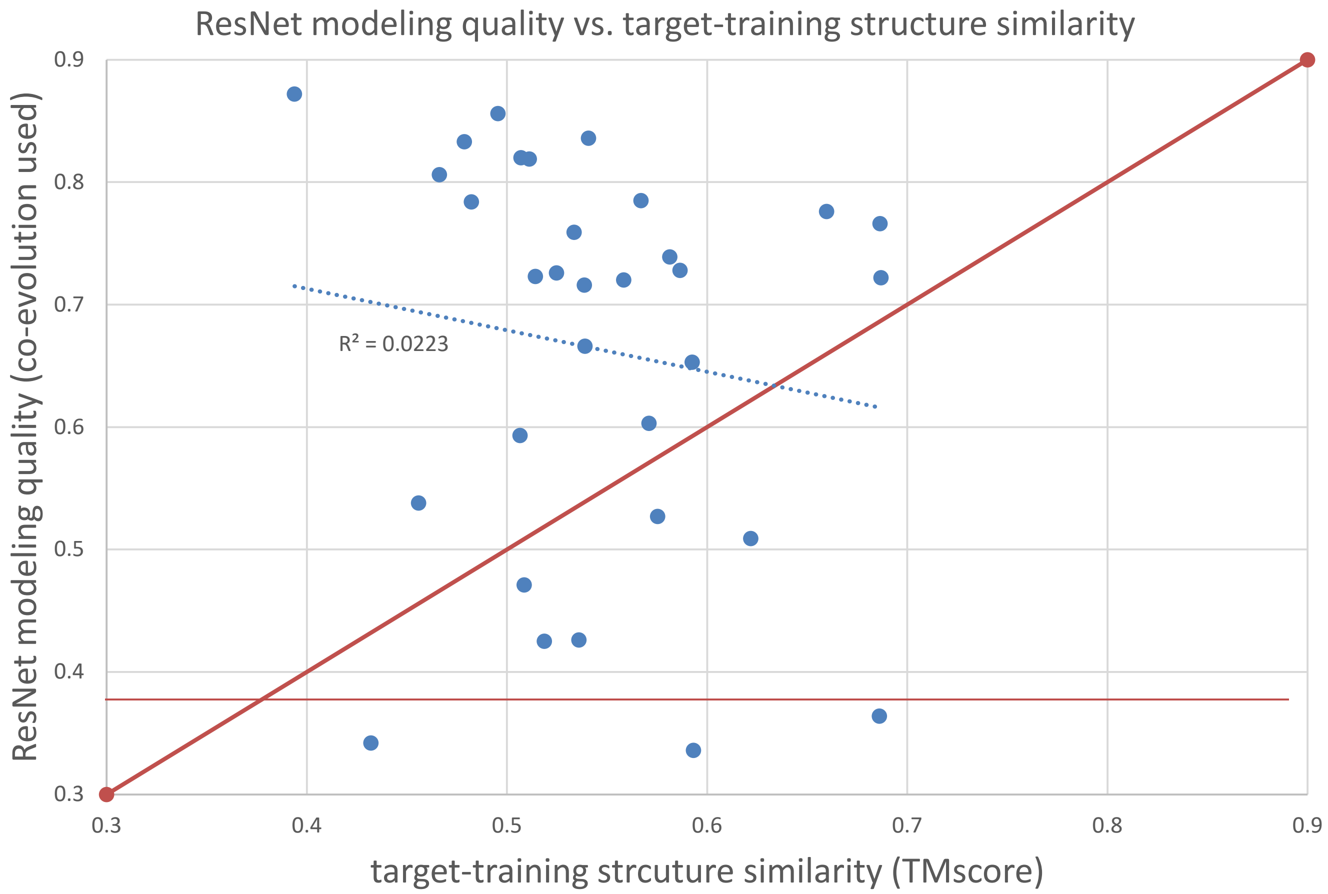
3D modeling accuracy on 32 CASP13 FM targets

	TMscore of 1st model
AlphaFold (CASP13)	0.582
trRosetta (after CASP13)	0.618
RaptorX (after CASP13)	0.640

Reference:

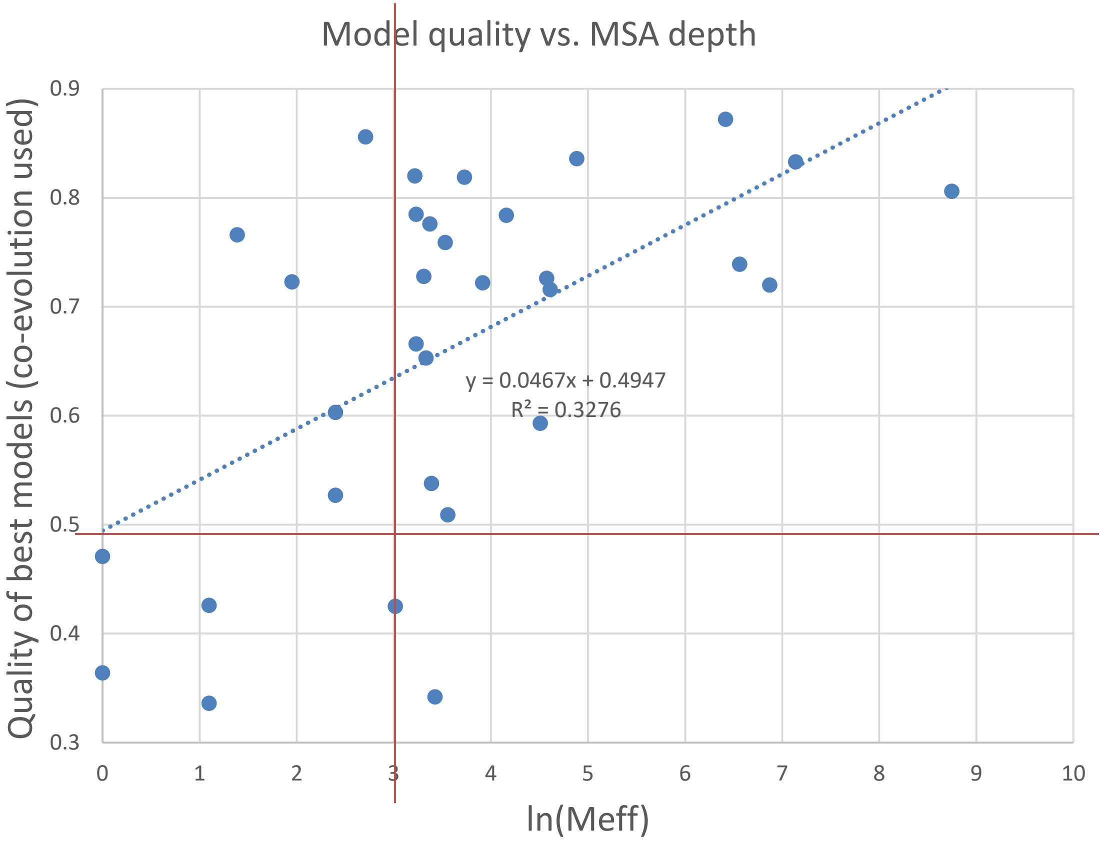
Jinbo Xu, Matthew Mcpartlon, Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information.
Nature Machine Intelligence, 2021

Deep Learning Can Predict Novel Folds



32 CASP13 FM targets: average TMscore ~0.64; 26 targets have predicted models with TMscore>0.5

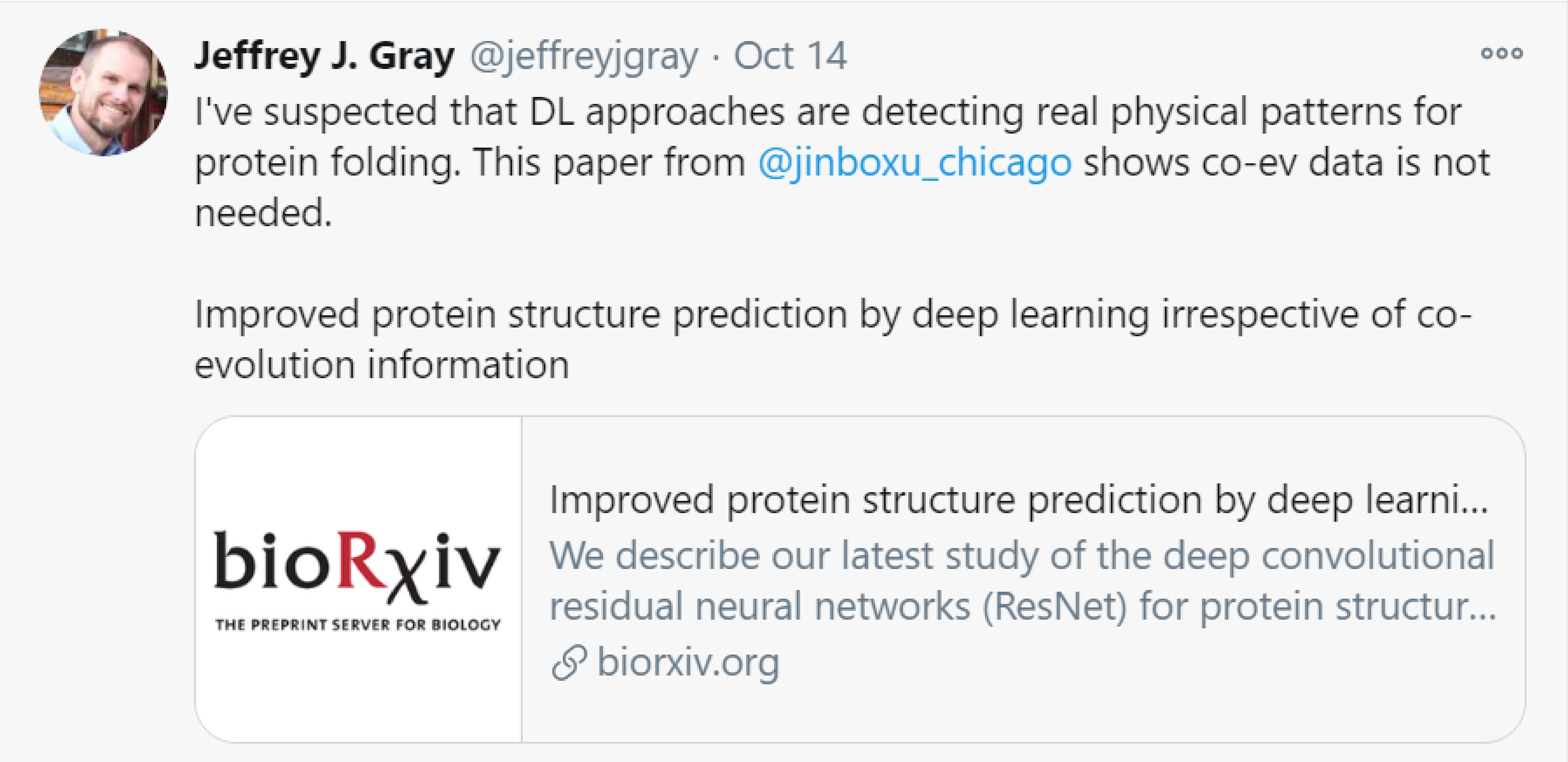
DL does not need many sequence homologs



32 CASP13 FM targets

What’s the role of deep learning?

- Denoise/amplify co-evolution signal ?
- Can DL work without co-evolution ?



Jeffrey J. Gray @jeffreyjgray · Oct 14

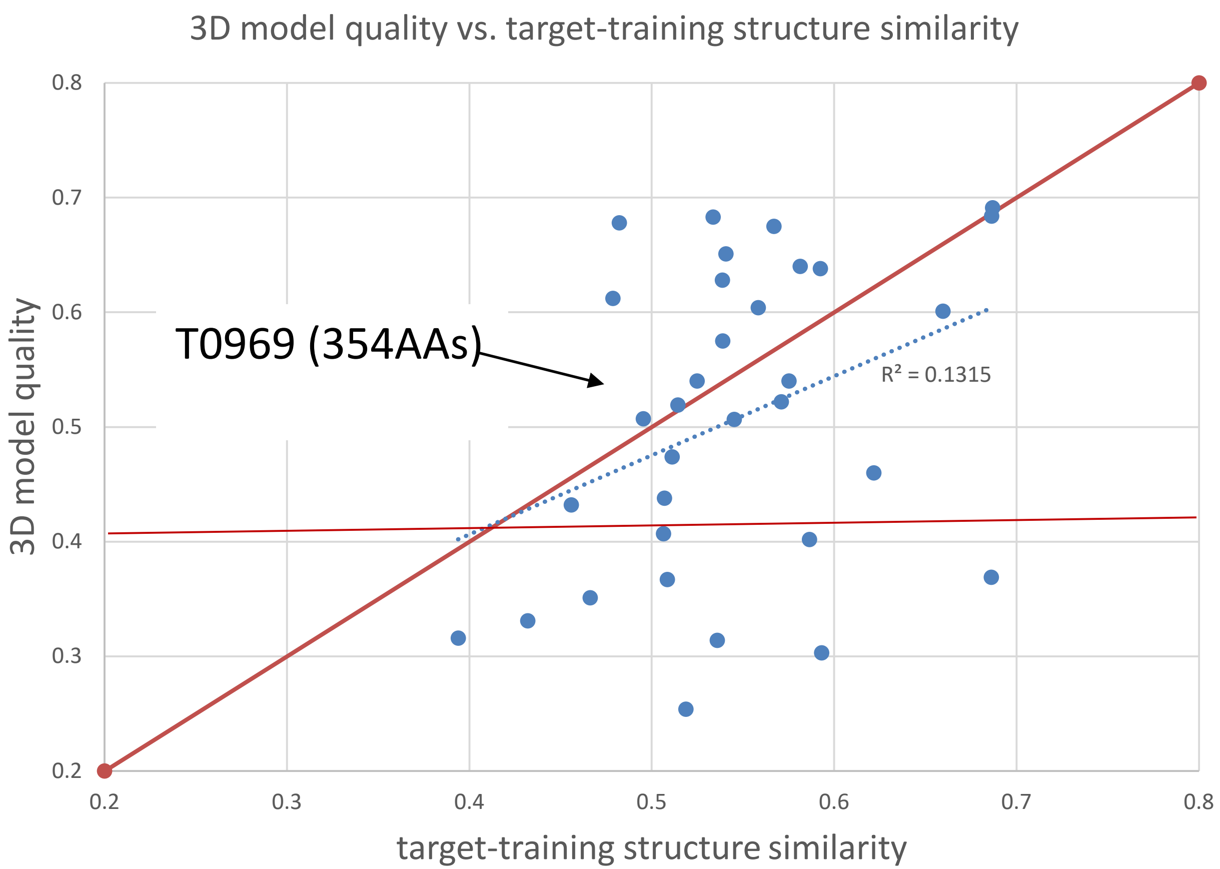
I've suspected that DL approaches are detecting real physical patterns for protein folding. This paper from @jinboxu_chicago shows co-ev data is not needed.

Improved protein structure prediction by deep learning irrespective of co-evolution information

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

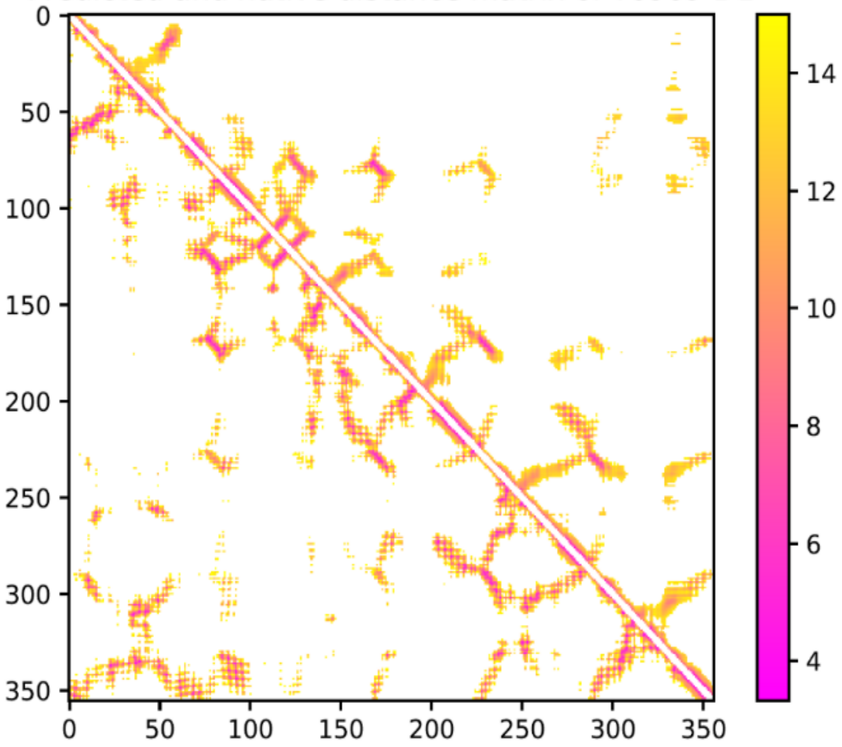
Improved protein structure prediction by deep learni...
We describe our latest study of the deep convolutional residual neural networks (ResNet) for protein structur...
biorxiv.org

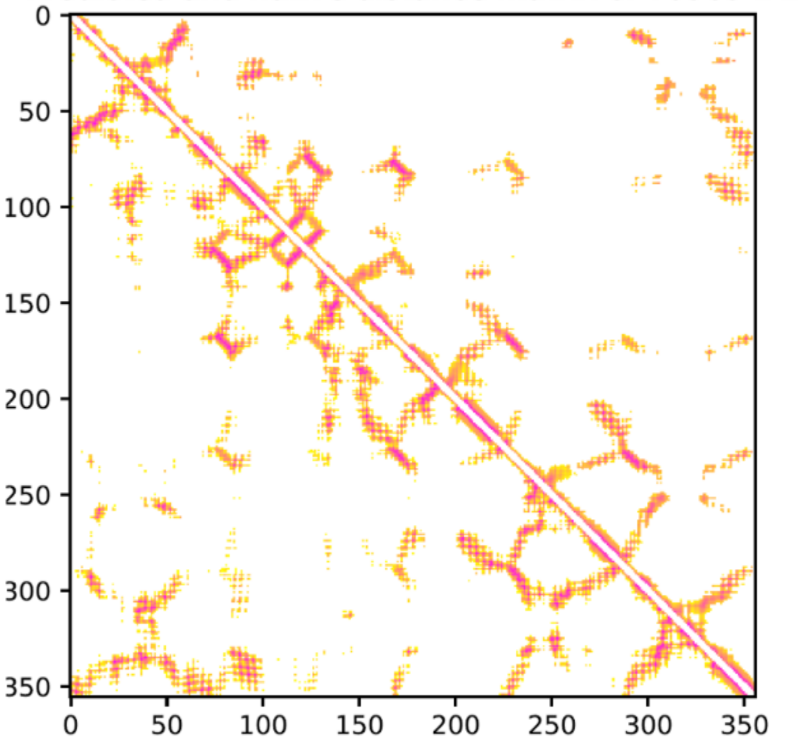
DL Can Predict Correct Folds Without Coevolution



CASP13 FM targets; Average TMscore ~0.5; 18 targets have predicted 3D models with TMscore > 0.5

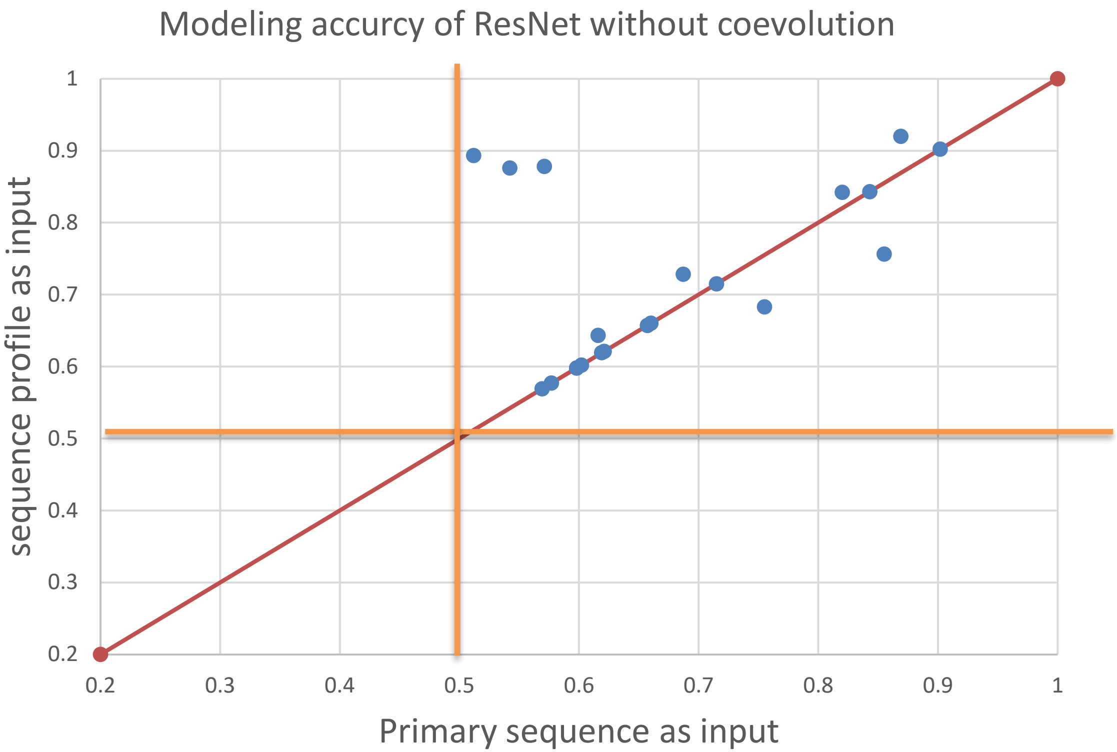
T0969


Predicted and native distance matrix of T0969-D1


Predicted and native distance matrix of T0969-D1

DL Can Predict Human Designed Proteins

designed proteins usually do not have coevolution or even evolutionary info

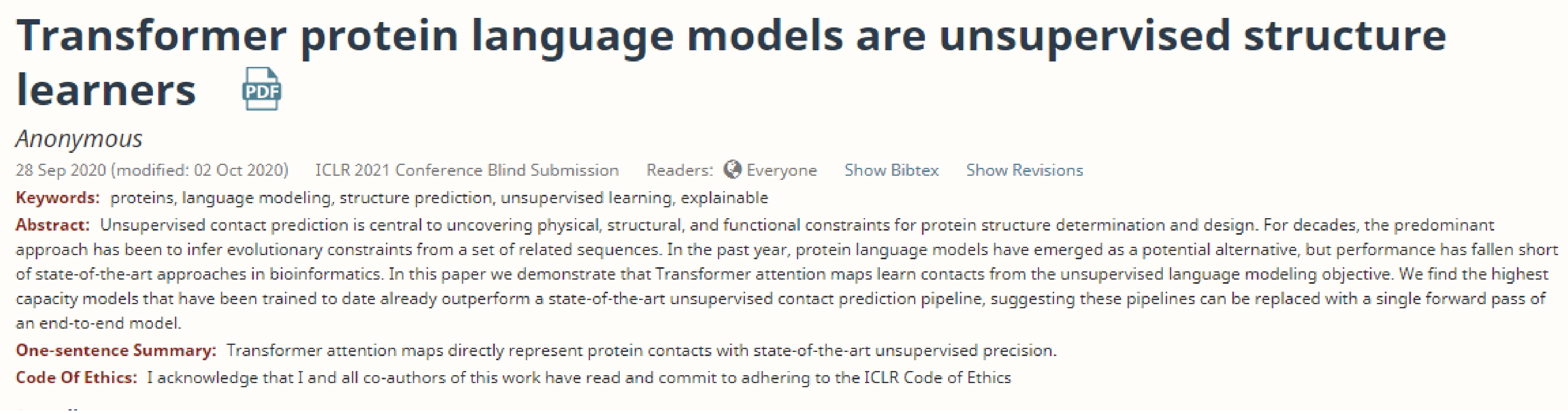


What's Next?

- Use sequence and structure info better
  - ➢ Self-supervised learning by Transformer (Facebook)
  - ➢ Template information (My group, DeepMind)
  - ➢ Learn sequence weights in an MSA (Kentaro Tomii, DeepMind)
- Better deep network architecture
  - ➢ ResNet + GAN (Haipeng Gong)
  - ➢ Replace direct-coupling analysis by ResNet(Dongbo Bu)
  - ➢ Transformer-like supervised learning

(1) ResNet --> Transformer

Facebook's work:



Transformer protein language models are unsupervised structure learners

Anonymous

28 Sep 2020 (modified: 02 Oct 2020) ICLR 2021 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Keywords: proteins, language modeling, structure prediction, unsupervised learning, explainable

Abstract: Unsupervised contact prediction is central to uncovering physical, structural, and functional constraints for protein structure determination and design. For decades, the predominant approach has been to infer evolutionary constraints from a set of related sequences. In the past year, protein language models have emerged as a potential alternative, but performance has fallen short of state-of-the-art approaches in bioinformatics. In this paper we demonstrate that Transformer attention maps learn contacts from the unsupervised language modeling objective. We find the highest capacity models that have been trained to date already outperform a state-of-the-art unsupervised contact prediction pipeline, suggesting these pipelines can be replaced with a single forward pass of an end-to-end model.

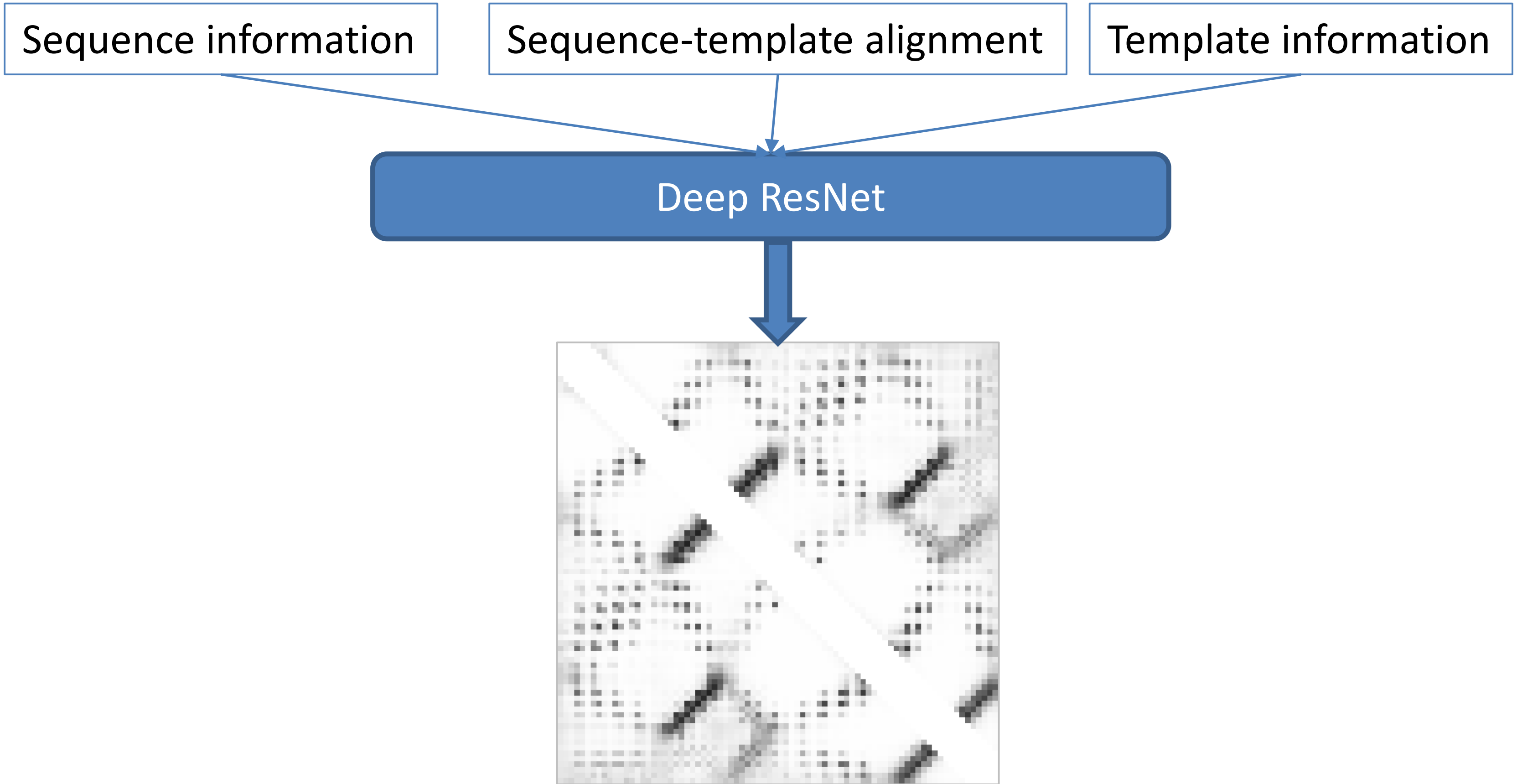
One-sentence Summary: Transformer attention maps directly represent protein contacts with state-of-the-art unsupervised precision.

Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics

DeepMind's work: a supervised Transformer-like network

(2) Integrate templates by DL

Started this idea in CASP13, but it took a long time to implement this idea well



Sequence information

Sequence-template alignment

Template information

Deep ResNet

Integrate Template by DL (Cont'd)

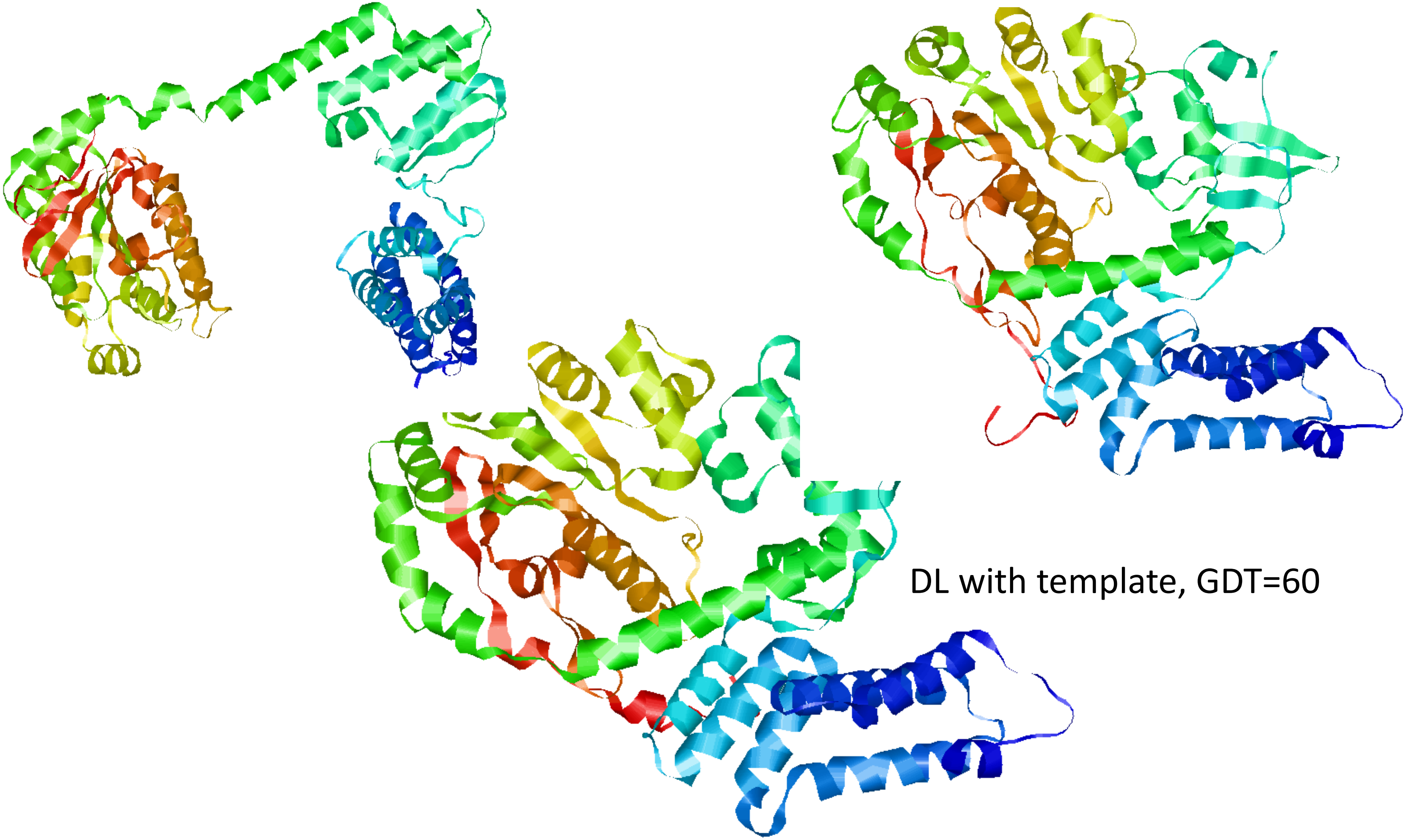
- Template-based modeling usually was the 1st choice
- Now template-free modeling outperforms template-based for unless very good templates (e.g., seq id > 35%)
- Even on some targets with very good templates, template-free modeling is better

When Templates Useful?

- When templates have seq id >40%, use MODELLER or RosettaCM
- Bad templates are not useful and even harmful, e.g., HHblits E-value>0.01
- Likely useful when templates with HHblits E-value <10^-5 and seq id < 35%, e.g., CASP13 TBM-hard and some TBM-easy targets

T0966

492 AAs, but <200 seq homologs, good template seq id 25%



DL without template, GDT=27

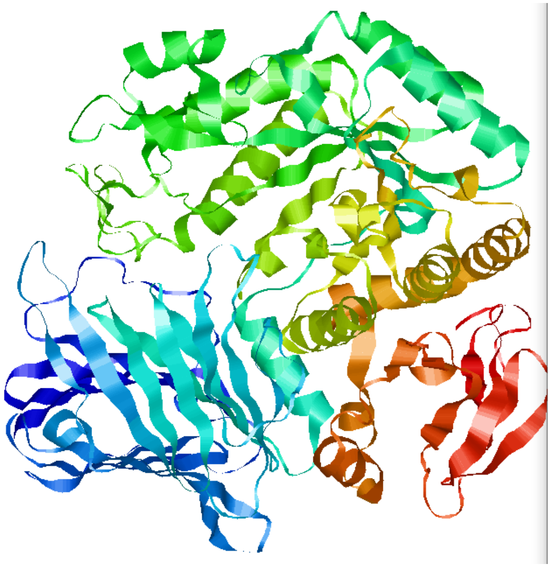
MODELLER with template, GDT=53

DL with template, GDT=60

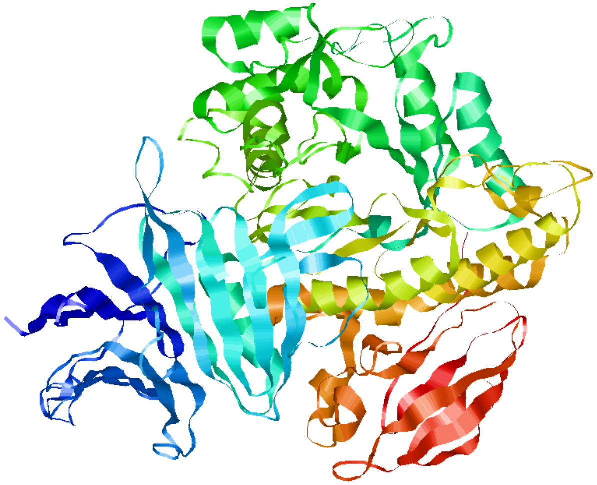
T1009

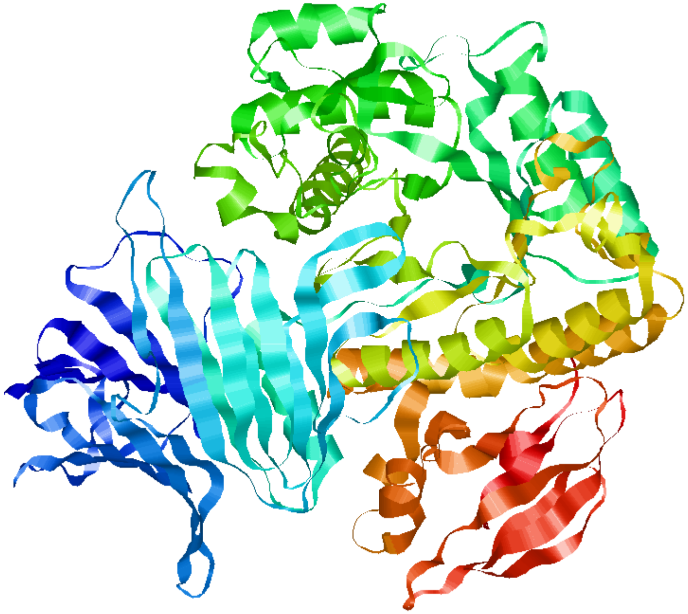
718 AAs, >13k seq homologs, good template 25% seq id

DL without template, GDT=60



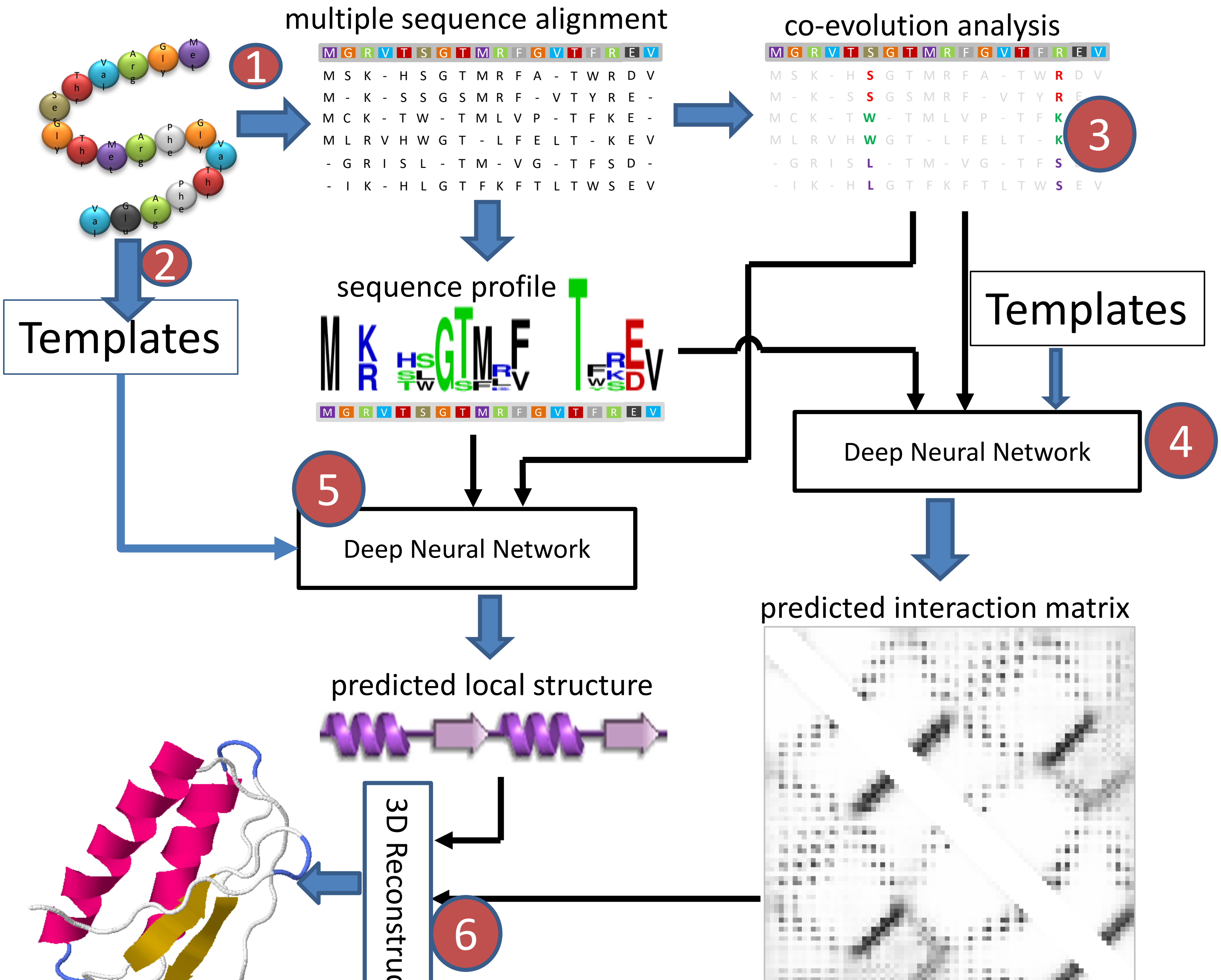
MODELLER with template, GDT=62





DL with template, GDT=68

(3) Towards An End-to-End Workflow



multiple sequence alignment

co-evolution analysis

sequence profile

Templates

Templates

Deep Neural Network

Deep Neural Network

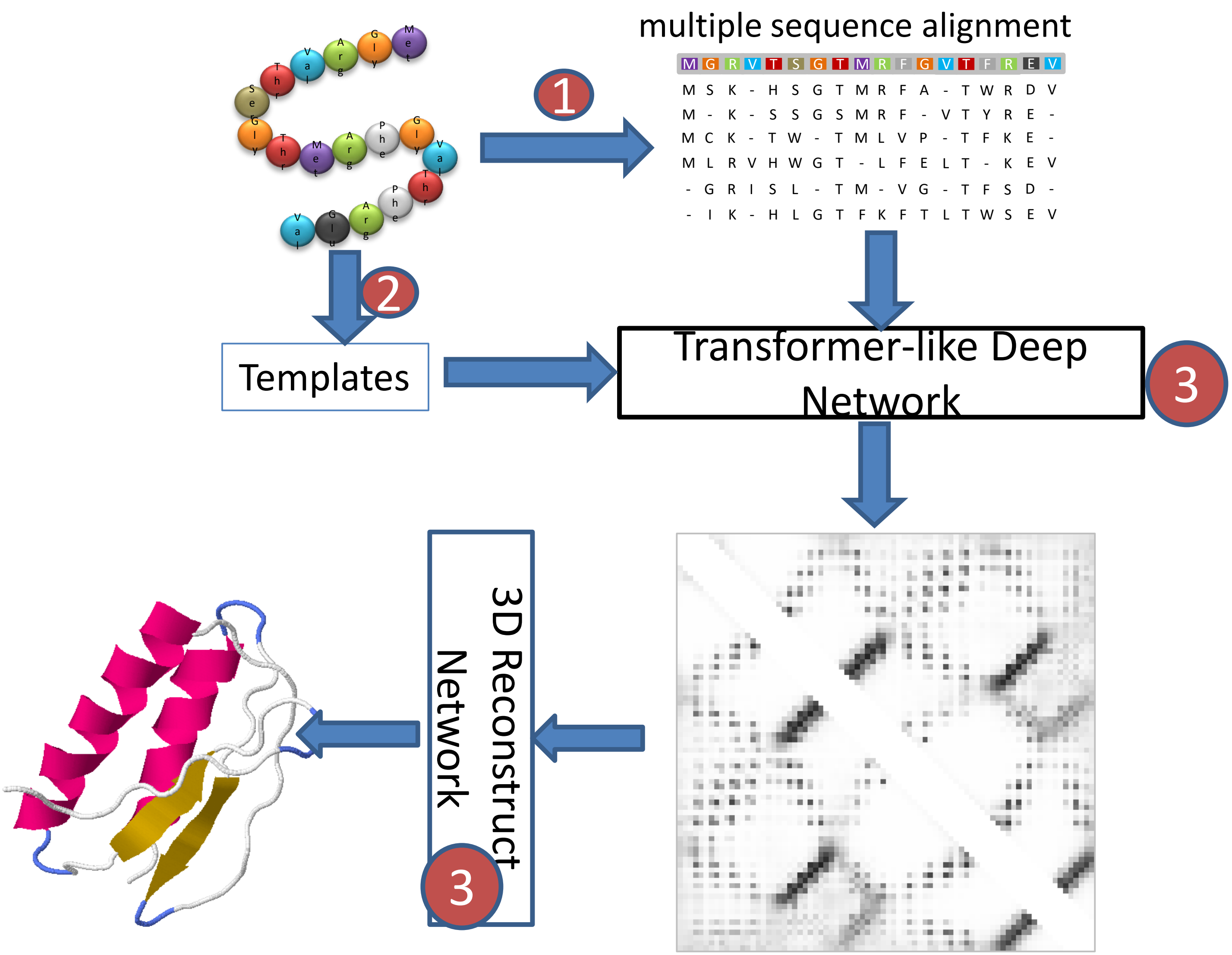
predicted interaction matrix

predicted local structure

3D Reconstruc

1
2
3
4
5
6
M G R V T S G T M R F G V T F R E V
M S K - H S G T M R F A - T W R D V
M - K - S S G S M R F - V T Y R E -
M C K - T W - T M L V P - T F K E -
M L R V H W G T - L F E L T - K E V
- G R I S L - T M - V G - T F S D -
- I K - H L G T F K F T L T W S E V
Met
Gly
Arg
Val
Thr
Ser
Phe
Glu

AlphaFold2: Almost End-to-End Workflow



multiple sequence alignment

M G R V T S G T M R F G V T F R E V

M S K - H S G T M R F A - T W R D V
M - K - S S G S M R F - V T Y R E -
M C K - T W - T M L V P - T F K E -
M L R V H W G T - L F E L T - K E V
- G R I S L - T M - V G - T F S D -
- I K - H L G T F K F T L T W S E V

1

2

Templates

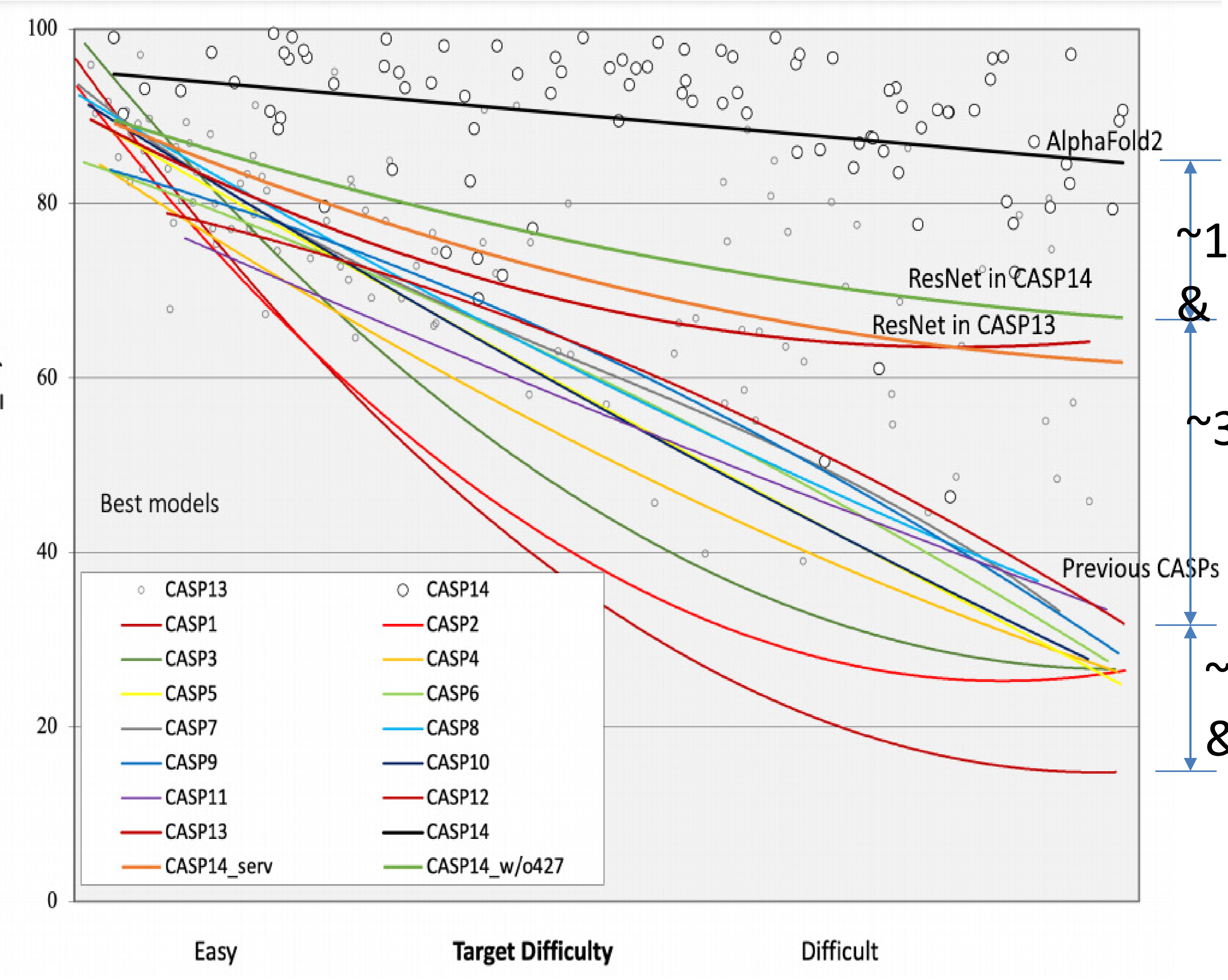
Transformer-like Deep Network

3

3D Reconstruct Network

3

CASP Progress on 3D Modeling



~19 by Transformer & End-to-End

~33 by ResNet

~17 by fragment & co-evolution

Summary

- Deep learning is powerful for protein folding
  - ResNet has success rate over 80%
  - End-to-End Transformer-like network much better
- Can fold very large proteins much faster than before
- Outperforms template-based modeling unless very good templates are available
- Integrate templates to Deep Learning for better
- More innovations to be seen