DeepPotential: Deep learning based inter-residue contact/distance prediction in CASP14

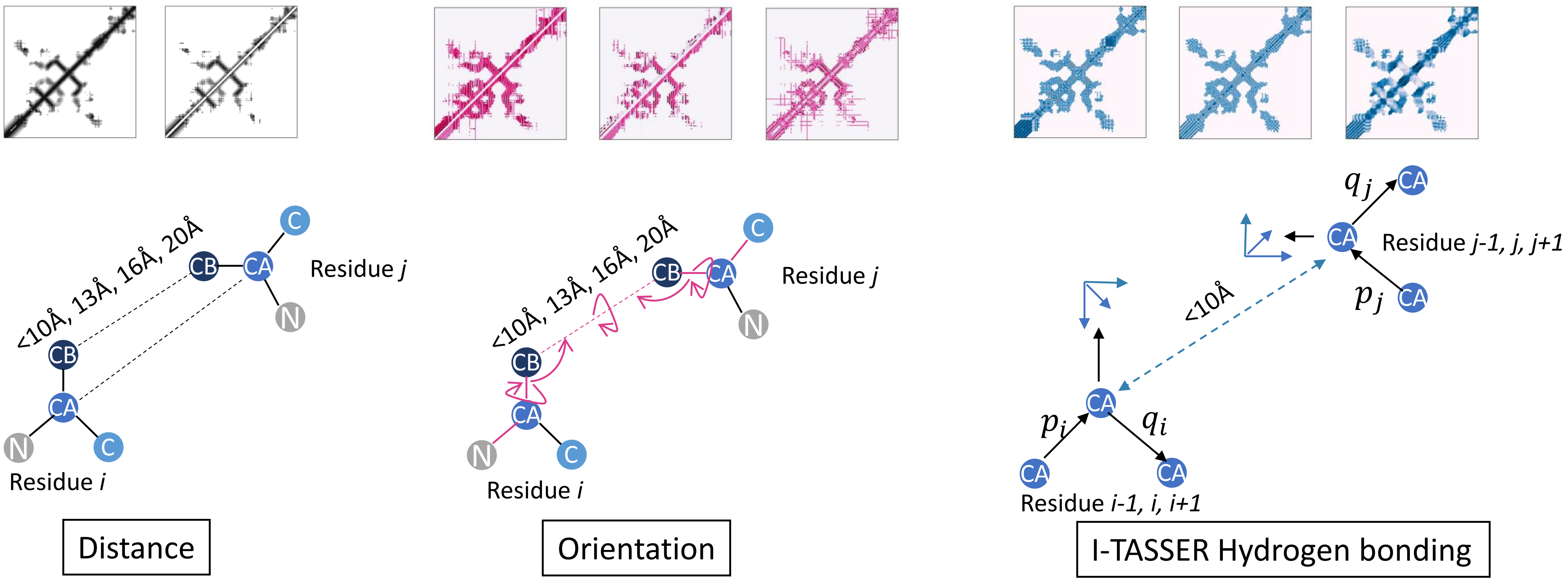
Yang Li, Chengxin Zhang, Wei Zheng, Xiaogen Zhou, Eric W. Bell, Dong-Jun Yu, and Yang Zhang

University of Michigan

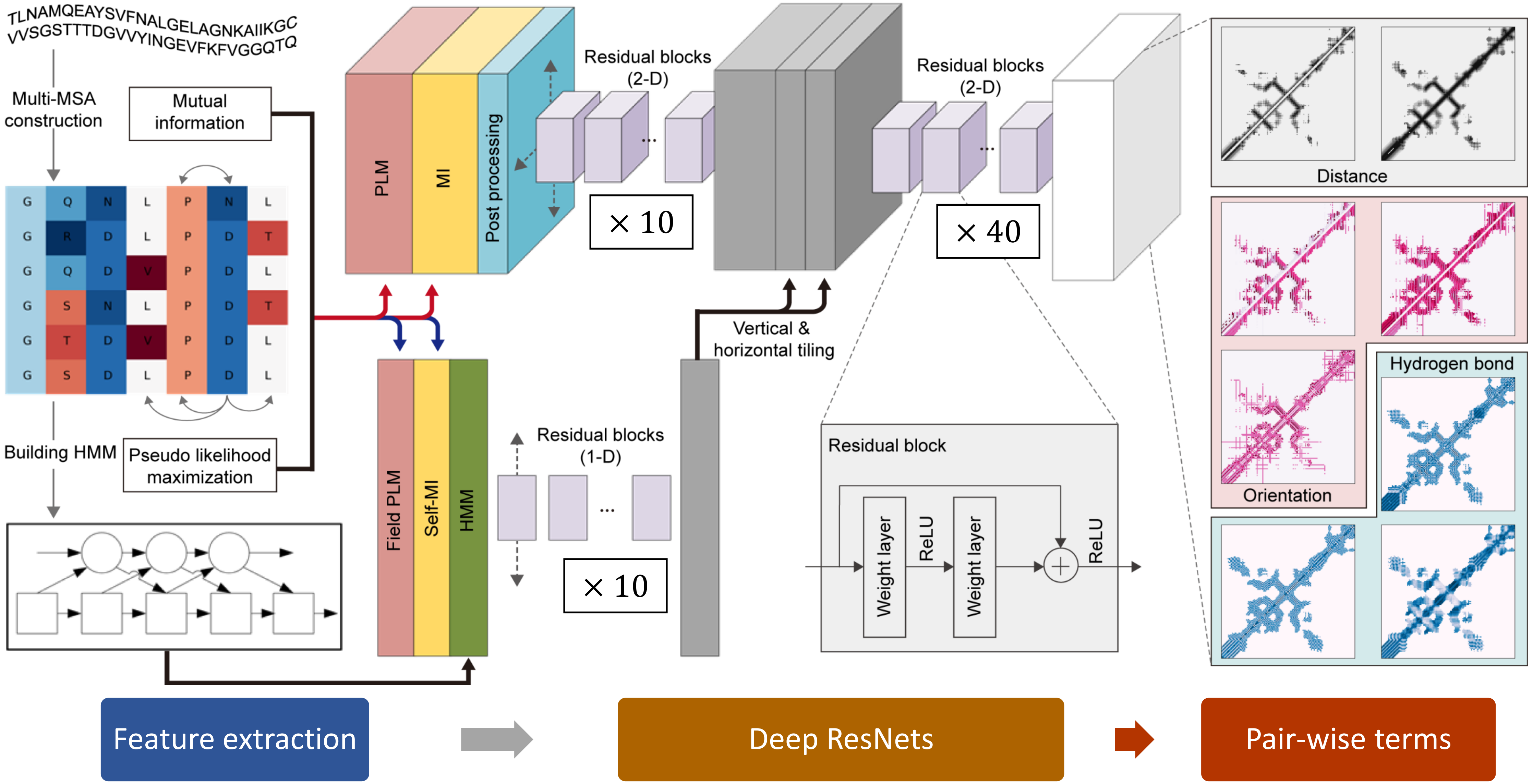
Nanjing University of Science and Technology

DeepPotential

Predicting (long-range) pair-wise **statistical potential terms** for protein structure prediction,

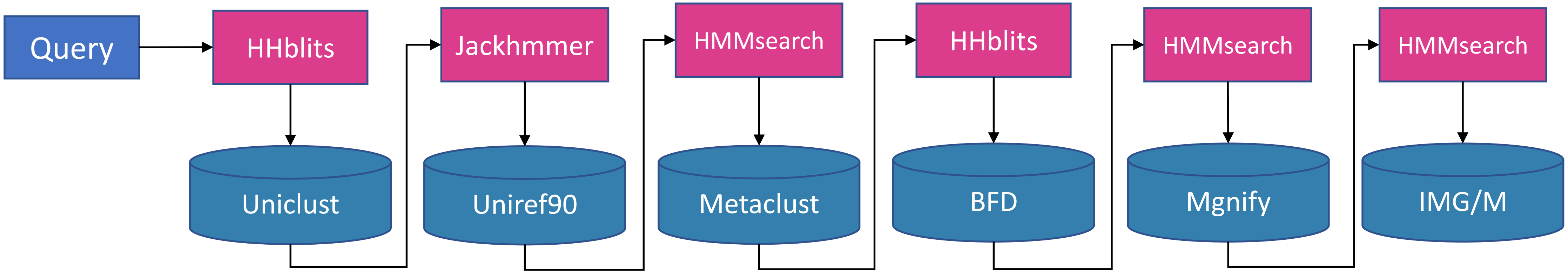


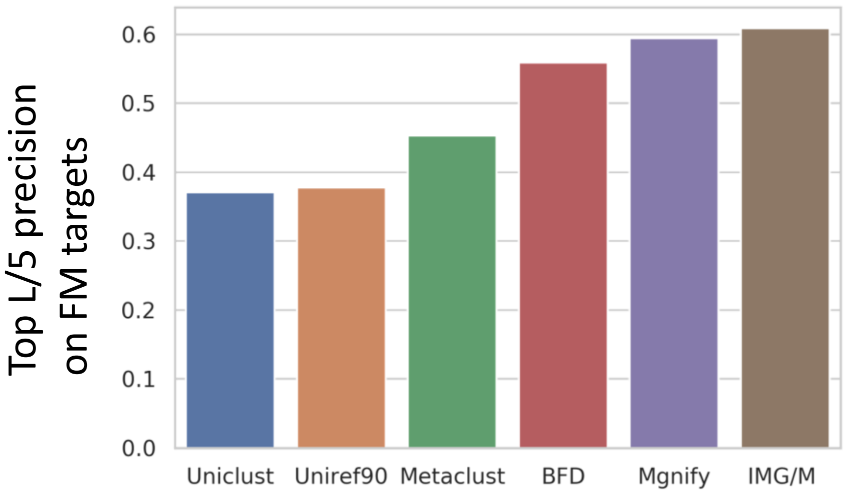
DeepPotential

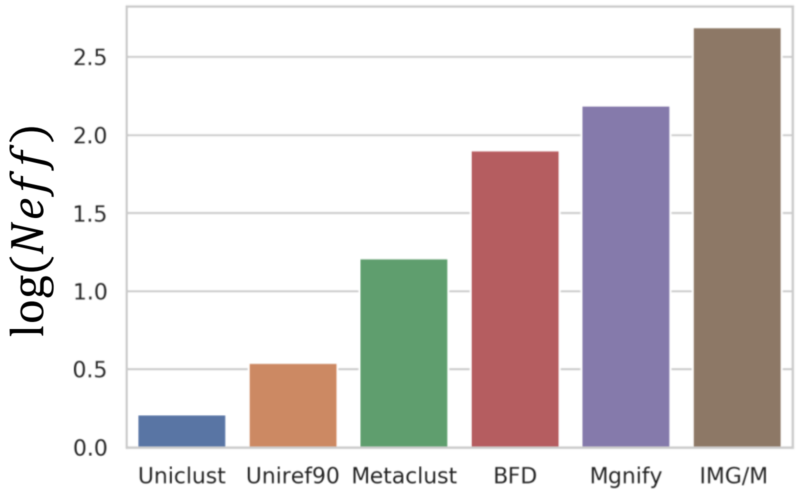


MSA construction

Progressive collection of MSA increasing accuracy of contact prediction



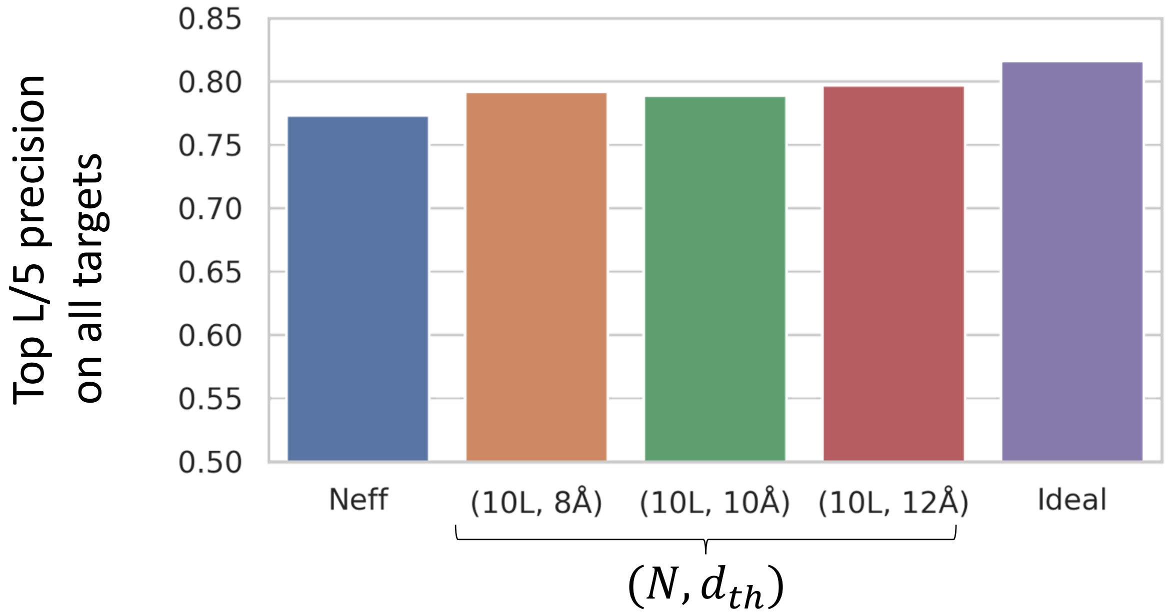




MSA selection

MSA selection based on confidence score outperforms based on Neff

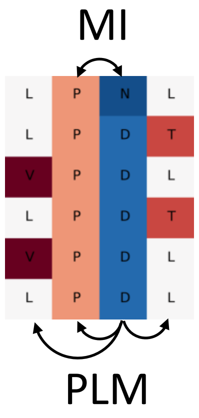
- Select MSA based on mean of top-$\boldsymbol{N}$ DeepPotential contact probabilities (defined at the threshold of $\boldsymbol{d_{th}}$, $p(x < \boldsymbol{d_{th}})$)
- Use the prediction from the selected MSA



In CASP14, two confidence score configurations are considered:

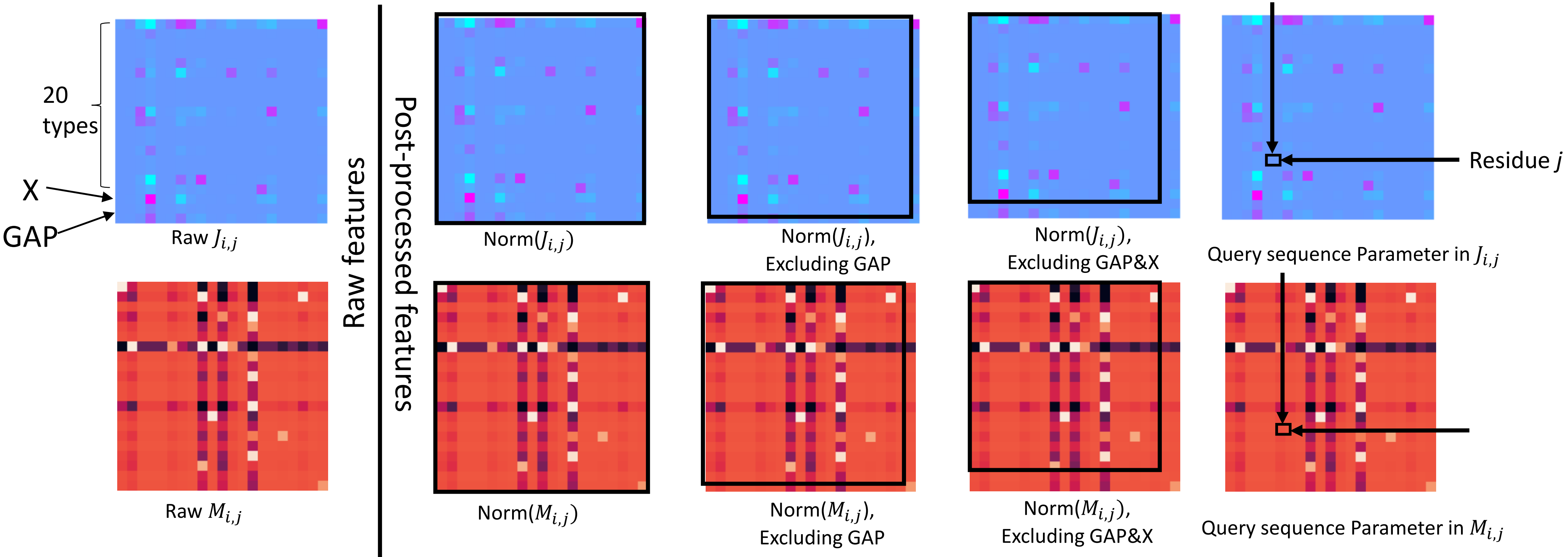
- $(\boldsymbol{N} = 10 \times L$, $\boldsymbol{d_{th}} = 12$Å), Group name: TripletRes
- $(\boldsymbol{N} = 10 \times L$, $\boldsymbol{d_{th}} = 8$Å), Group name: DeepPotential

Feature extraction



Co-evolutionary features:

- Couplings matrix $\boldsymbol{J}$ ($\boldsymbol{J} \in \boldsymbol{R}^{L \times L \times 22 \times 22}$) of Pseudolikelihood maximization (PLM)
- Raw Mutual information matrix (MI): $\boldsymbol{M}$ ($\boldsymbol{M} \in \boldsymbol{R}^{L \times L \times 22 \times 22}$);
- And their post-processing. $(L \times L \times (4+4))$



Training

Training data:

- 26,151 structures from PDB, by 11/12/2019
- Sequence identity cut-off of 35%
- Maximum length of 1000
- Training MSA: HHblits against Uniclust only

Loss function

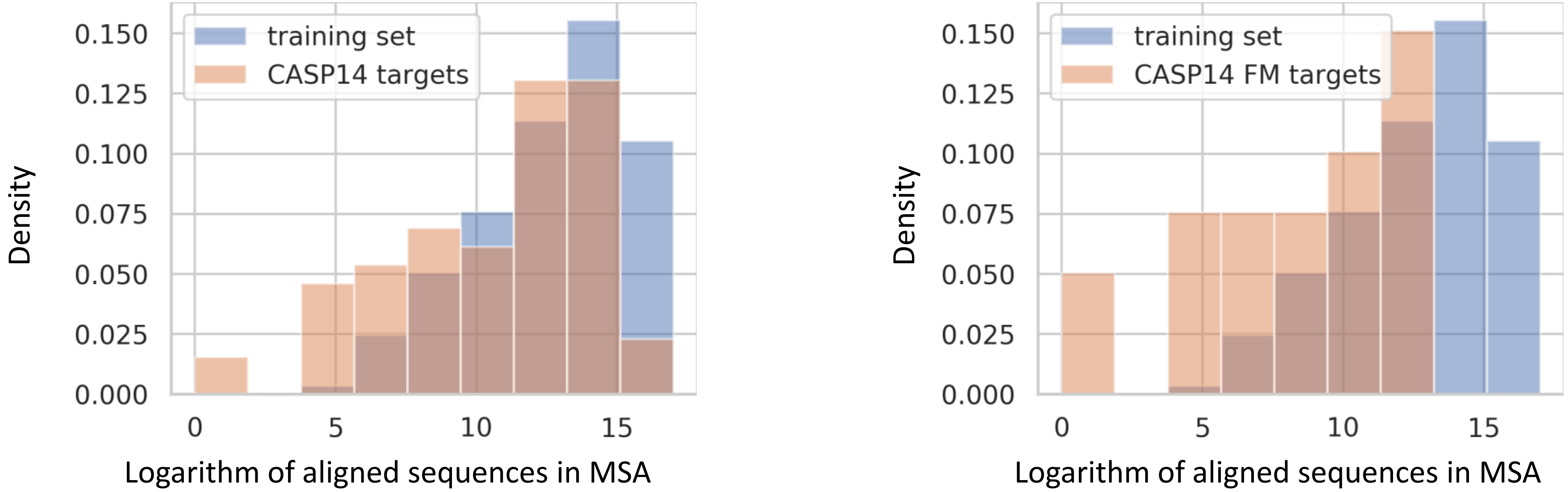
- Discretizing prediction terms into bins
- Neg-log likelihood of all prediction terms
- $Loss = -\sum_{n=1}^{N} \sum_{i,j} \sum_{t \in T} w_t \log P(data_n^t(i,j) | \boldsymbol{J}, \boldsymbol{M})$
  - n, i, j enumerates all residue pairs in the training set
  - $w_t = 1$ for all $t \in \{distance\ terms; orientation\ terms; Hbond\ terms\}$

Approximations: Independent distributed in

- $p(data) = \prod_{n=1}^{N} p(data_n)$ ← *Samples*
- $= \prod_{n=1}^{N} \prod_{t \in T} p(data_n^t)$ ← *Prediction terms*
- $= \prod_{n=1}^{N} \prod_{t \in T} \prod_{i,j} p(data_n^t(i,j))$ ← *Residue pairs (pixels)*

Training

Generalization ability of the model



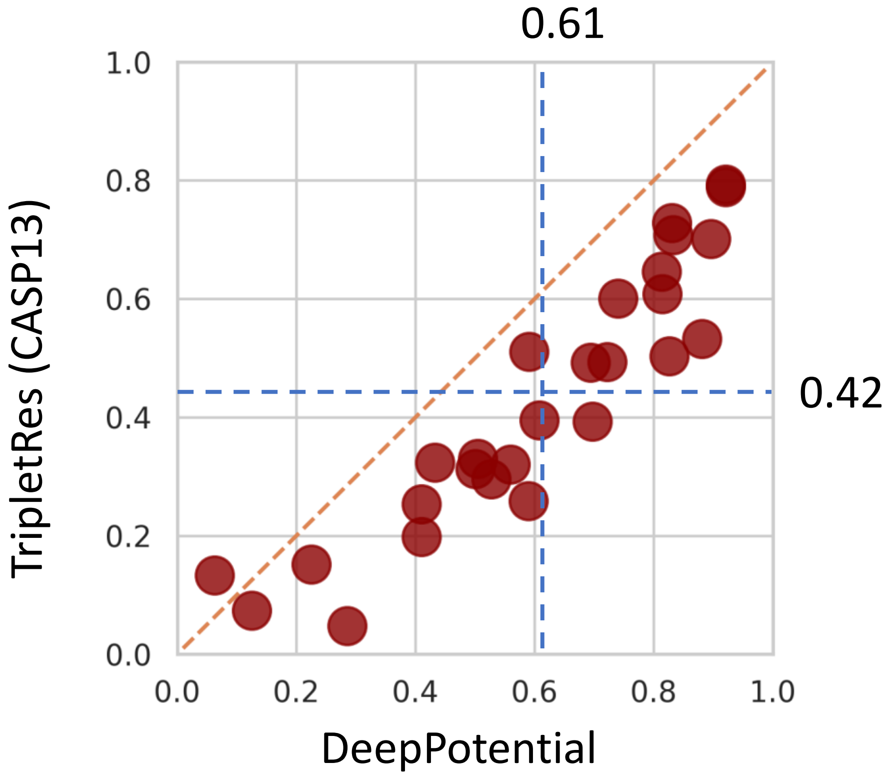
- Sub-sampling MSAs during the training
- Larger weights on shallow MSAs

The finale prediction is the ensemble of 15 diverse models, with different combination of terms and thresholds

Results

Results in contact prediction on CASP13 targets

- Head-to-head comparison of long-range top-*L* precision on 27 CASP13 FM targets



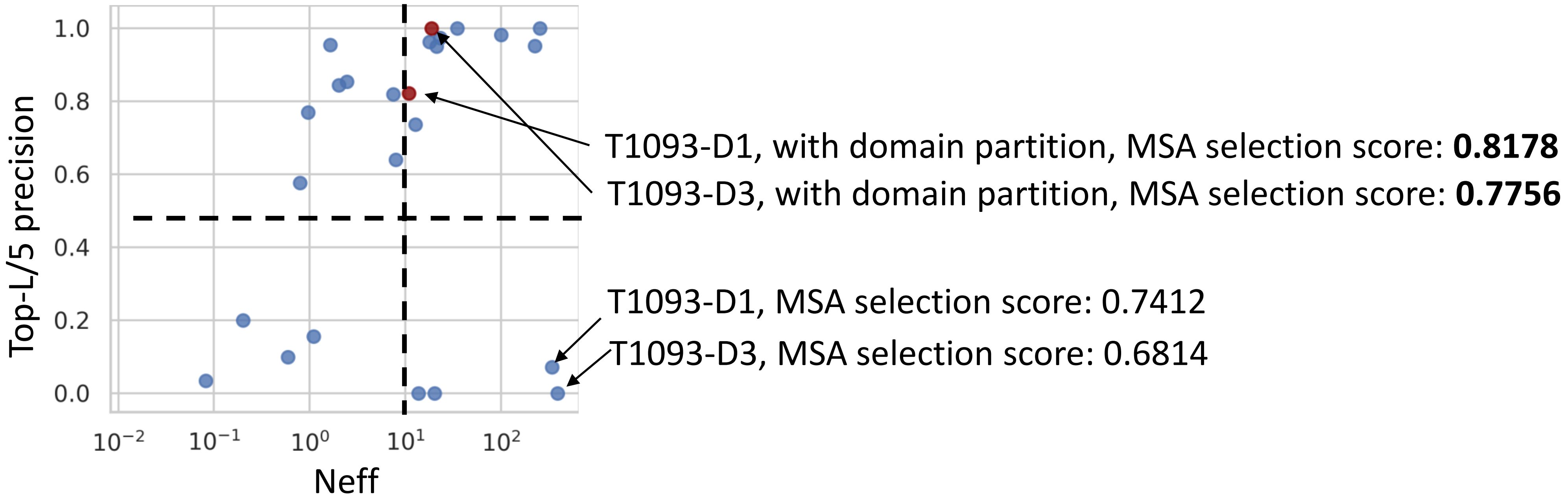
DeepPotential is over 40% higher than CASP13 version of TripletRes

Results

Results of DeepPotential in CASP14

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |d^i_{expectation} - d^i_{experimental}|$$

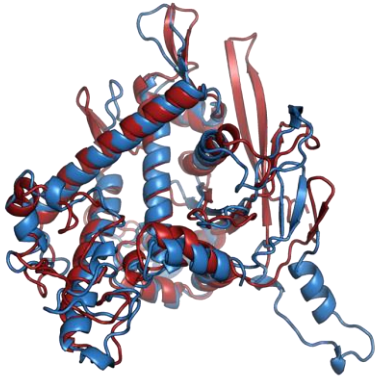
MSA selection	Contact precision (long range)				Mean Absolute Error (long range)		
	Top L/10	Top L/5	Top L/2	Top L	Top L	Top 2L	Top 5L
$N = 10 \times L$ $d_{th} = 8Å$	**65.53**	**61.31**	**50.96**	**37.66**	**2.68**	2.89	**3.23**
$N = 10 \times L$ $d_{th} = 12Å$	62.67	59.01	48.16	36.59	2.69	**2.87**	3.25



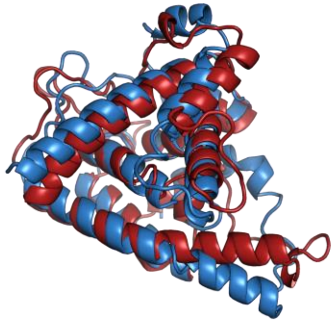
Results

DeepPotential is capable of folding high-accuracy protein structures

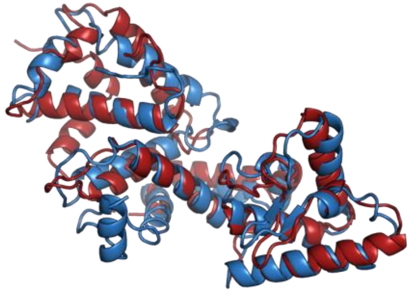
9 FM targets with contact precision over 0.8 and Zhang-Server has a TM-score over 0.5



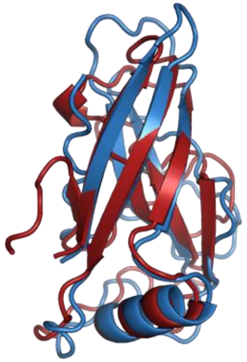
T1037-D1, MAE=0.948
TM-score=0.680



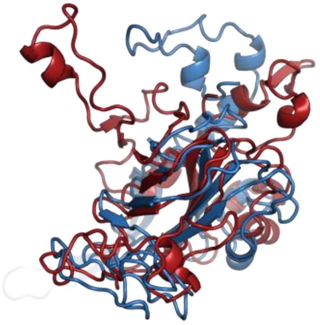
T1041-D1, MAE=0.760
TM-score=0.722



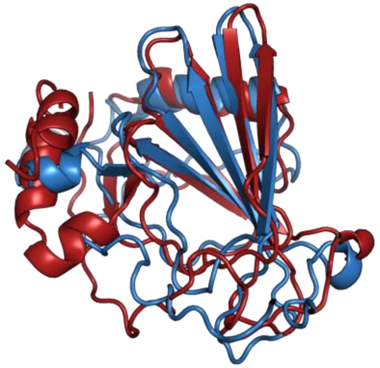
T1042-D1. MAE=1.344
TM-score=0.730



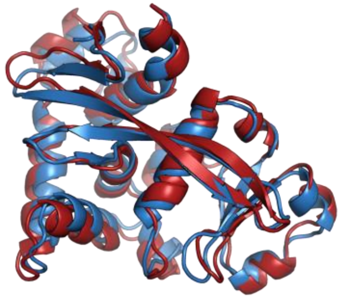
T1049-D1, MAE=1.561
TM-score=0.675



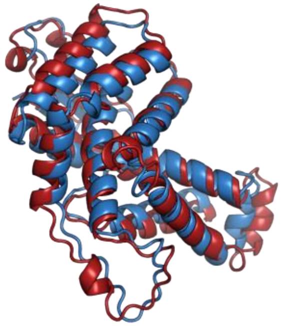
T1061-D2, MAE=1.109
TM-score=0.527



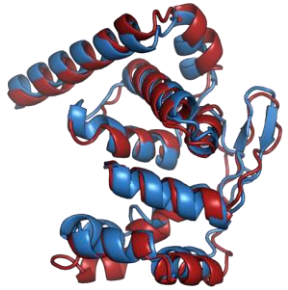
T1090-D1, MAE=1.157
TM-score=0.656



T1094-D2, MAE=1.062
TM-score=**0.914**



T1096-D1, MAE=1.189
TM-score=0.835



T1096-D2, MAE=1.454
TM-score=0.833

Native

Zhang-Server (DeepPotential + I-TASSER)

MAE: Top-5L long range MAE

Summary

What was working?

- More data help the training
- Constructing deeper MSA
- MSA selection by top-N contact scores
- Various prediction tasks
- Raw coevolution/multi-view feature fusion

What went wrong?

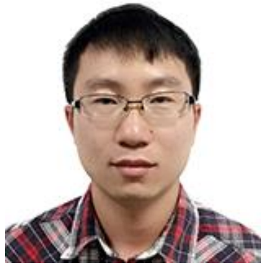
- Limited computational resources, trainable with single GPU (10GB)
  - RAW Precision matrix (PRE in TripletRes (CASP13)) was discarded
  - Deeper/wider neural networks was not considered
- Tuning weight of distance term should help distance/contact accuracy
- Overconservative domain partition.

Acknowledgements

Zhanglab members

Chengxin Zhang

Wei Zheng



Xiaogen Zhou

Eric W. Bell

Dong-Jun Yu

Yang Zhang

Special thanks to

- The authors of *trRosetta* for insightful discussion

Thank you!