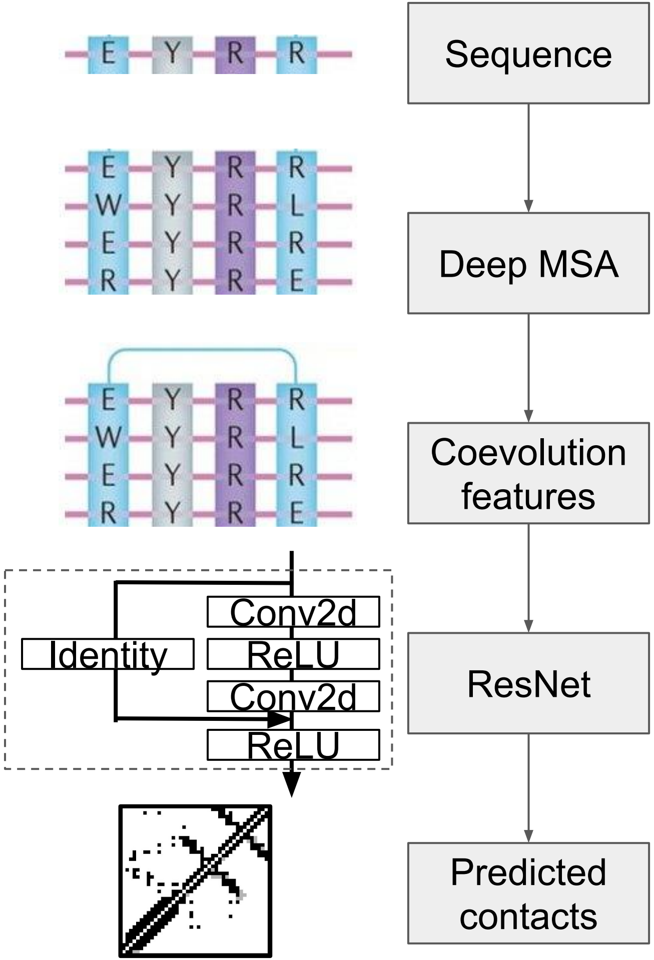
ResTriplet/TripletRes: Learning contact-maps from a triplet of coevolutionary matrices

Eric W. Bell, Yang Li, Chengxin Zhang, Dong-Jun Yu, Yang Zhang

Department of Computational Medicine and Bioinformatics, University of Michigan - Ann Arbor

ResTriplet/TripletRes Method overview

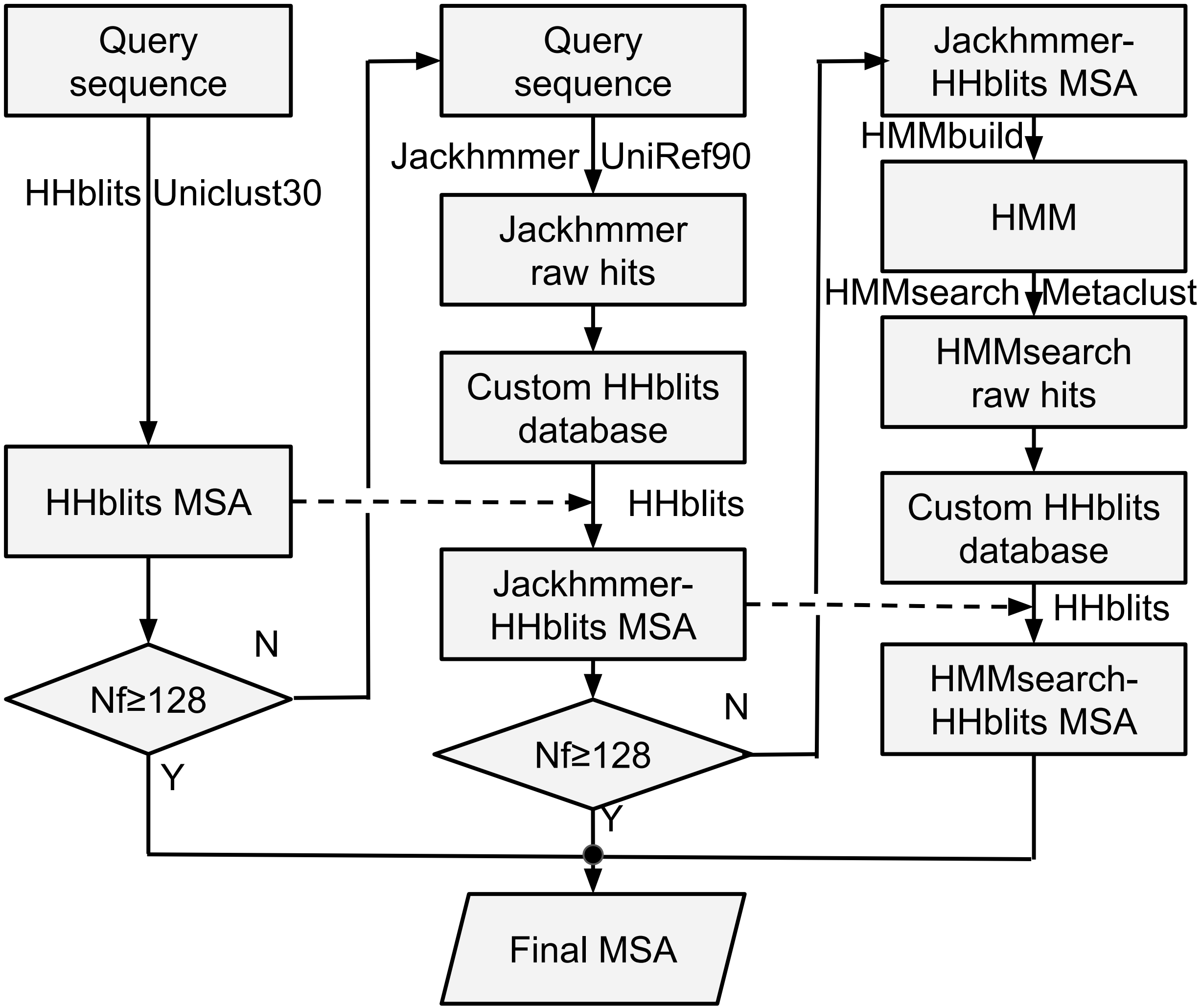


1. Deep MSA: build MSA from incremental sequence searching protocols

2. **Triple** coevolution features: covariance matrix, precision matrix, and pseudolikelihood maximization

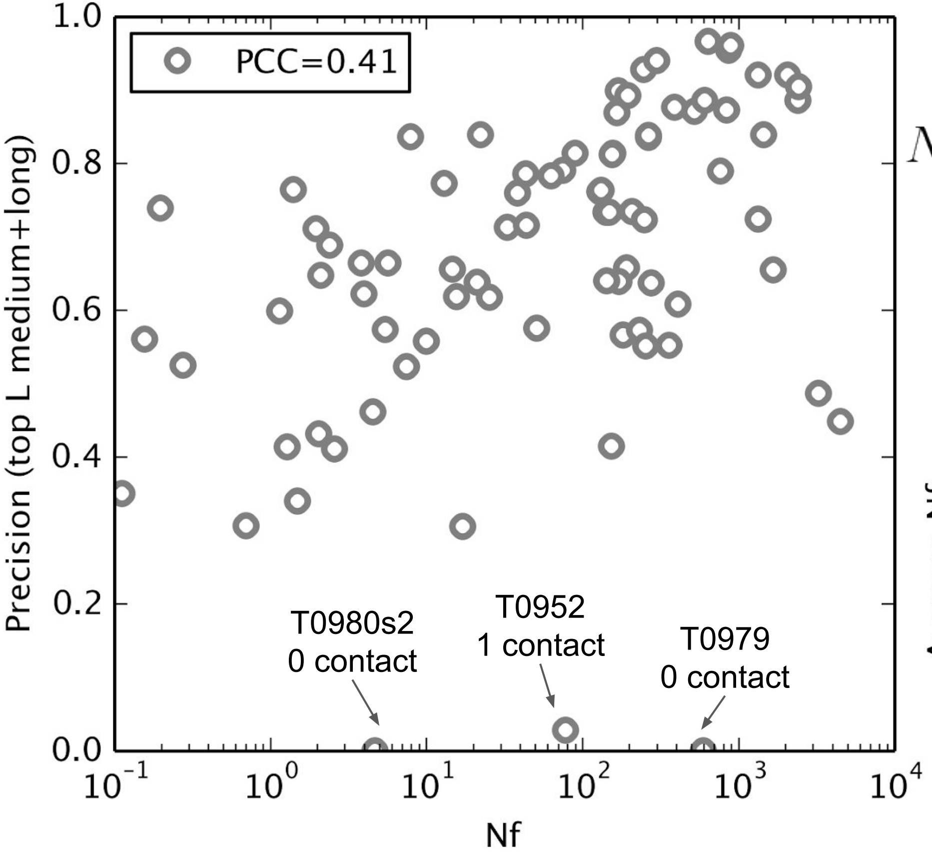
3. **Res**Net: fully convolutional neural network with residual blocks

Step 1: Deep MSA

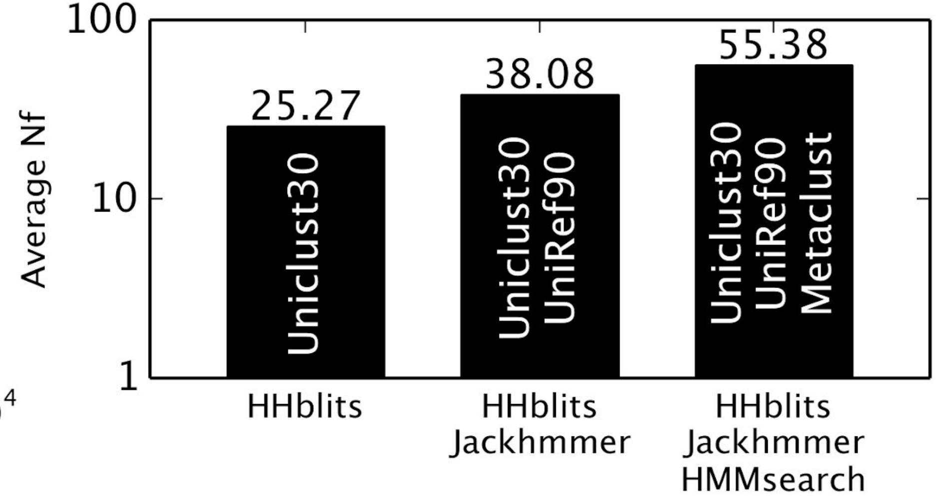


Nf: number of effective sequences in MSA

Step 1: Effect of MSA on contact prediction



$$Nf = \frac{1}{\sqrt{L}} \cdot \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} \mathbb{I}[S_{m,n} \geq 0.8]}$$



Step 2: Three feature matrices derived from MSA

1. COVariance matrix (COV) *S*:

$$S_{i,j}(a,b) = f_{i,j}(a,b) - f_i(a) \cdot f_j(b)$$

2. PREcison matrix (PRE) *θ*:

$$\theta = \underset{\theta}{argmin} \left(tr(S \cdot \theta) - log(det(\theta)) + \rho \cdot \|\theta\|_2^2 \right)$$

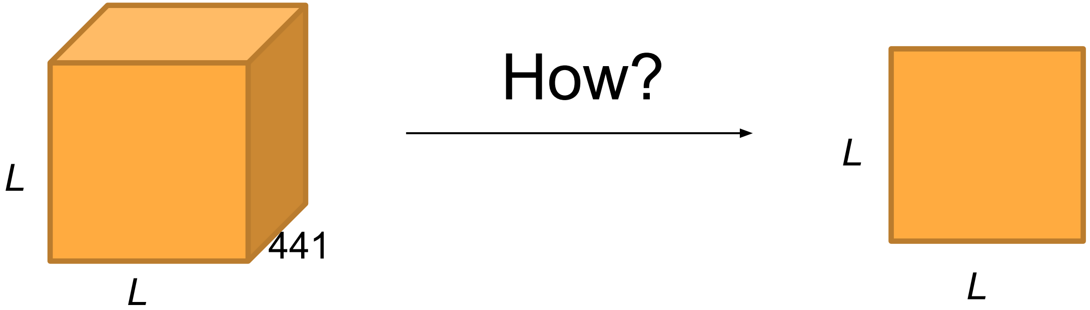
3. Pseudo-Likelihood Maximization (PLM) σ:

$$P(X|\sigma) = \prod_{n=1}^{N} \prod_{i=1}^{L} P(x_i^n | [x_1^n, ..., x_{i-1}^n, x_{i+1}^n, ..., x_L^N], \sigma)$$

$$= \prod_{n=1}^{N} \prod_{i=1}^{L} \frac{1}{Z_i^n} \cdot exp \left(\sigma_i(x_i^n) + \sum_{j=1, j \neq i}^{L} \sigma_{i,j}(x_i^n, x_j^n) \right)$$

Step 3: Predicting contact-map from features

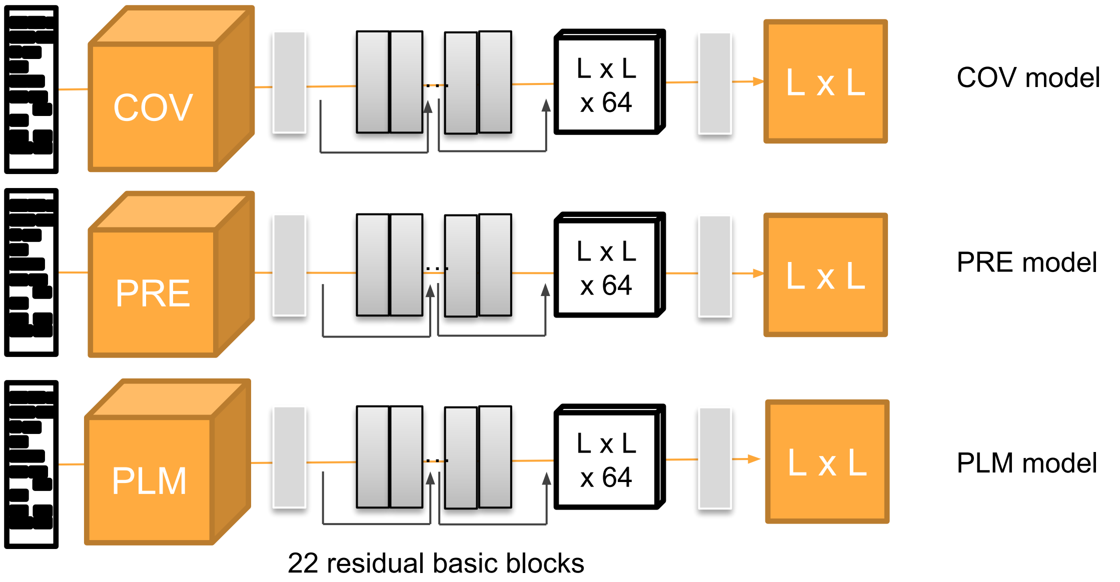
How do we convert $L \times L \times 441$ (=21 × 21) evolutionary coupling features to $L \times L \times 1$ contact map?



- L1 norm (λ=1) or L2 norm (λ=2): $$C_{i,j} = \left(\sum_{a,b} (\theta_{i,j}(a,b))^{\lambda} \right)^{1/\lambda}$$
- Or just leave it to deep learning: ResTriplet/TripletRes

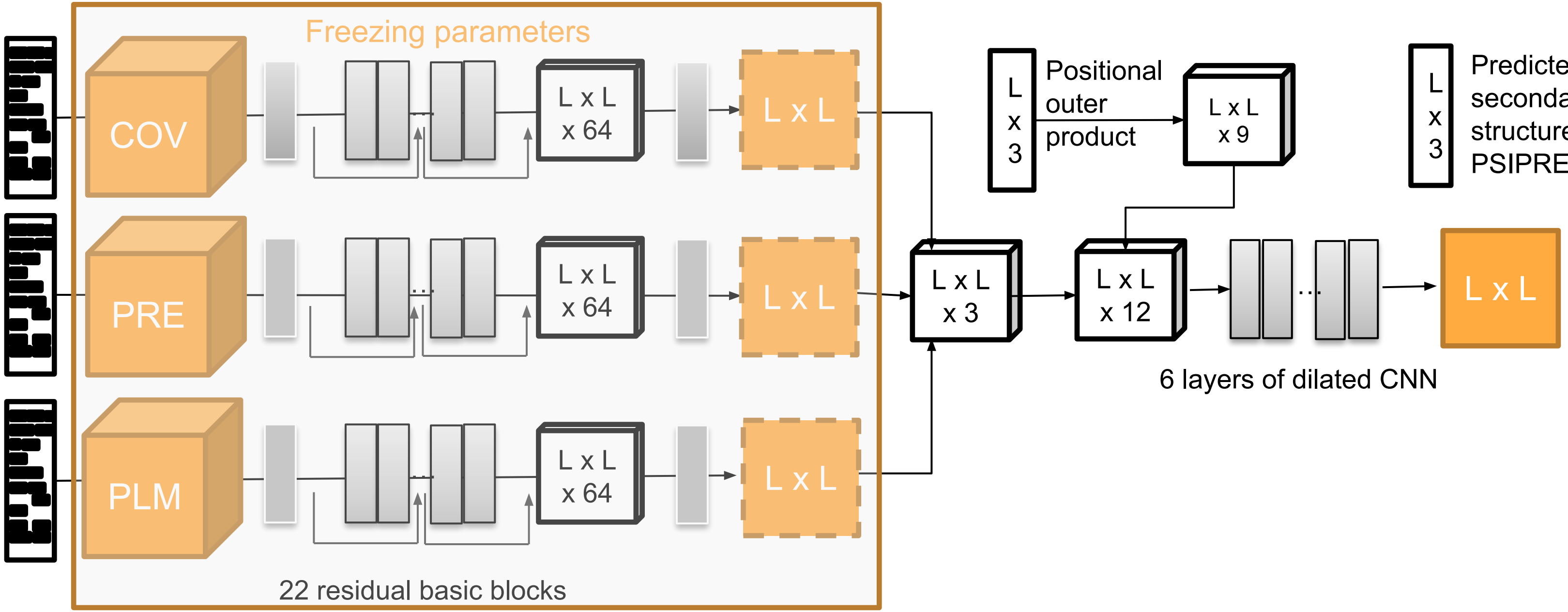
Step 3: ResTriplet neural network architecture

- First, train CNN models on COV, PRE and PLM features, separately.



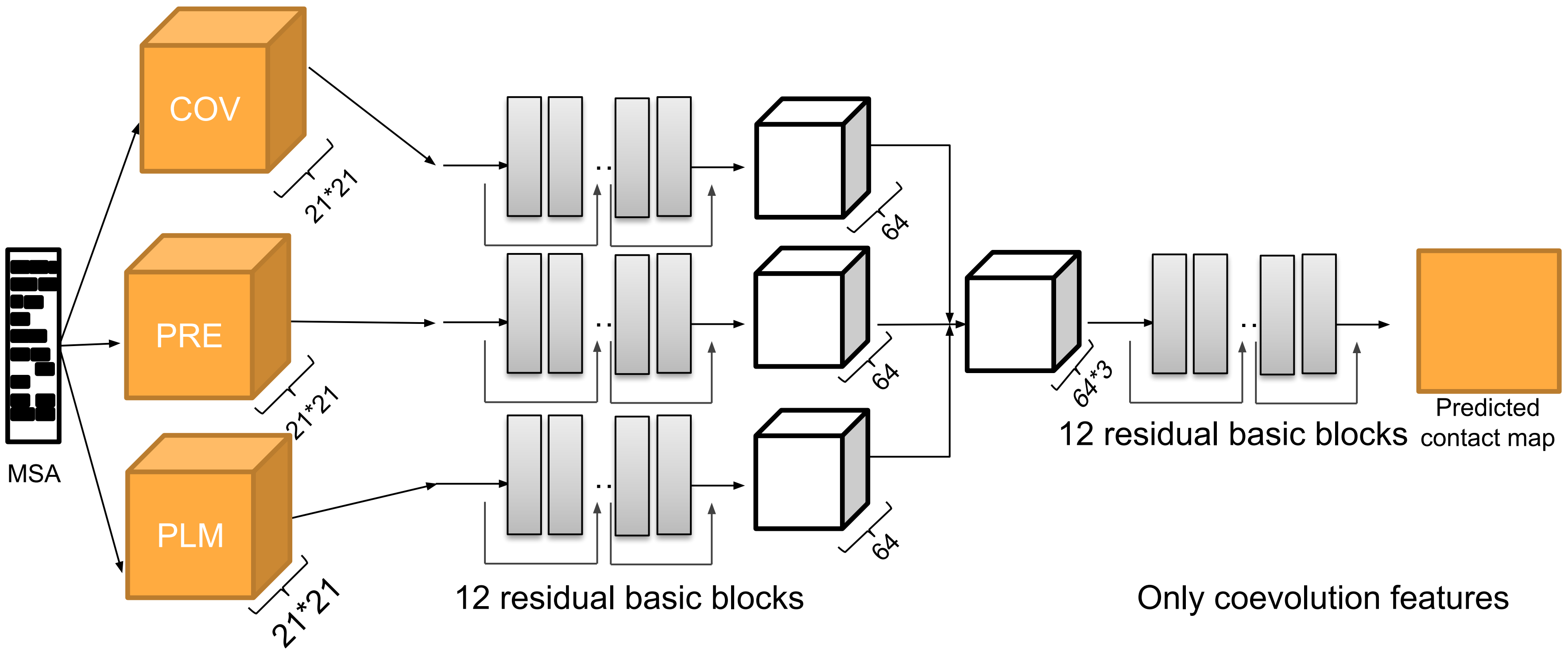
Step 3: ResTriplet neural network architecture

- First, train CNN models on COV, PRE and PLM features, separately.
- Second, stack 3 models with another dilated CNN model, with additional secondary structure features.

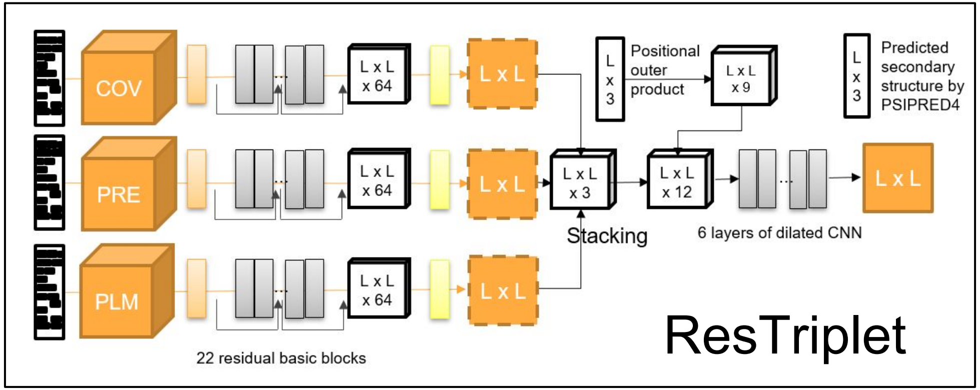


Step 3: TripletRes neural network architecture

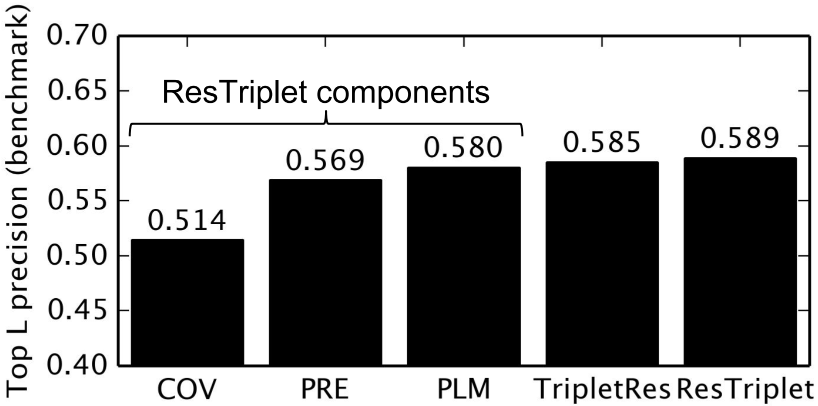
Train all CNN models together, in an end-to-end fashion.

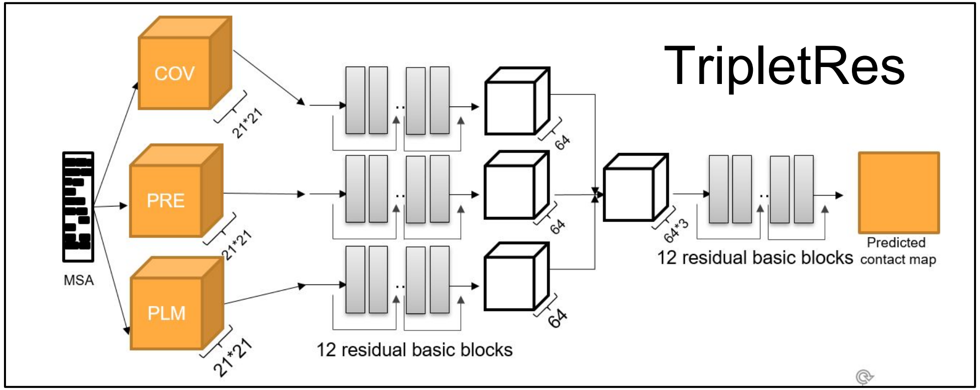


Step 3: ResTriplet vs TripletRes: neural network architecture



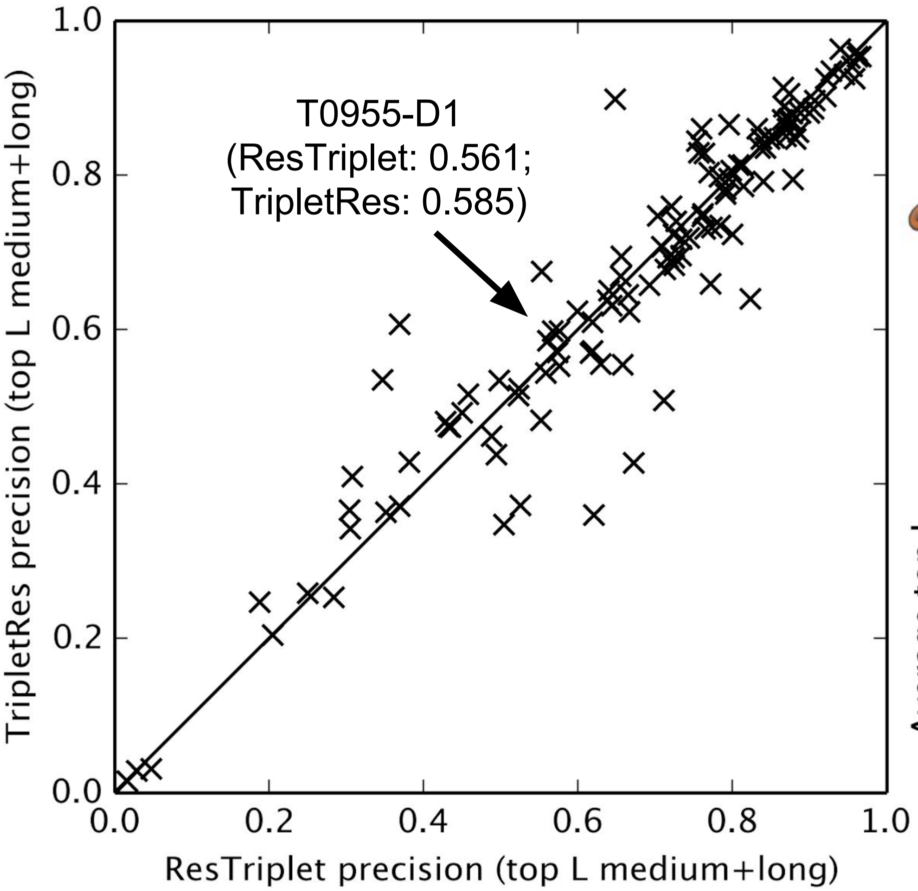
- Coevolution features + predicted secondary structure feature
- Training 4 models separately
- Can be trained with 1 GPU

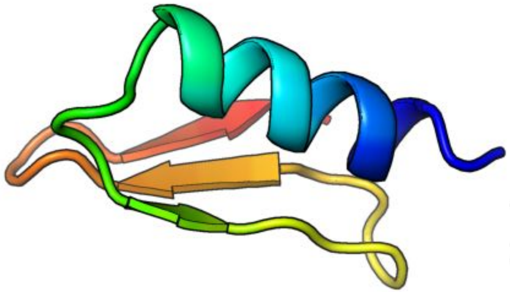




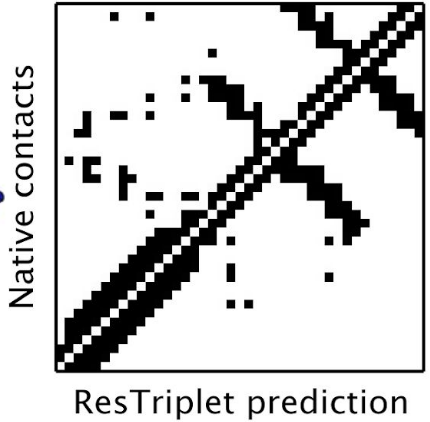
- ONLY coevolution features
- End-to-end training
- Requires 4 GPUs for training

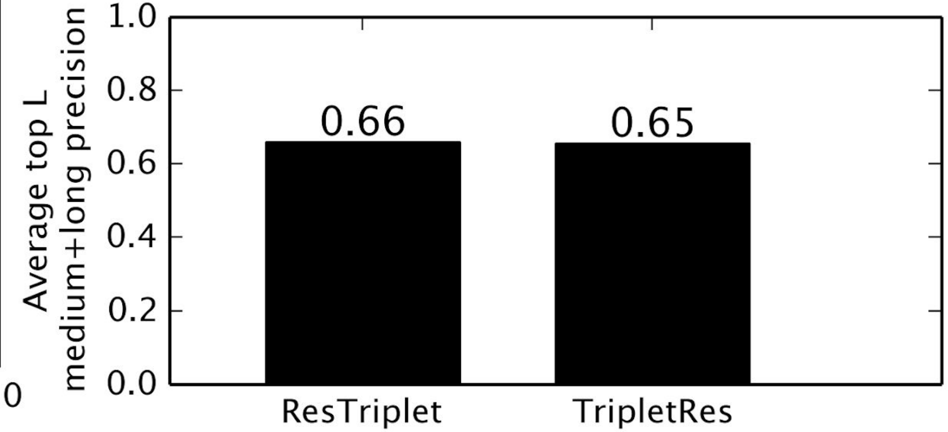
Result of ResTriplet/TripletRes on CASP13 Targets



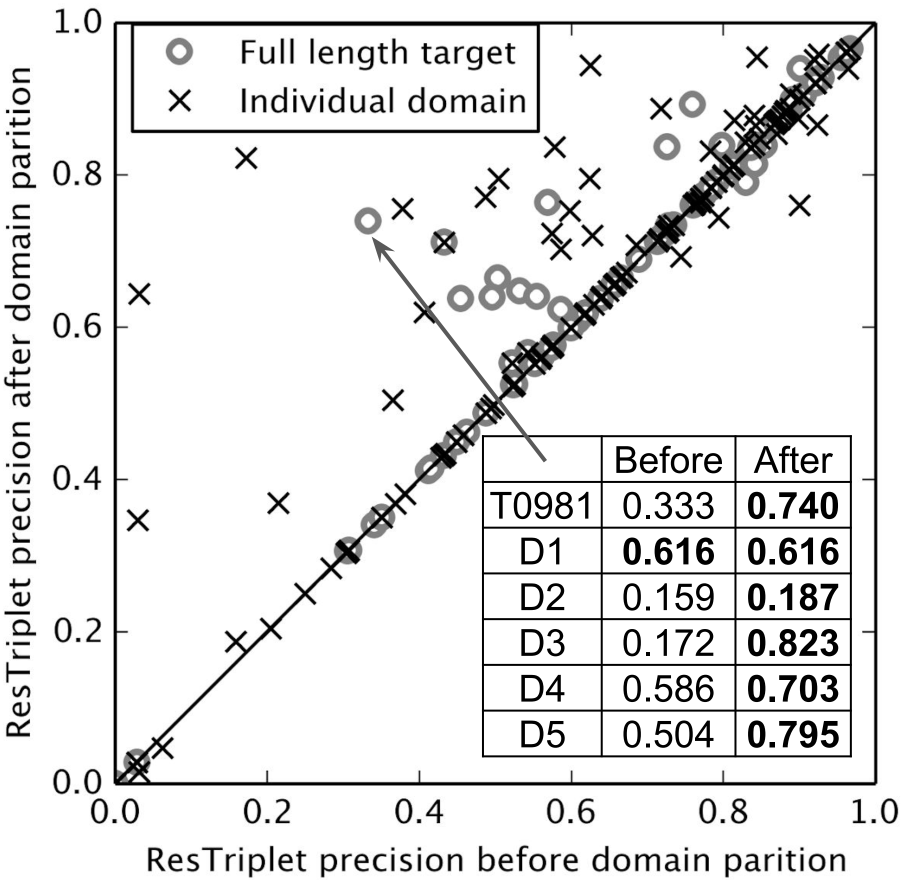


T0955-D1 (designed protein;
single sequence in MSA)

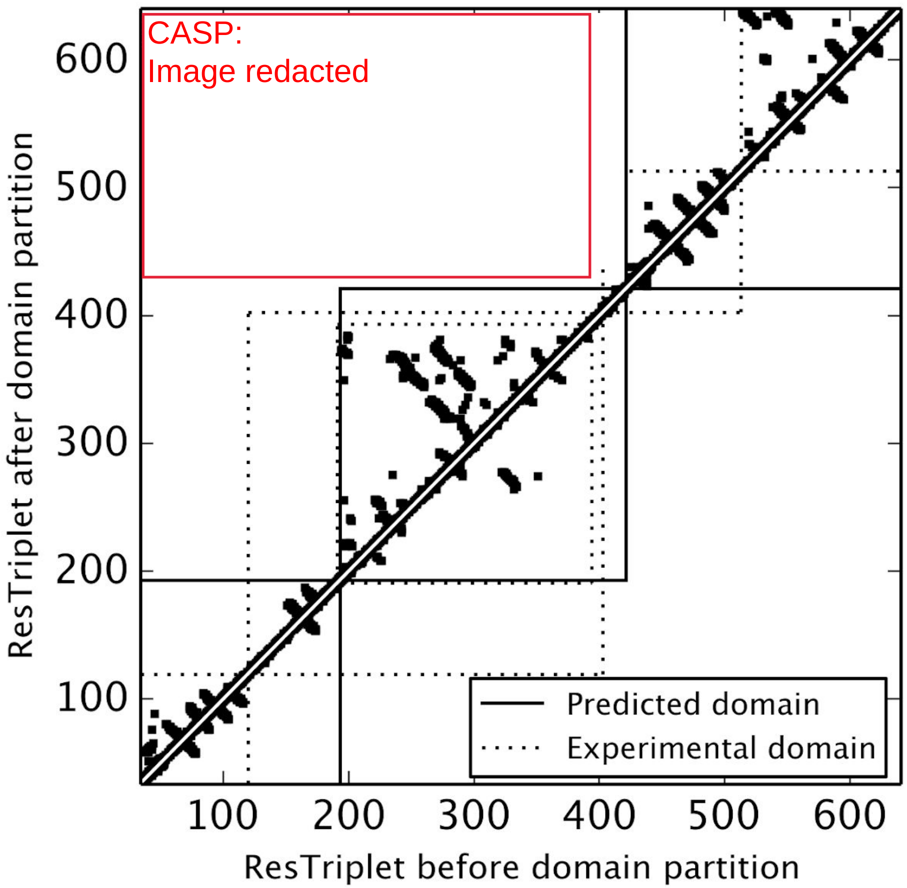




Effect of Domain Partition on Contact Prediction



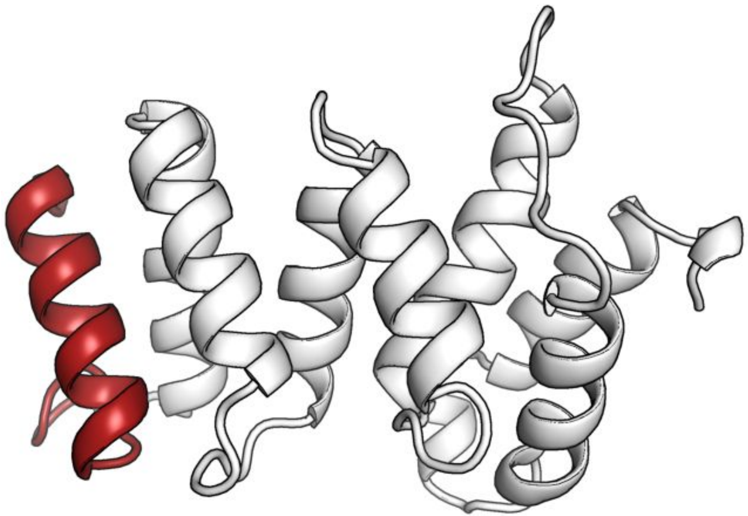
	Before	After
T0981	0.333	**0.740**
D1	**0.616**	**0.616**
D2	0.159	**0.187**
D3	0.172	**0.823**
D4	0.586	**0.703**
D5	0.504	**0.795**

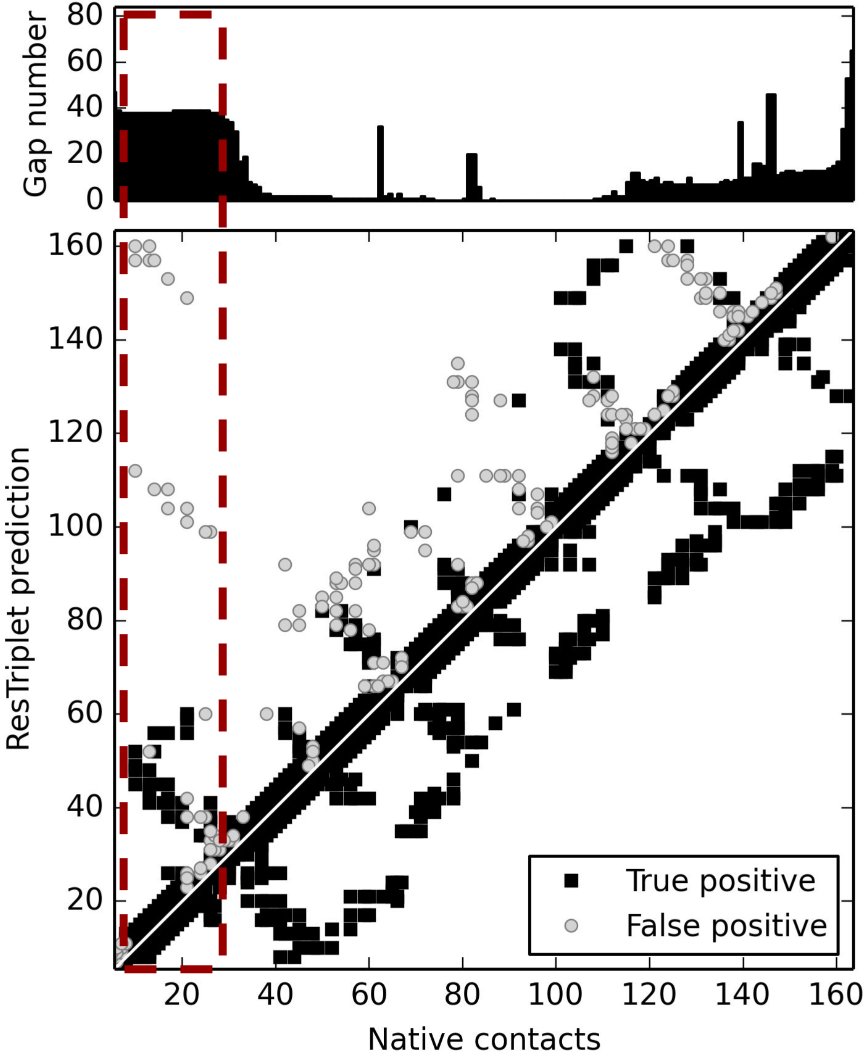


What went wrong?

T0957s2-D1: top *L* long range accuracy 0.342 (ResTriplet) and 0.394 (TripletRes)

Incorrectly predicted long range contacts for the first helix of T0957s2-D1 caused mainly by long stretch of gaps in MSA.





Summary

What went right?

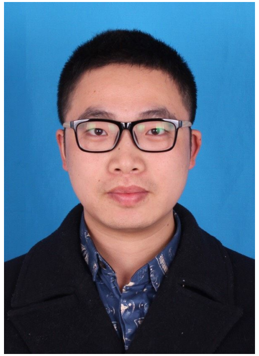
- DCA features (PRE and PLM) outperforms covariance feature (COV).
- Multiple feature fusion/ensemble with deep convolutional neural networks leads to highly accurate contact prediction.
- With the set of coevolutionary features, predicted one-dimensional features (secondary structure, sequence profile, solvent accessibility etc) is not strictly required for deep learning.
- Domain partition (even when domain boundary is not exact) improves precision.
- Combination of diverse multiple sequence alignment generation protocols (search algorithms and sequence databases) improves contact prediction.

What went wrong?

- How to appropriately consider large gaps in MSA is still an open question.

Acknowledgements

Zhang lab Research Group


Yang Li


Chengxin Zhang


Dong-Jun Yu


Yang Zhang

...and all other current and former members!

Funding and Resources

NSFC

NIH NIGMS

NSF

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

XSEDE
Extreme Science and Engineering Discovery Environment

Thank You!