Structural Bioinformatics

GENOME 541

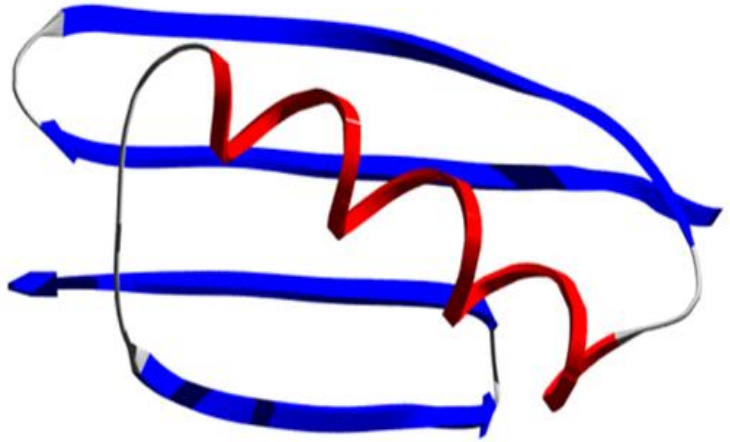
Spring 2025

Lecture 3: Protein Structure Prediction

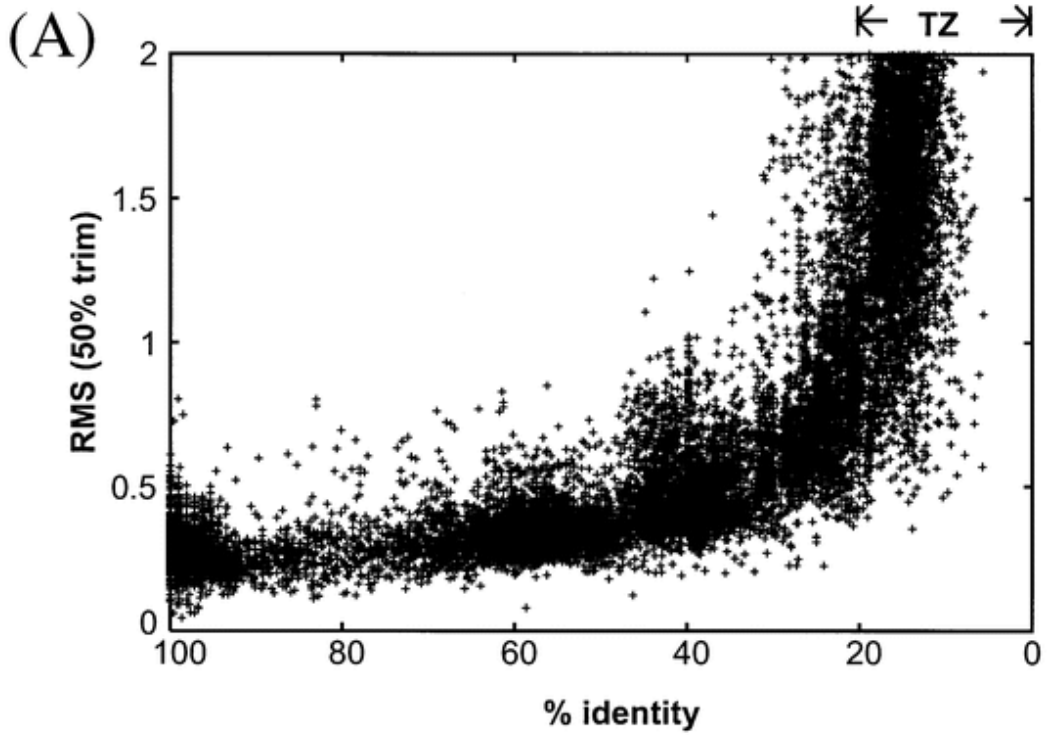
Frank DiMaio (dimaio@uw.edu)

Structure Prediction

DEIVKMSPIIRFYSSGNAGLRTYIGDHK
SCVMCTYWQNLLTYESGILLPQRSRTSR



Prediction Strategies



Wilson, Kreychman, Gerstein (2000)

Homology Modeling

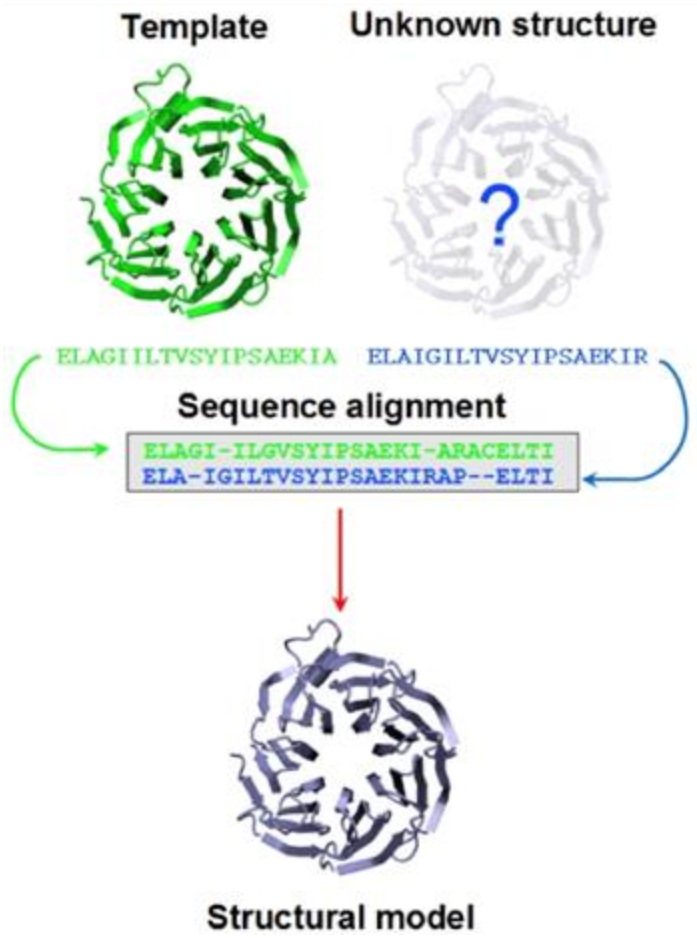
- Proteins that share similar sequences share similar folds.
- Use known structures as the starting point for model building.

De Novo Structure Prediction

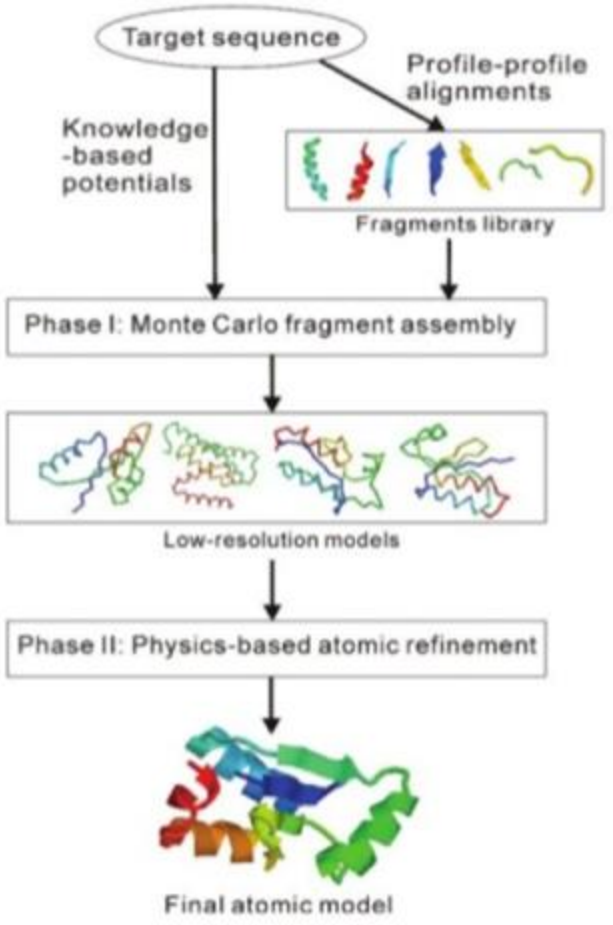
- Do not rely on global similarity with proteins of known structure
- Folds the protein from the unfolded state.

A (very) brief introduction to pre-ML protein structure prediction

Template-based Modeling



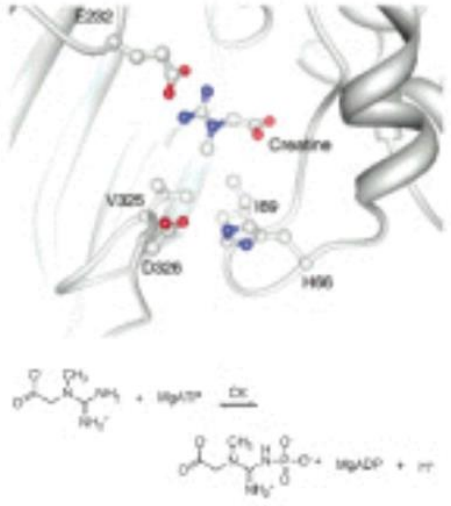
(Template-)Free Modeling



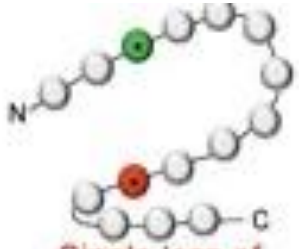
Coevolution guided modeling

Sequence → Structure → Function

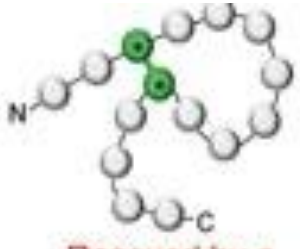
MPFGNTHNKFKL
NYKPEEEYPDLSK
HNNHMAKVLTLE
LYKKLRDKETPSGF
TVDDVIQTGVDNP
GHPFIMTVGCVAG
DEESYEVFKELFDPI
ISDRHGGYKPTD...



Initial sequence

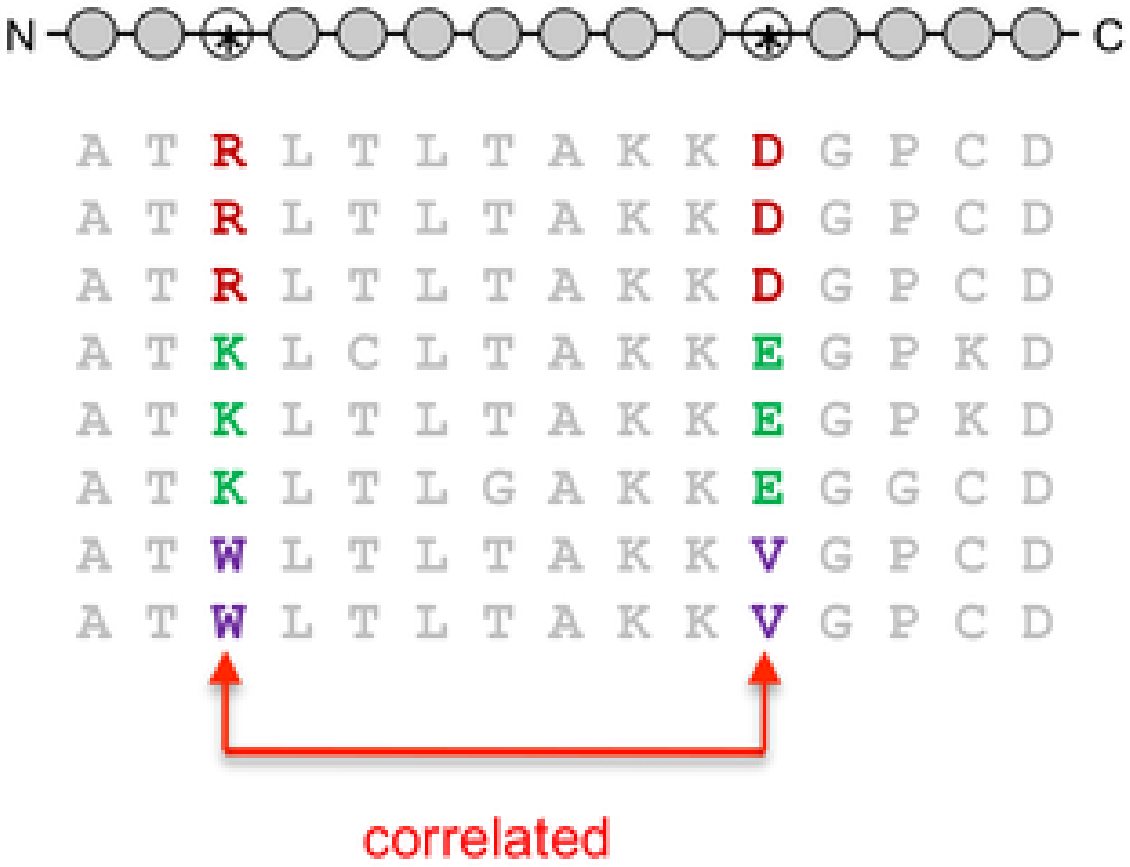


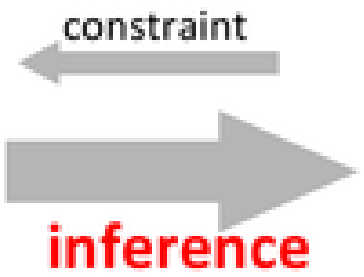
Single loss of function mutation

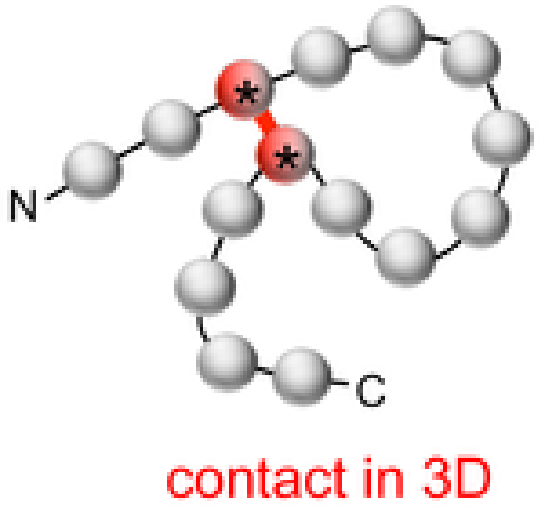


Rescued by a compensating mutation

Correlated mutations carry information about distance relationships in protein structure.







Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766. https://doi.org/10.1371/journal.pone.0028766
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

Learning the DCA (direct coupling analysis) matrix

The essence of DCA is then to assume that the rows, i.e. our aligned homologous proteins, are independent events drawn from a Potts-model probability distribution,

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \frac{1}{2} \sum_{i,j=1}^{N} J_{ij}(\sigma_i, \sigma_j) \right), \tag{1}$$

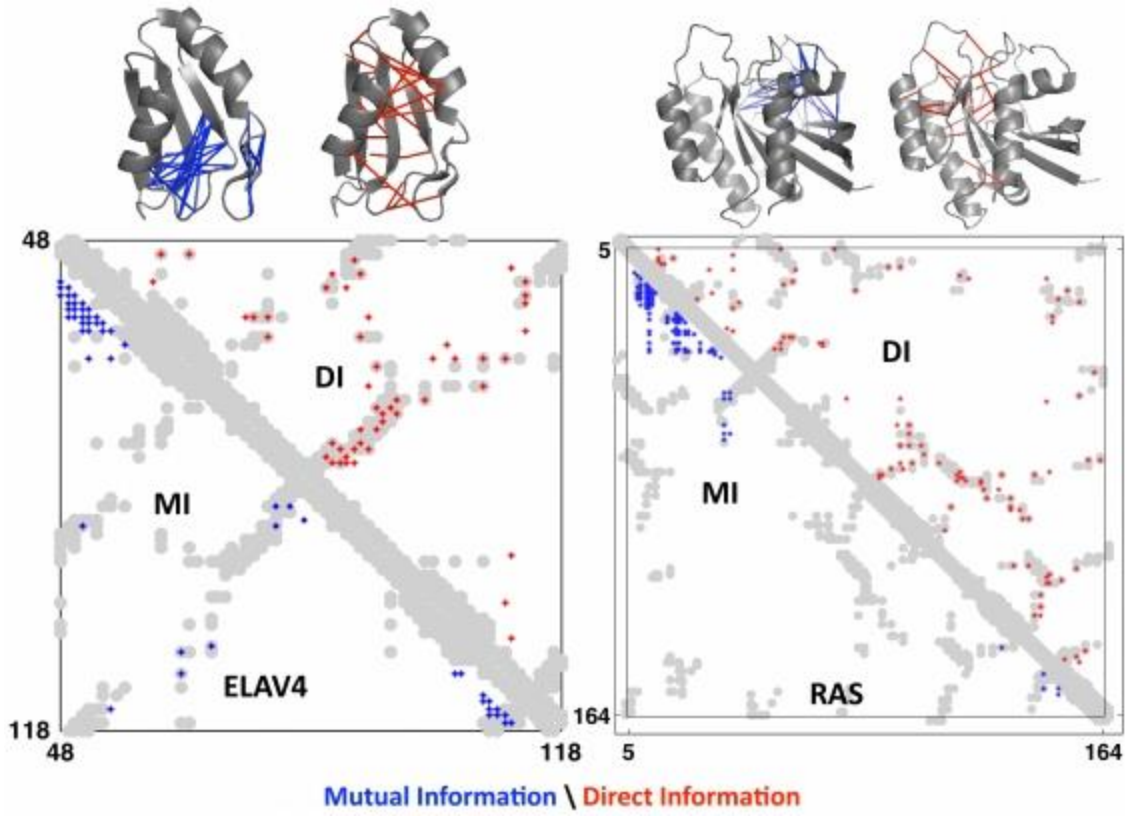
and to use the interaction parameters $\mathbf{J}_{ij}$ as predictions of spatial proximity among amino-acid pairs in the protein structure.

Problem: **Z** cannot be tractably computed

Solutions:

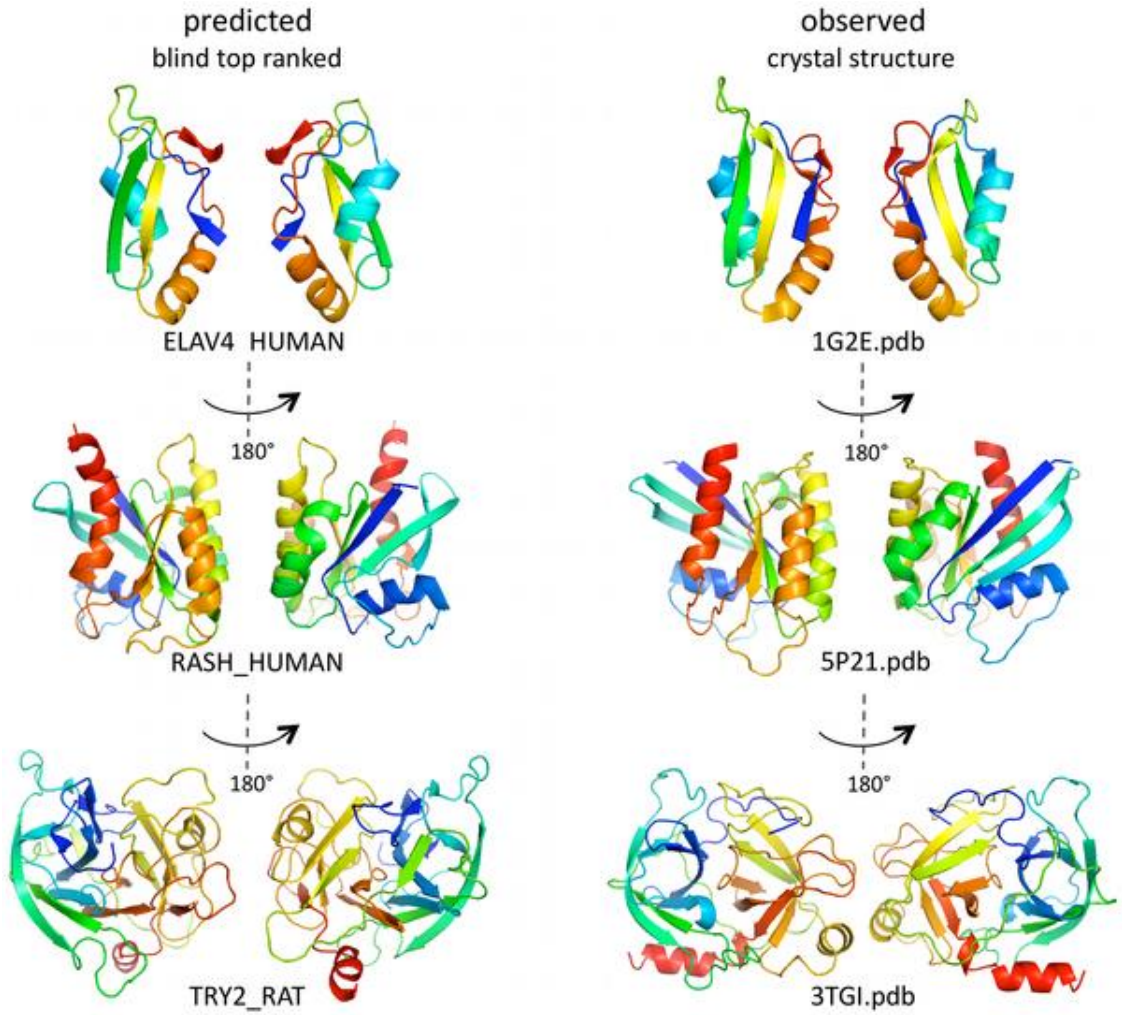
- Mean-field approach (mfDCA) (https://www.pnas.org/content/108/49/E1293)
- Pseudo-likelihood (plmDCA) (https://journals.aps.org/pre/abstract/10.1103/PhysRevE.87.012707)

Correlated mutations carry information about distance relationships in protein structure.



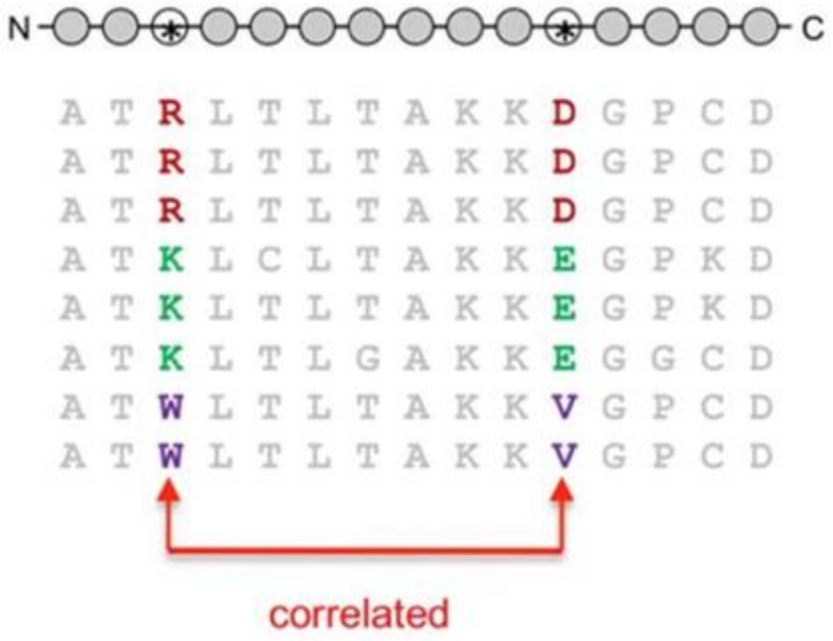
$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\exp\left(\sum_{i=1}^{L}\left[\mathbf{v}_i(x_i)+\sum_{j>i}^{L}\mathbf{w}_{ij}(x_i,x_j)\right]\right).$$

Predicted 3D structures for three representative proteins.

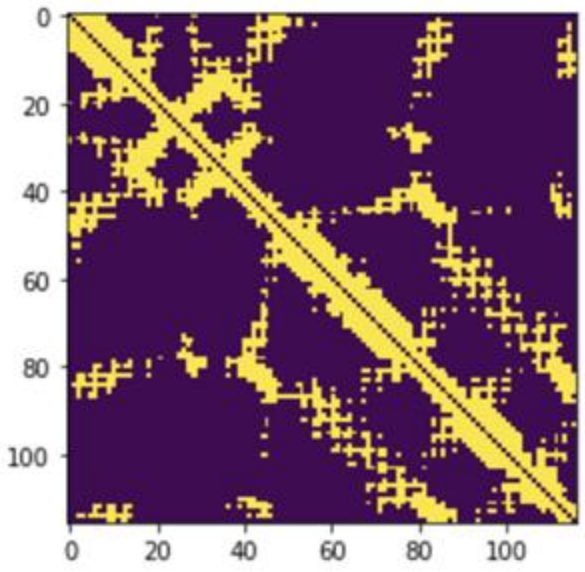


Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766. https://doi.org/10.1371/journal.pone.0028766
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

Coevolution guided modeling

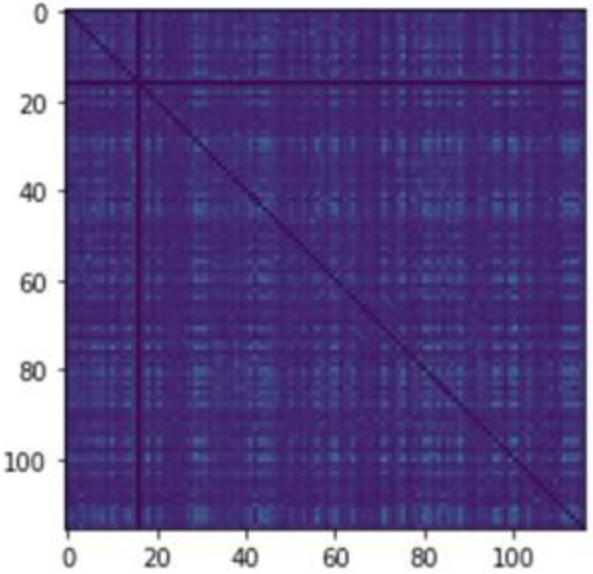


Native contact map



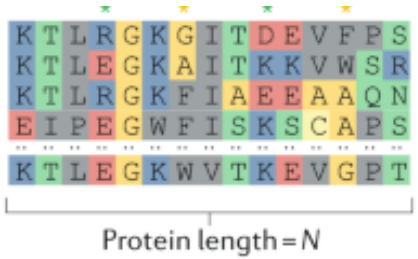
GREMLIN predictions on shallow MSAs

(Nseq=36, Nf=2.3)

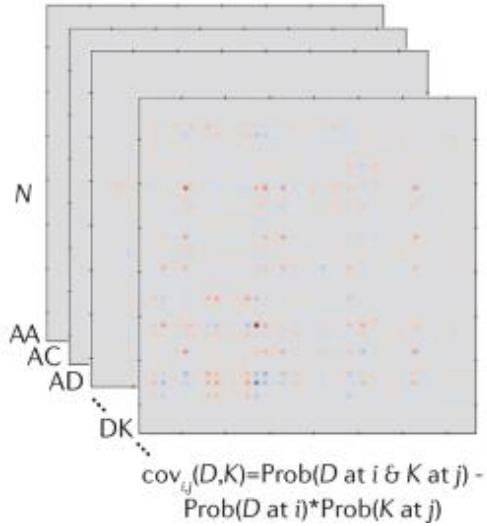


Contact maps = Computer Images?

MSA

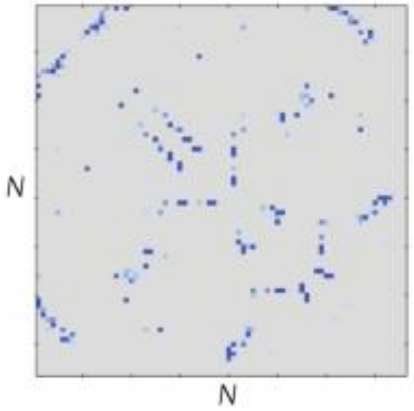


Covariances

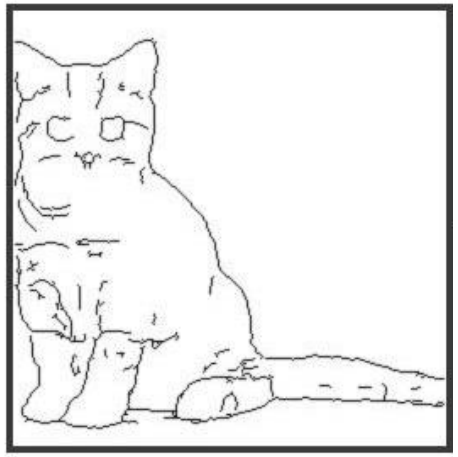


Contact prediction

Contacts



INPUT



AI-based
image processing

OUTPUT

PLOS COMPUTATIONAL BIOLOGY

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*

Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America

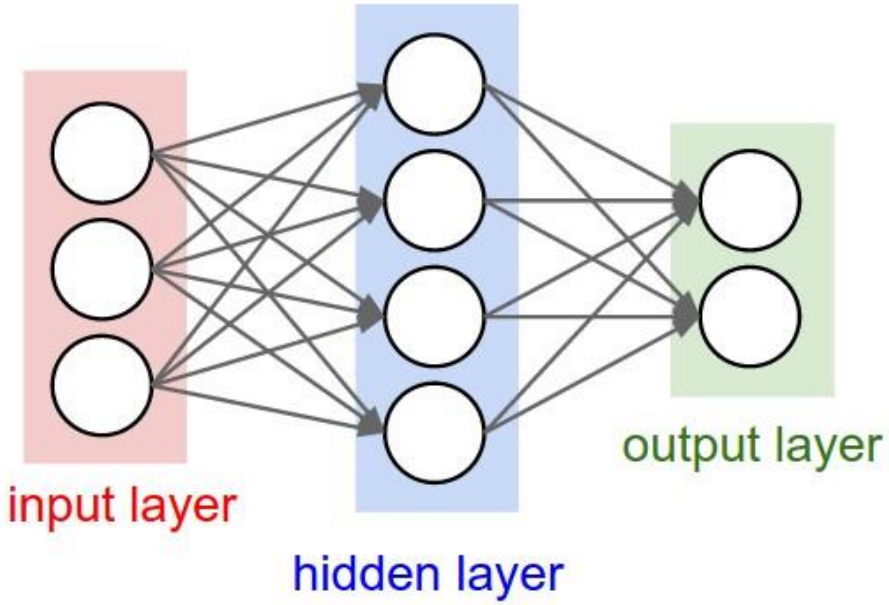
☯ These authors contributed equally to this work.
* jinboxu@gmail.com

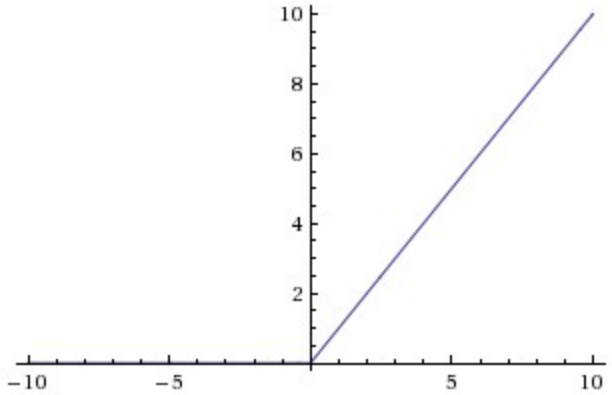
Abstract

Motivation

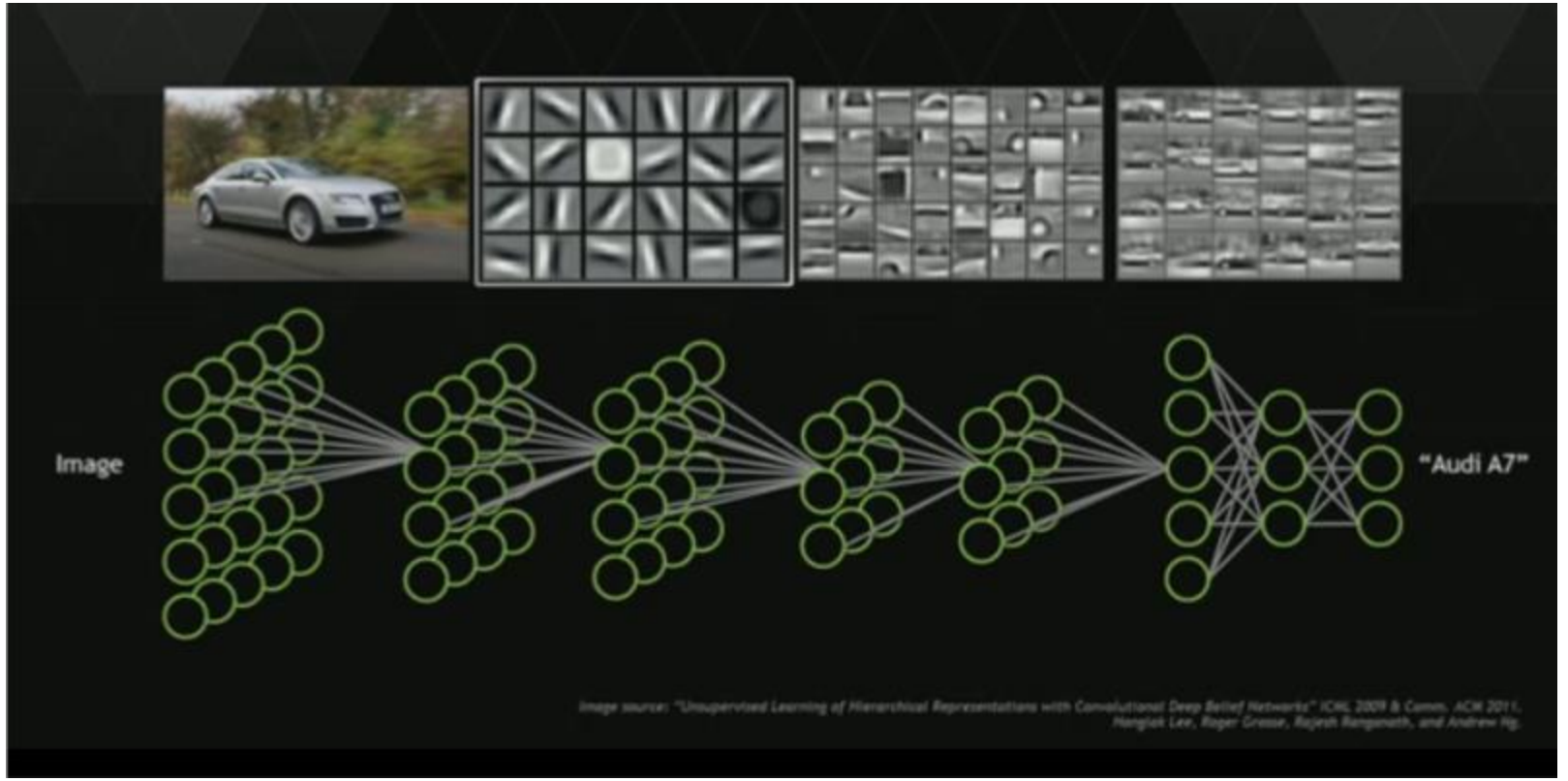
Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still of low quality and not very useful for de novo structure prediction.

Neural networks

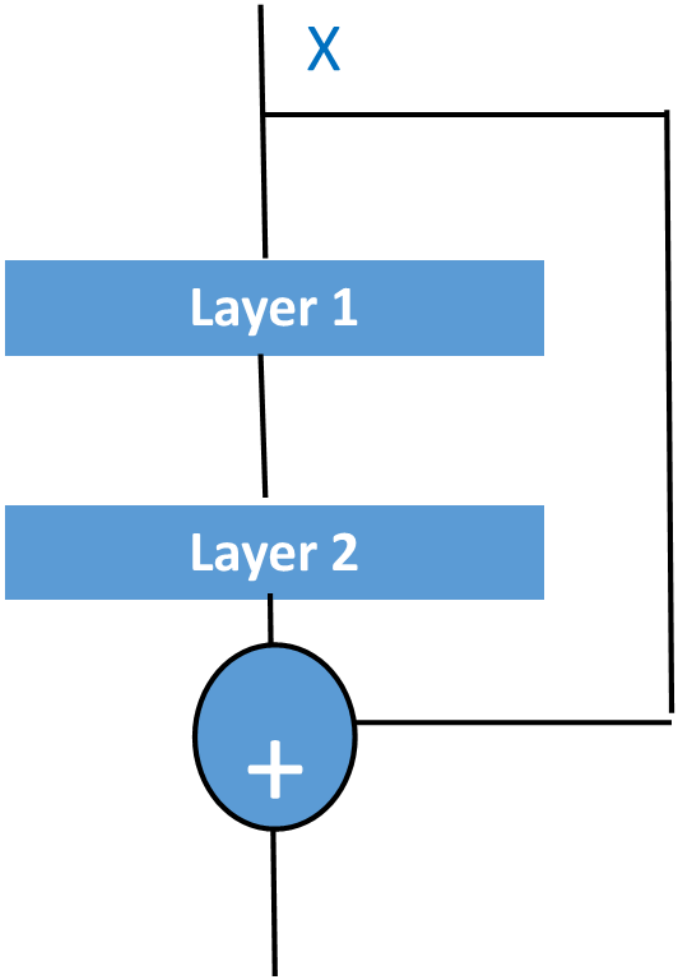


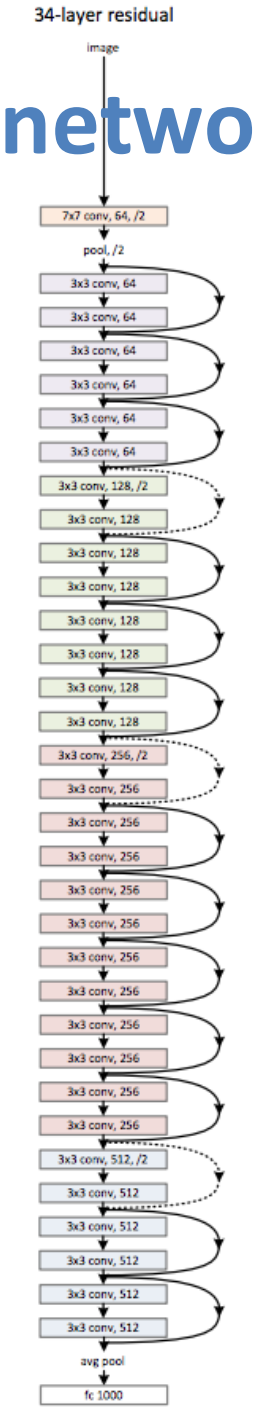


Convolutional neural networks

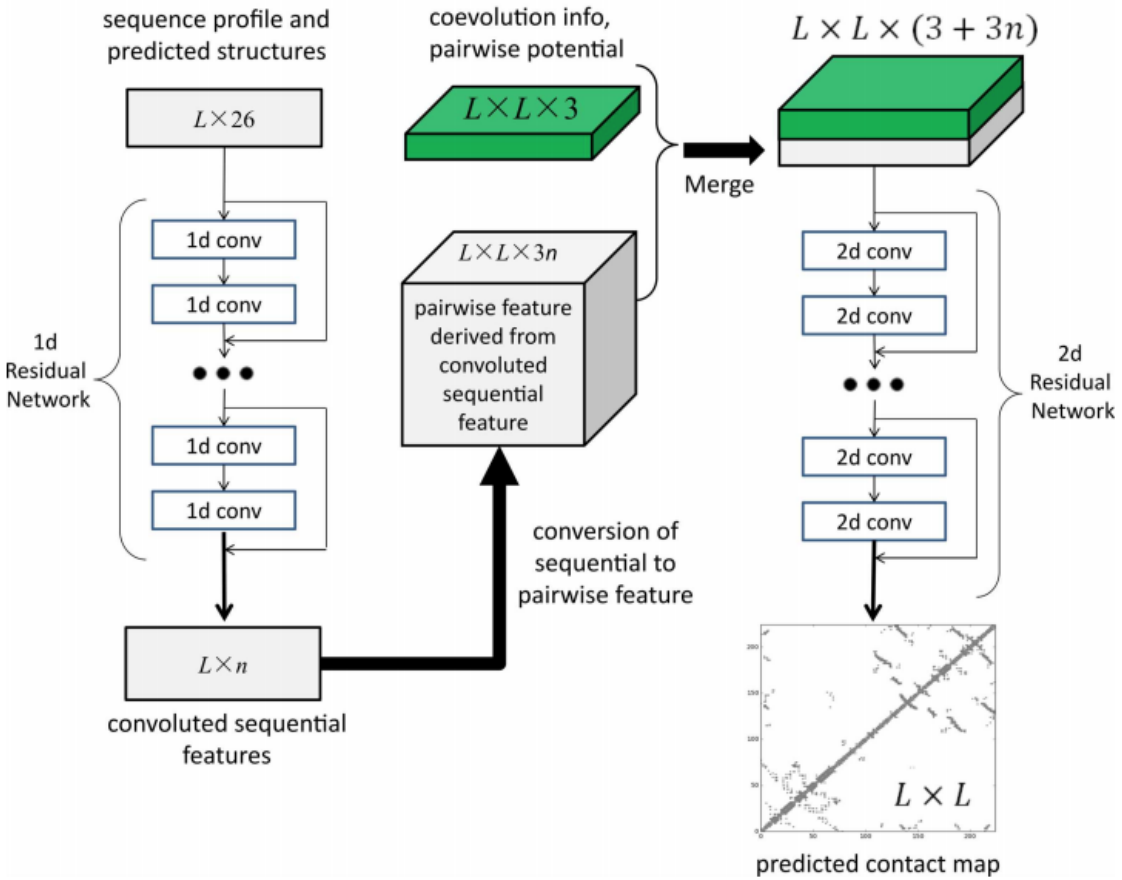


Residual connections allow deep networks





Learning a contact map from co-evolving residues



Inferring better contact maps (I)

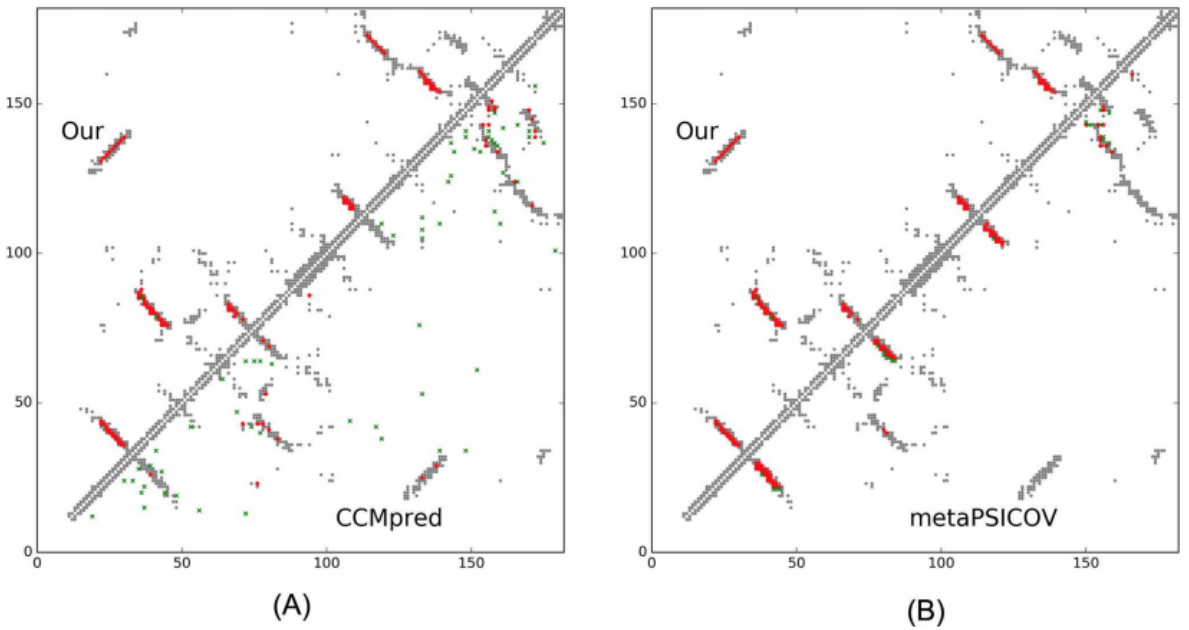
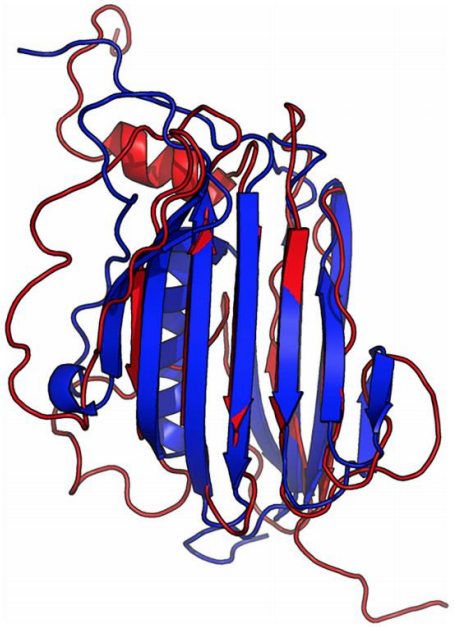


Fig 6. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 2nc8A. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The comparison between our prediction (in upper-left triangle) and MetaPSICOV (in lower-right triangle).



Inferring better contact maps (II)

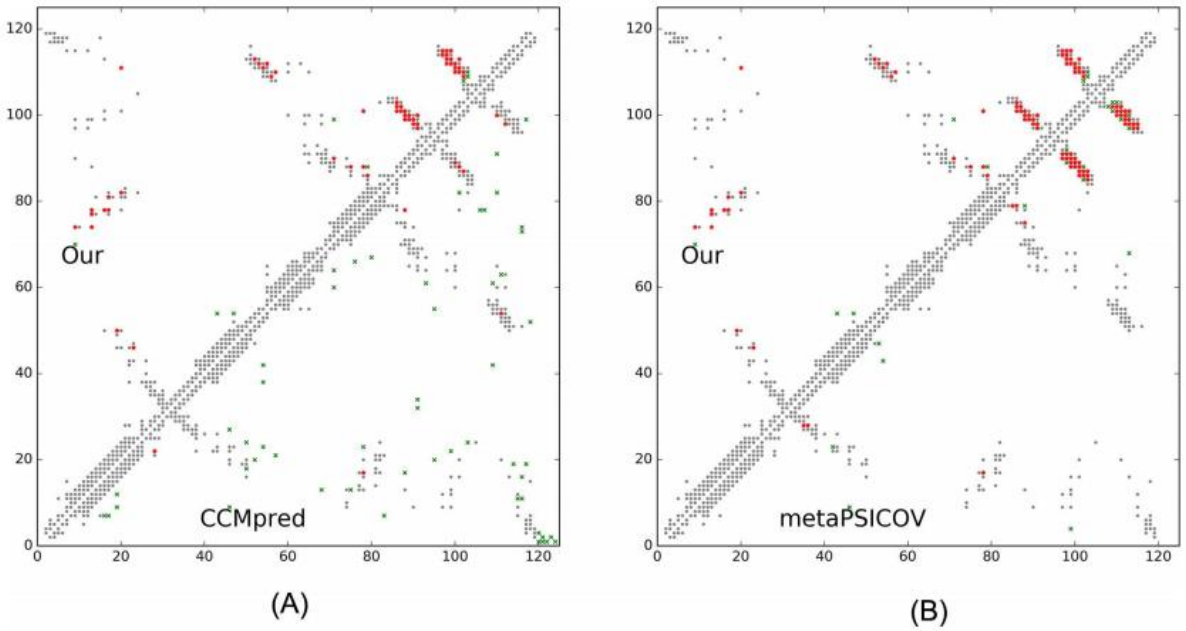
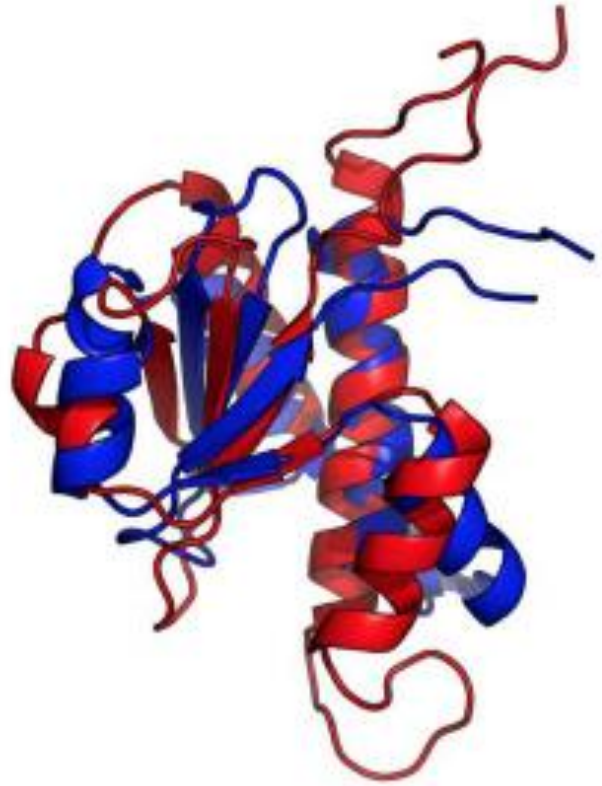
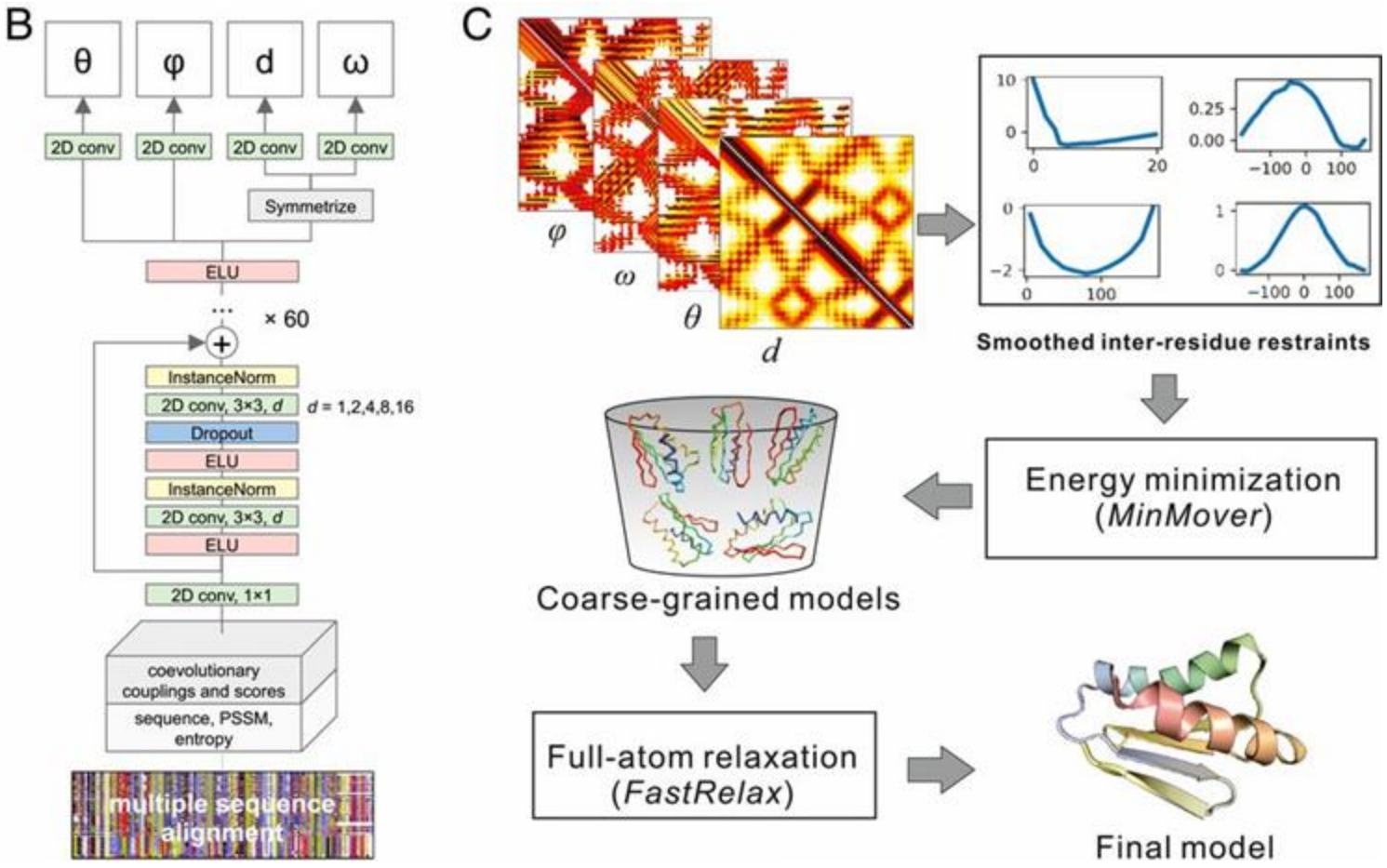


Fig 9. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 5dcjA. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The comparison between our prediction (in upper-left triangle) and MetaPSICOV (in lower-right triangle).



trRosetta



Improved protein structure prediction using predicted interresidue orientations

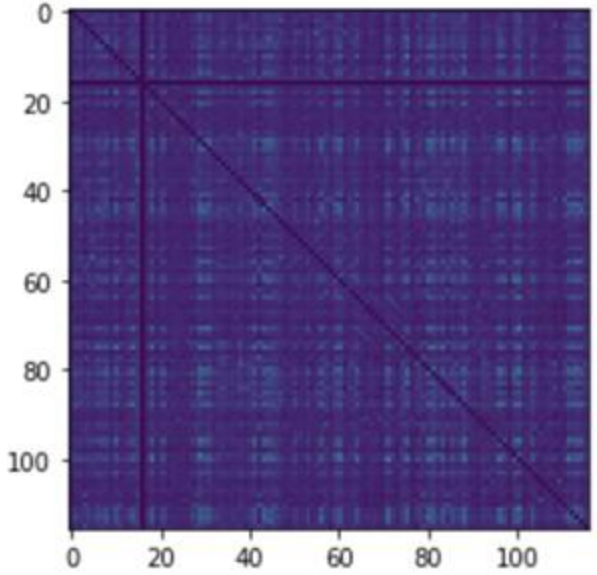
Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker

PNAS January 21, 2020 117 (3) 1496-1503; first published January 2, 2020 https://doi-org.offcampus.lib.washington.edu/10.1073/pnas.1914677117

Discovering hidden patterns with a learned model

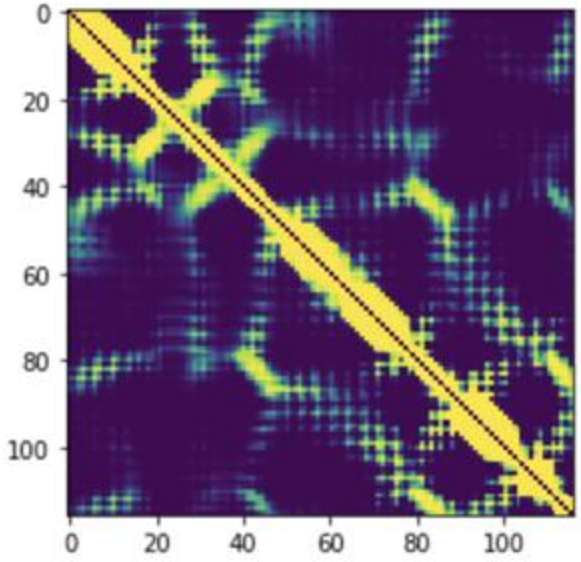
Gremlin predictions on shallow MSAs

(Nseq=36, Nf=2.3)

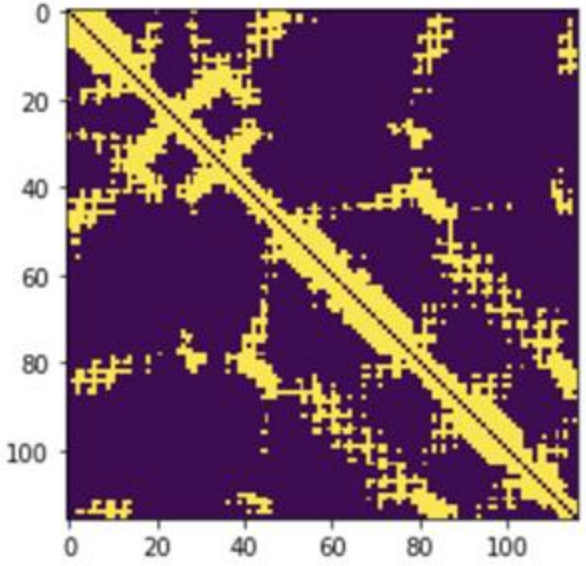


trRosetta predictions on shallow MSAs

(Nseq=36, Nf=2.3)



Native contact map



Improving protein structure prediction

Free modeling accuracy in CASP

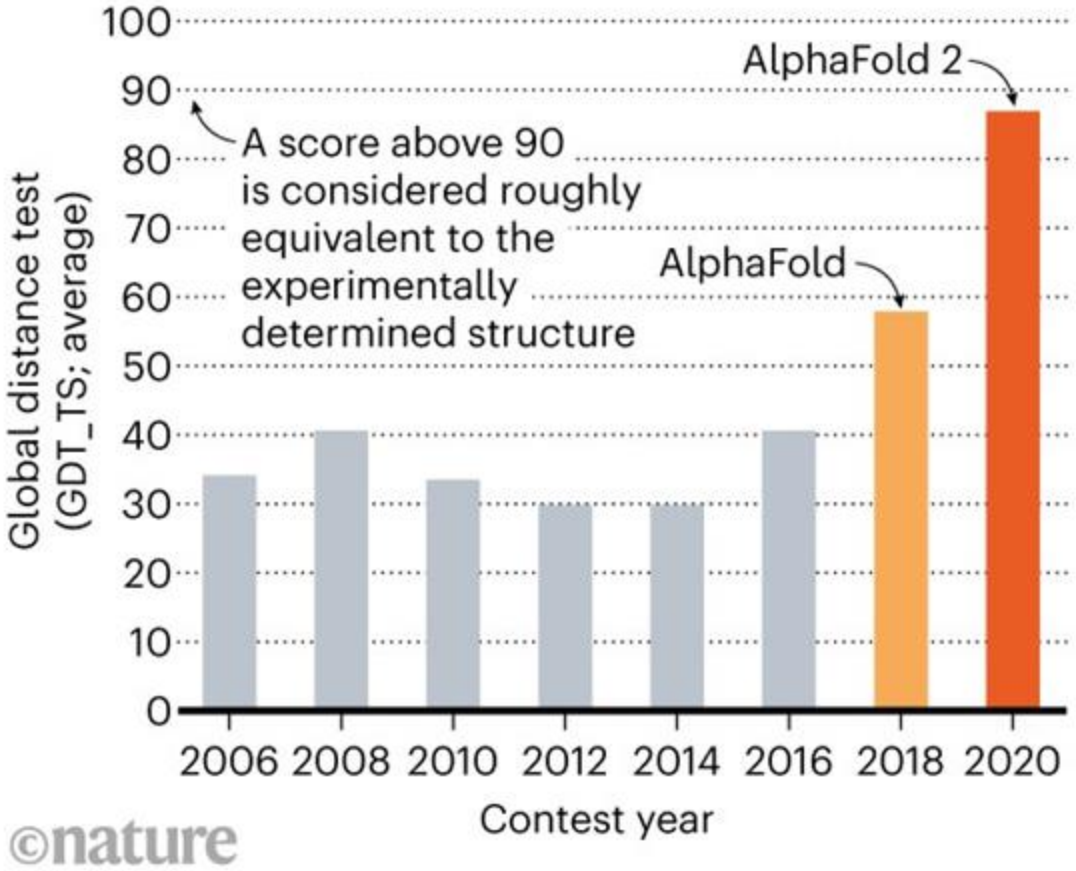
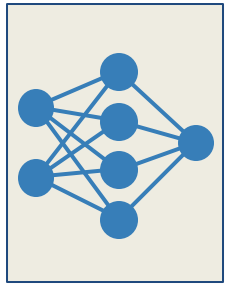


Figure from Callaway E. *Nature*, 2020

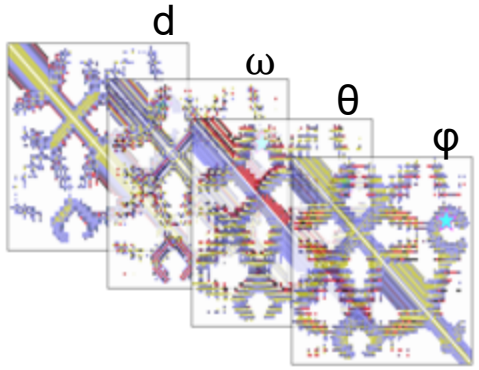
A differentiable end-to-end structure predictor

trRosetta

input MSA

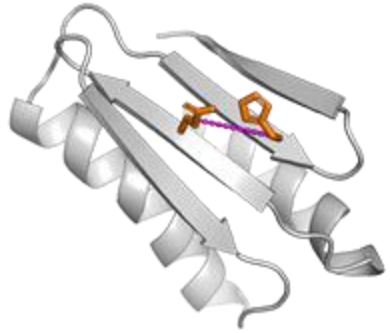


trRosetta



Fold by minimization & relax

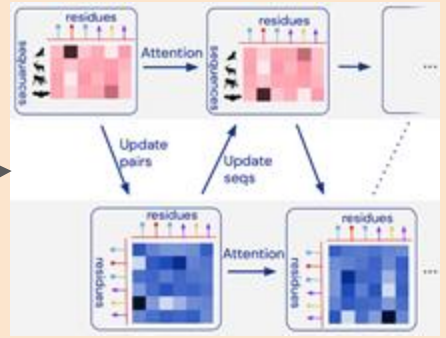
PyRosetta



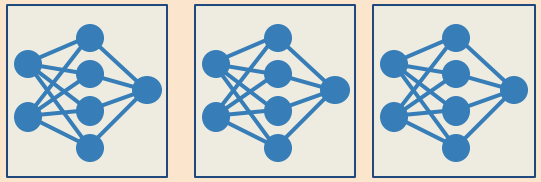
DeepMind

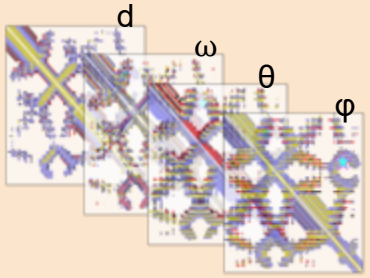
input MSA

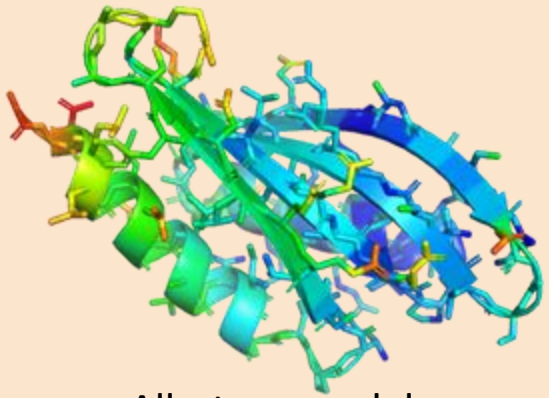
Attention-based "trunk"



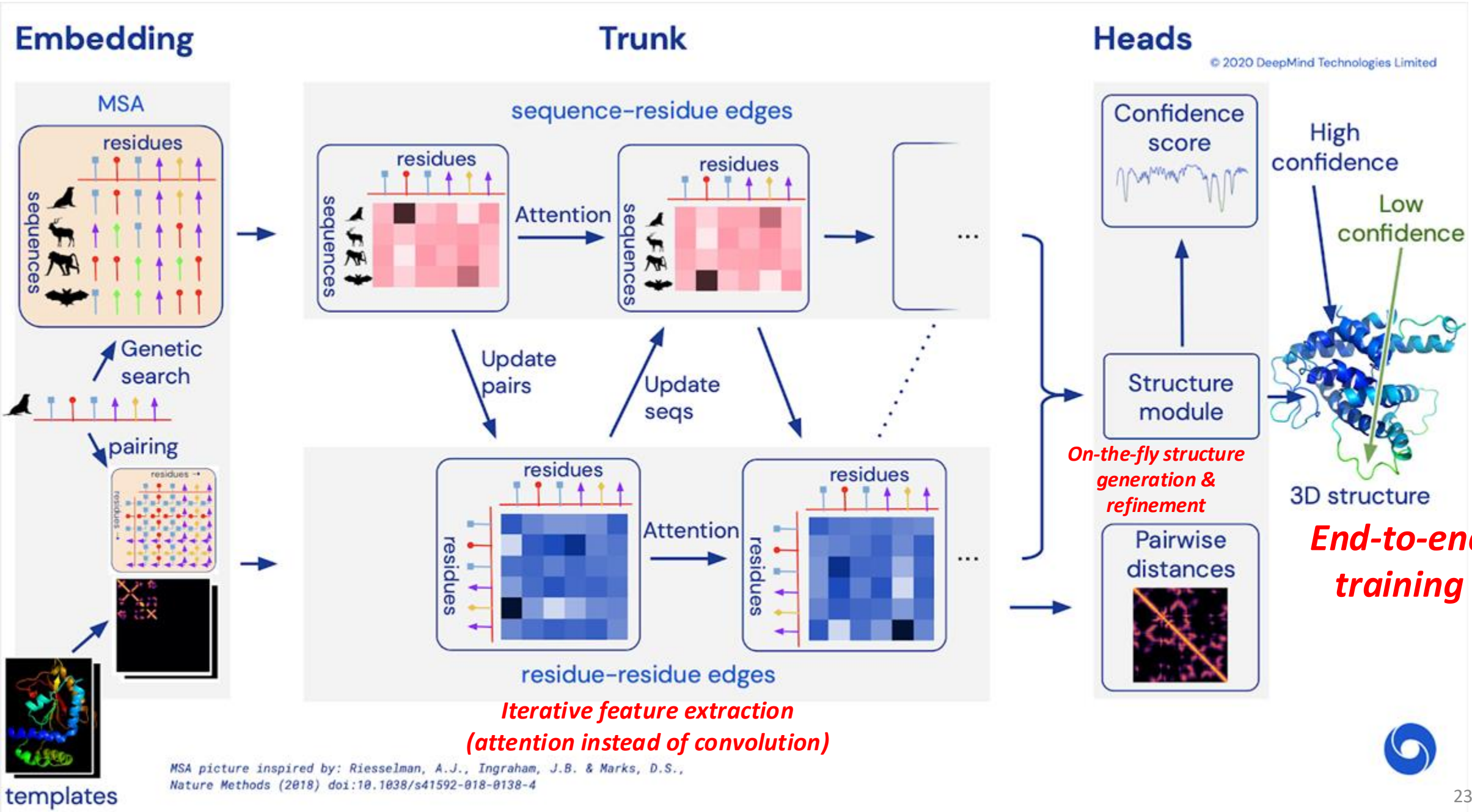
SE3-invariant transformer layers



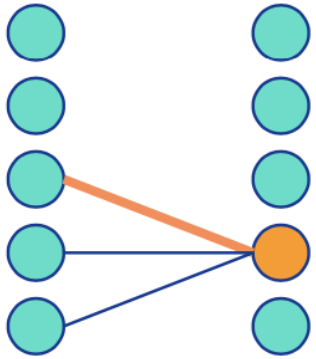




All-atom model + predicted errors

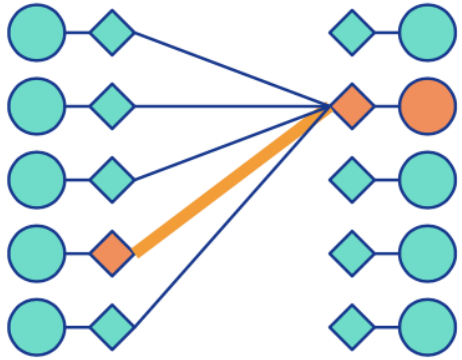

Embedding
Trunk
Heads
© 2020 DeepMind Technologies Limited
MSA
residues
sequences
Genetic search
pairing
templates
sequence-residue edges
Attention
Update pairs
Update seqs
residue-residue edges
Confidence score
High confidence
Low confidence
Structure module
3D structure
On-the-fly structure generation & refinement
Pairwise distances
End-to-end training
Iterative feature extraction
(attention instead of convolution)
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S.,
Nature Methods (2018) doi:10.1038/s41592-018-0138-4

What would be a proper inductive bias for protein structure prediction?



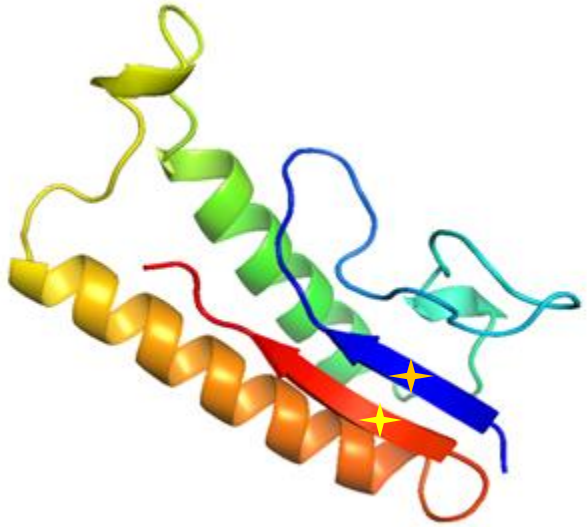
Convolutional Networks (e.g. computer vision)

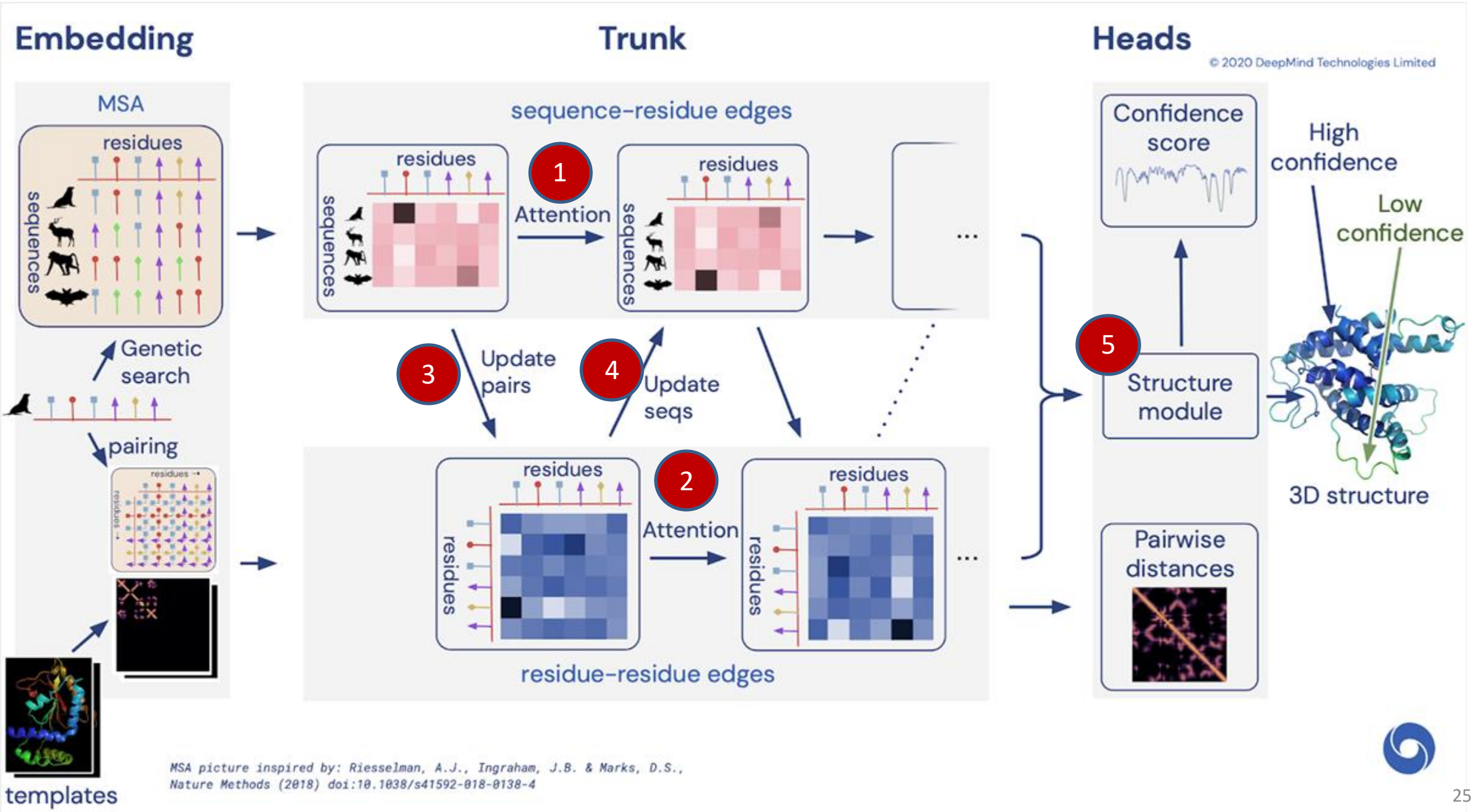
- data in regular grid
- information flow to local neighbours



Attention Module (e.g. language)

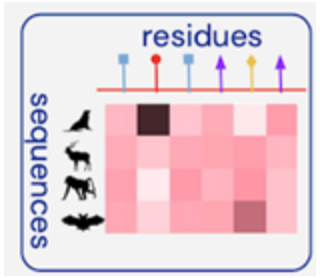
- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)



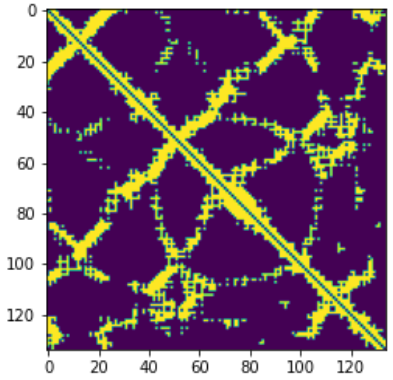

Embedding
Trunk
Heads
© 2020 DeepMind Technologies Limited
MSA
residues
sequences
Genetic search
pairing
templates
sequence-residue edges
residues
sequences
1
Attention
3
Update pairs
4
Update seqs
2
Attention
residue-residue edges
...
Confidence score
High confidence
Low confidence
5
Structure module
3D structure
Pairwise distances
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

Component 1: MSA updates via self-attention

MSA (M, L, K) → Attention over L (**Transformer w/ softly tied attention**) → Attention over M (w/ Performer) → MSA (M, L, K)

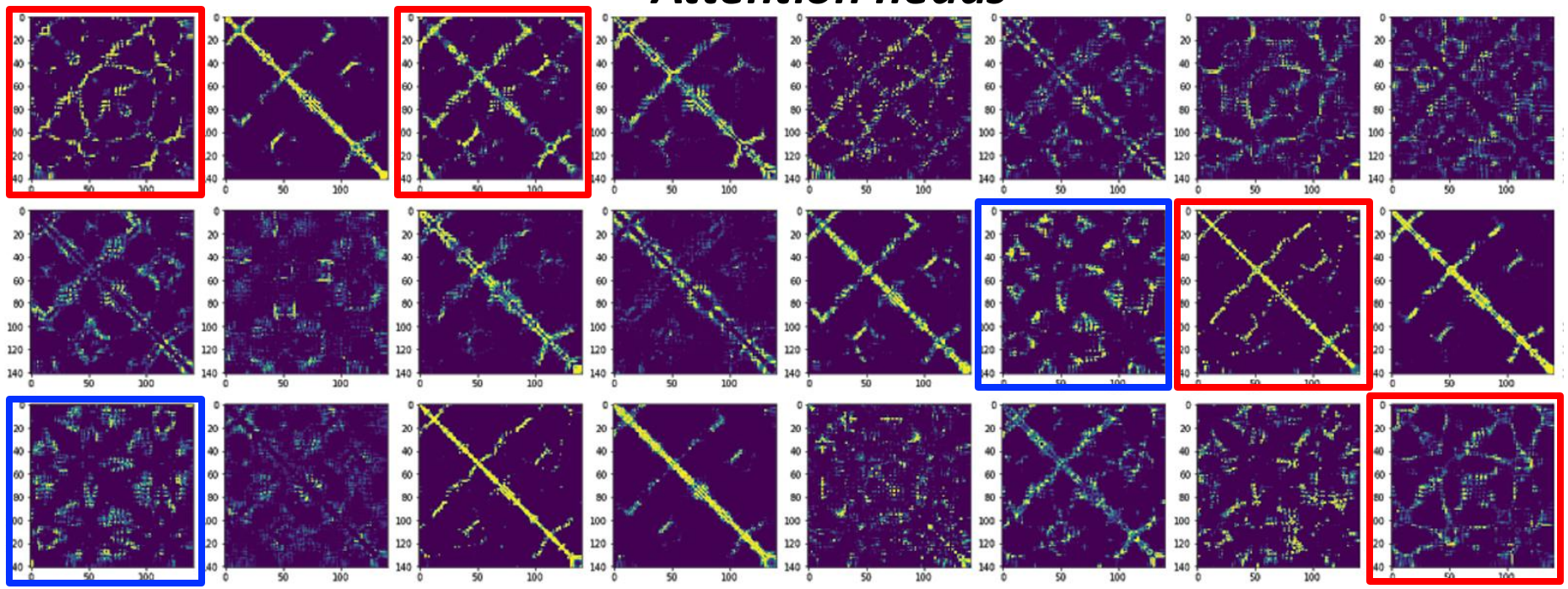


True contact map

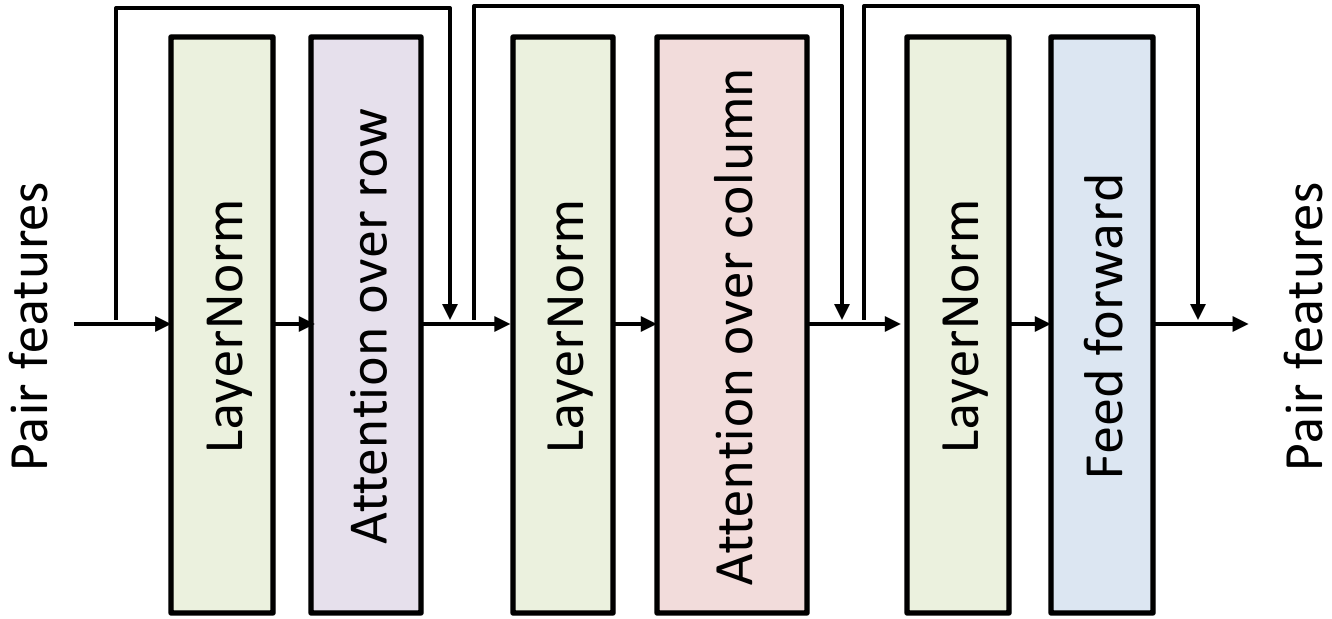


Attention heads

Last 3 blocks

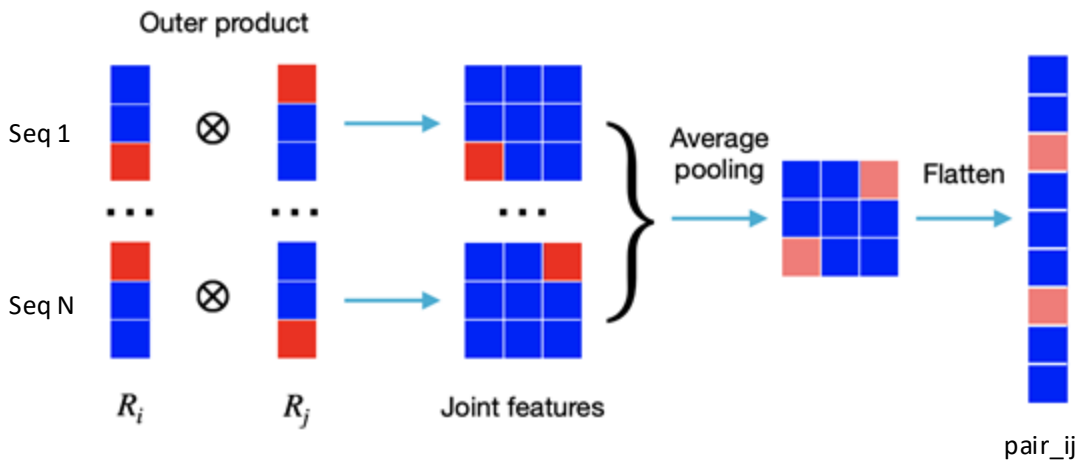


Component 2: Update pair features via self-attention

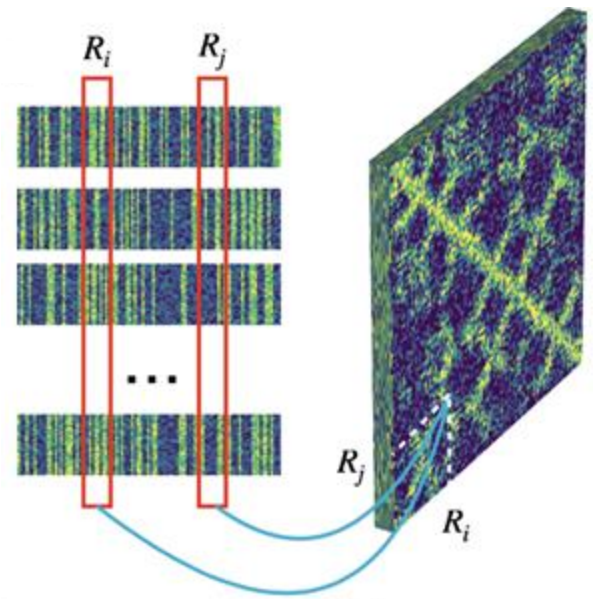


Axial Attention (attention over rows then columns)
to reduce memory requirements & computation time

Component 3: Extract pair features from MSA



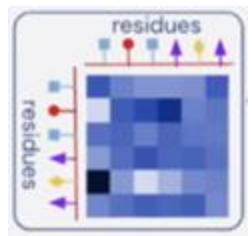
Non-interacting pairs → Broader distribution
Interacting pairs (co-mutating) → Sharper distribution



Outer product & aggregate

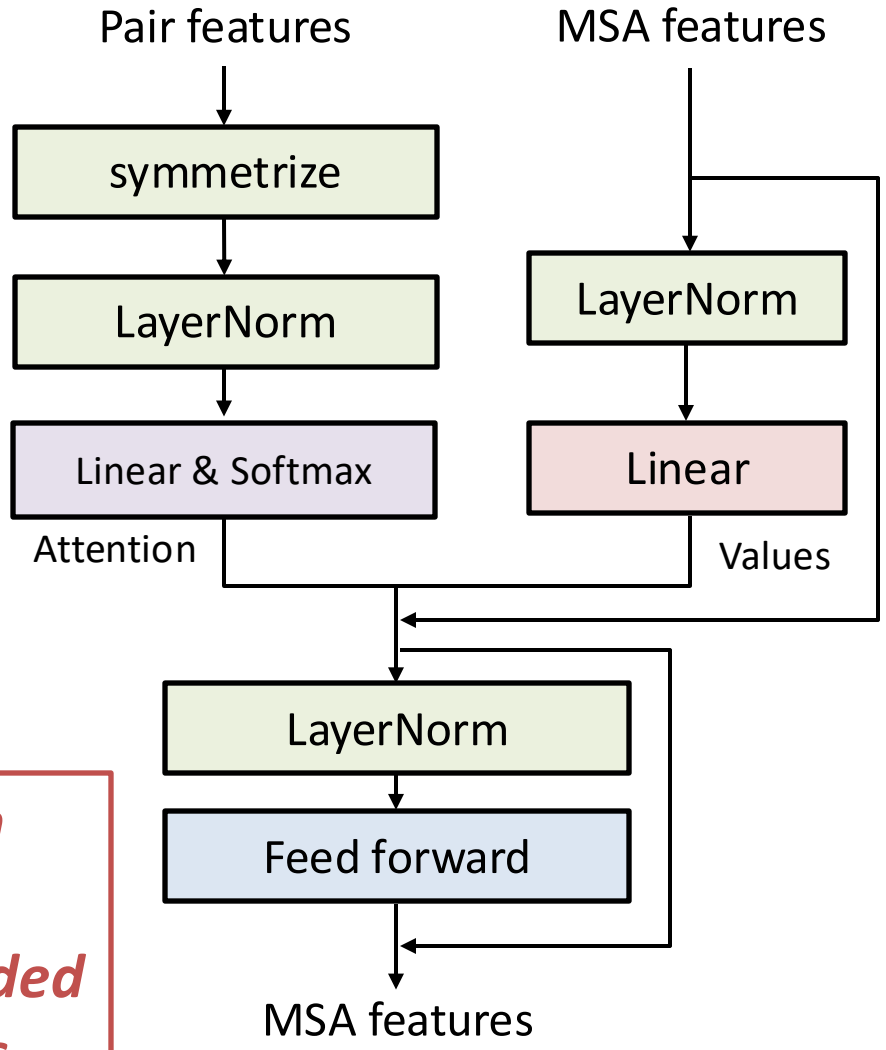
Concat to original pair

& ResNet



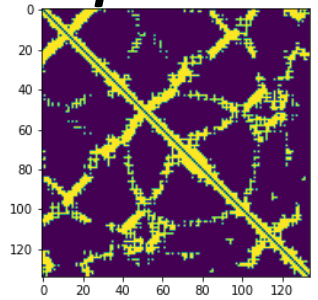
Ju, Fusong, et al. "CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction." *bioRxiv* (2020).

Component 4: Update MSA based on pair features



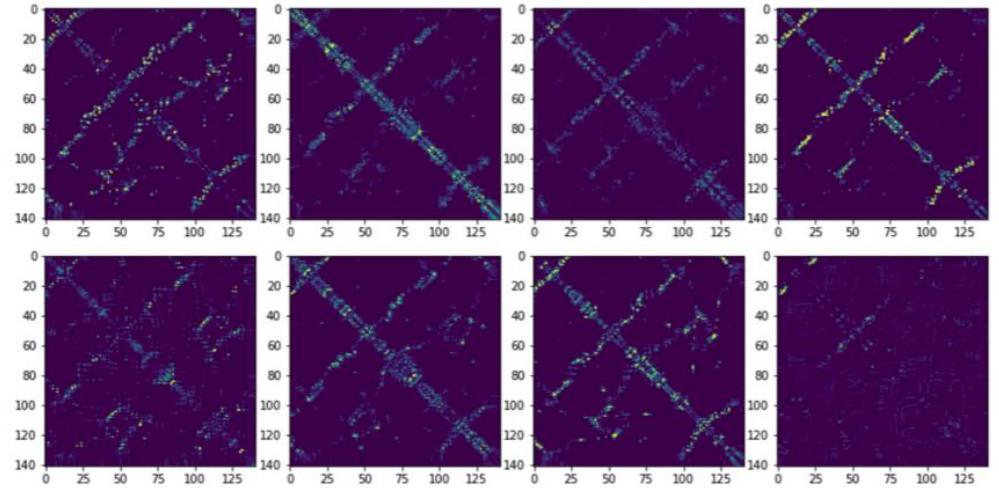
Attention from structure information encoded in pair features

True contact map

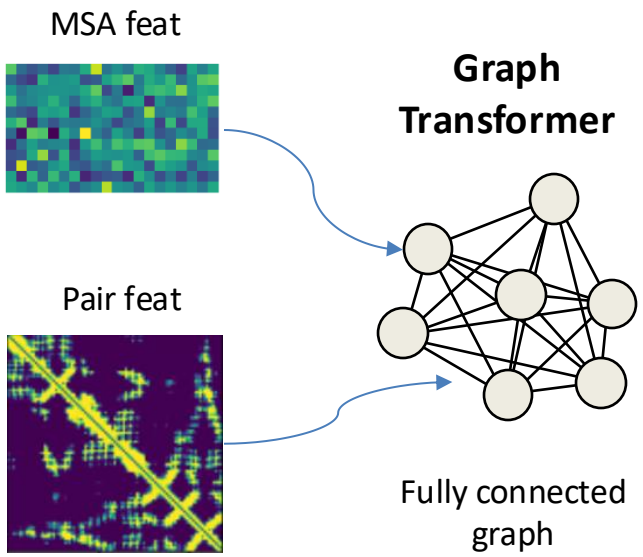


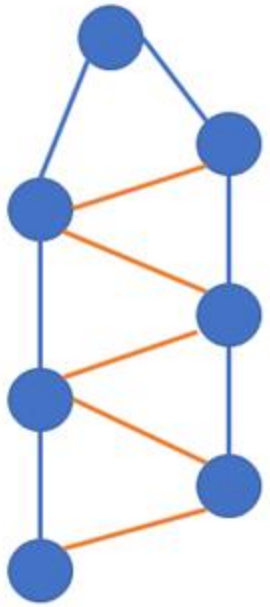
Attention heads

Last 2 blocks

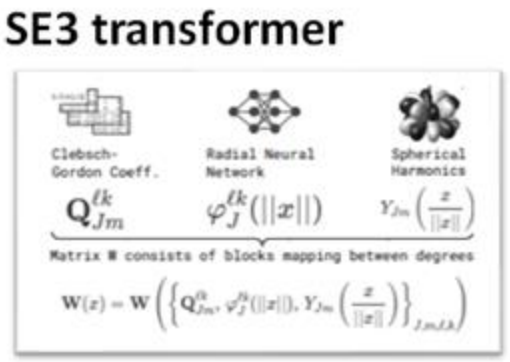


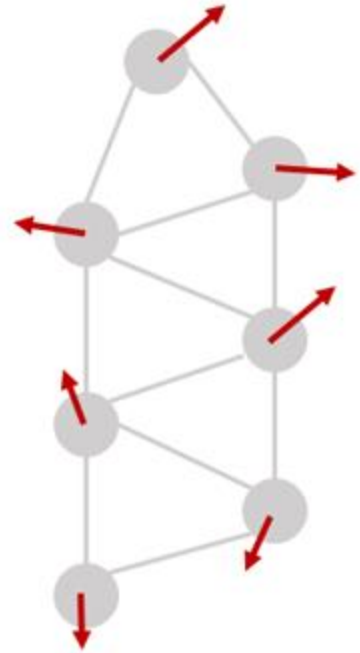
Component 5: SE(3)-Transformer for structure refinement





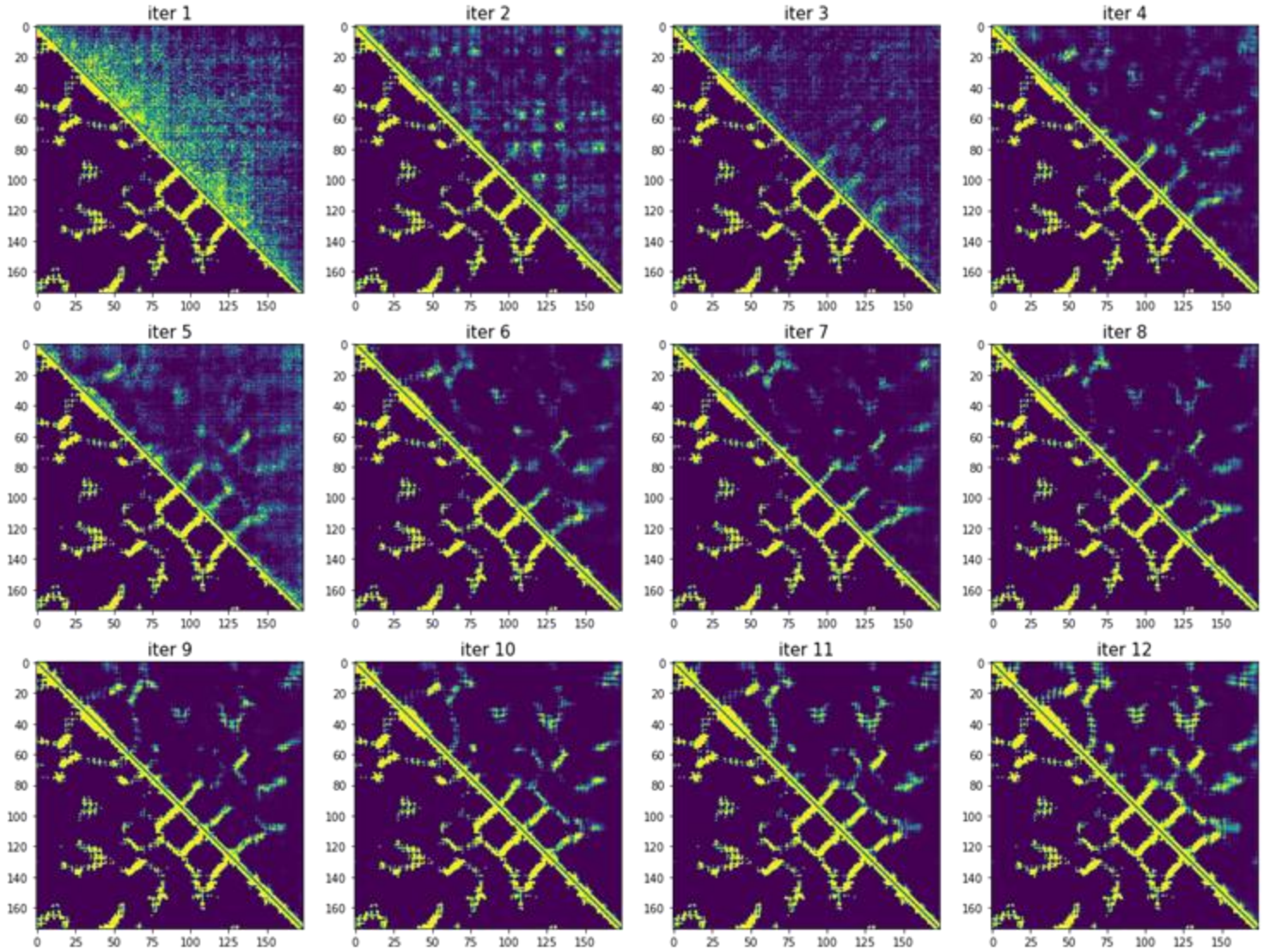
- Initial backbone Cartesian coordinates
- Node + edge information from trunk
- Graph connecting nearby residues



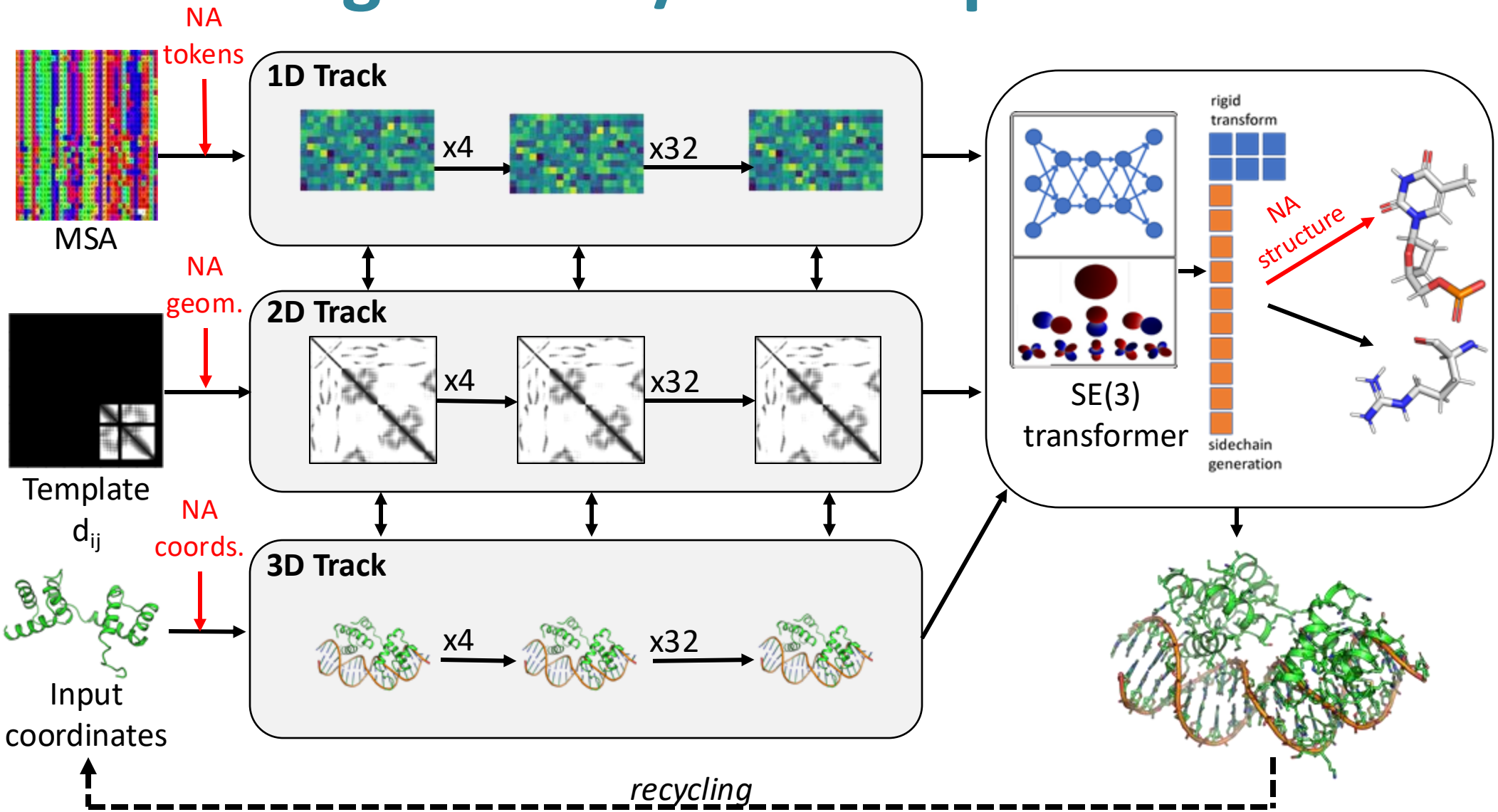


- Predicted offset vectors
- Predicted per-residue errors

What happens during iteration?

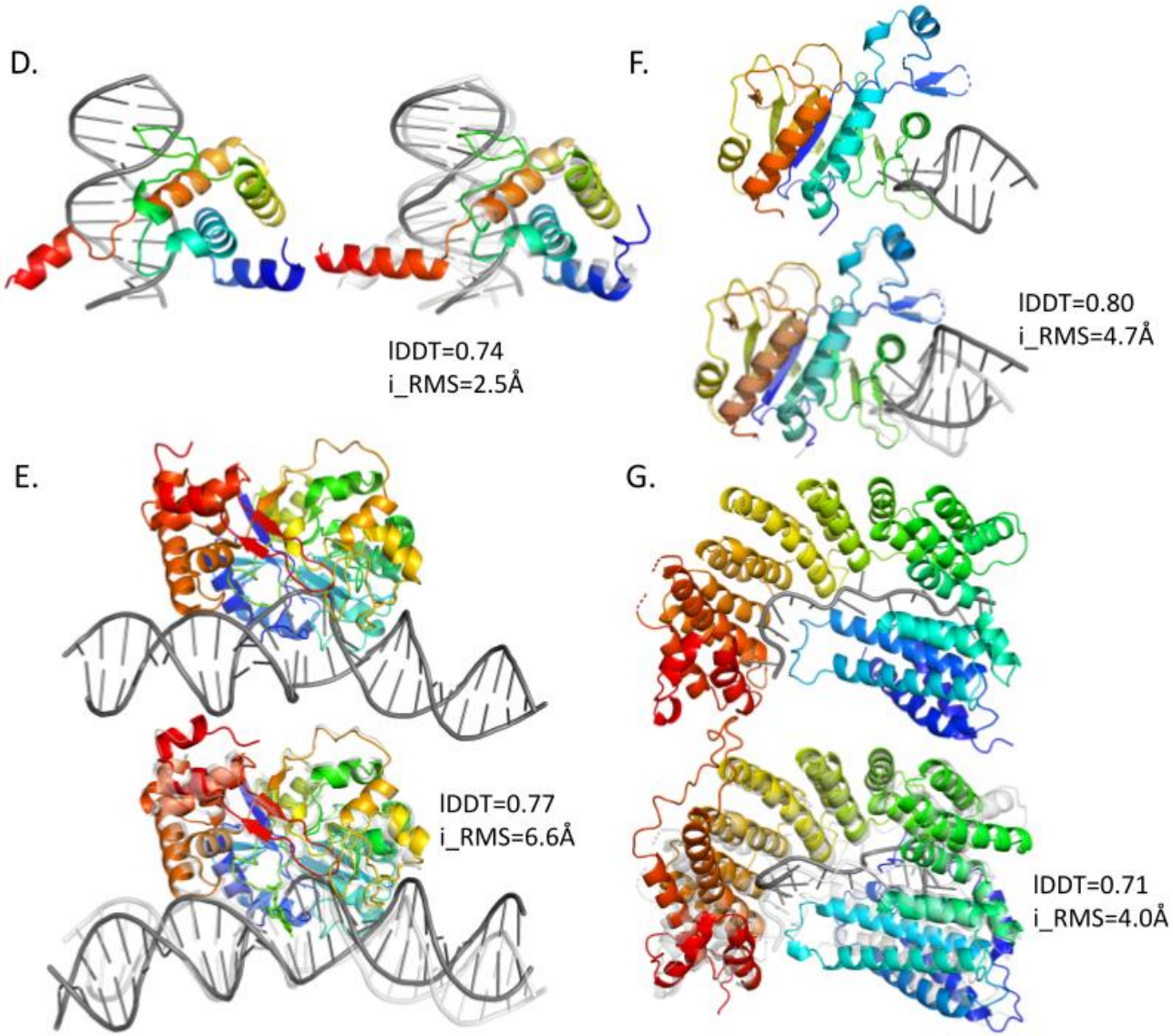


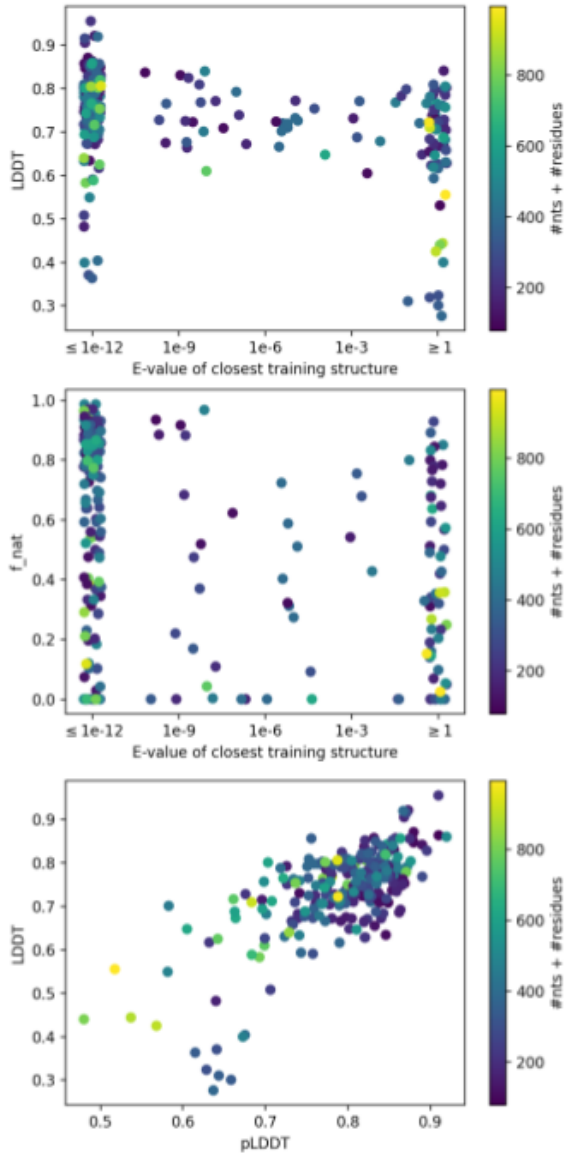
Predicting Protein/NA complexes with RF2



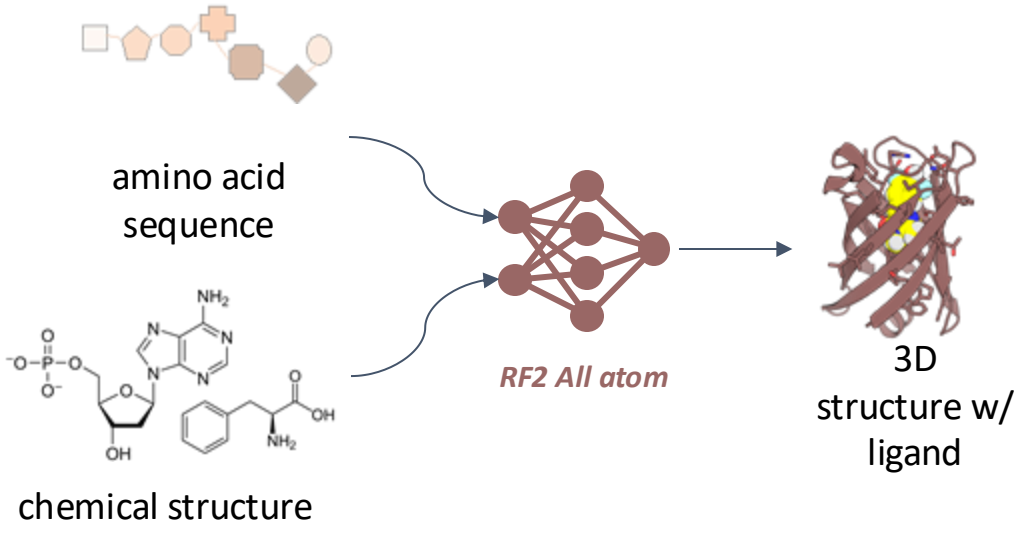
Baek, McHugh, *et al.*
Nature Methods (in press)

Predicting Protein/NA complexes with RF2





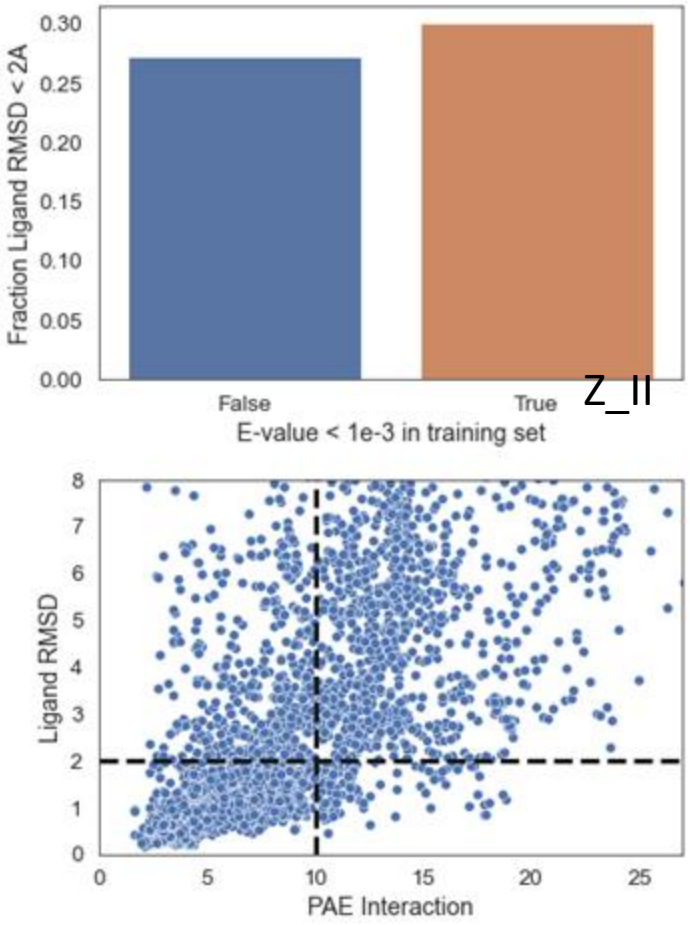
RoseTTAFold Allatom



- Data represented as residues (protein/NA) and/or atoms (small molecules/PTMs)
- Models proteins, DNA/RNA, small molecule ligands, unnatural amino acids, glycosylations, other PTMs

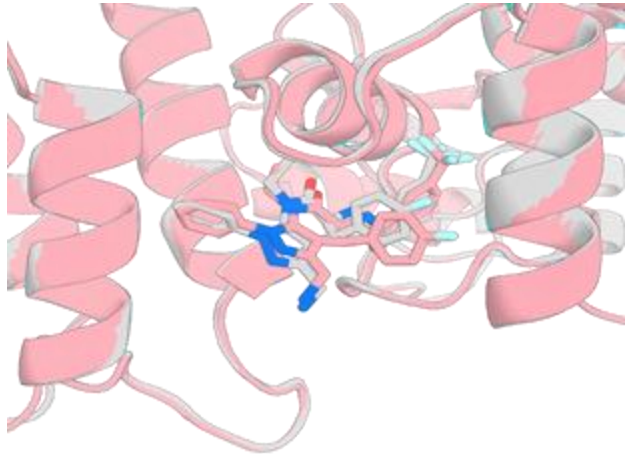
Rohith Krishna (Baker lab)

RF-allatom predicts protein small molecule complexes



Z_II

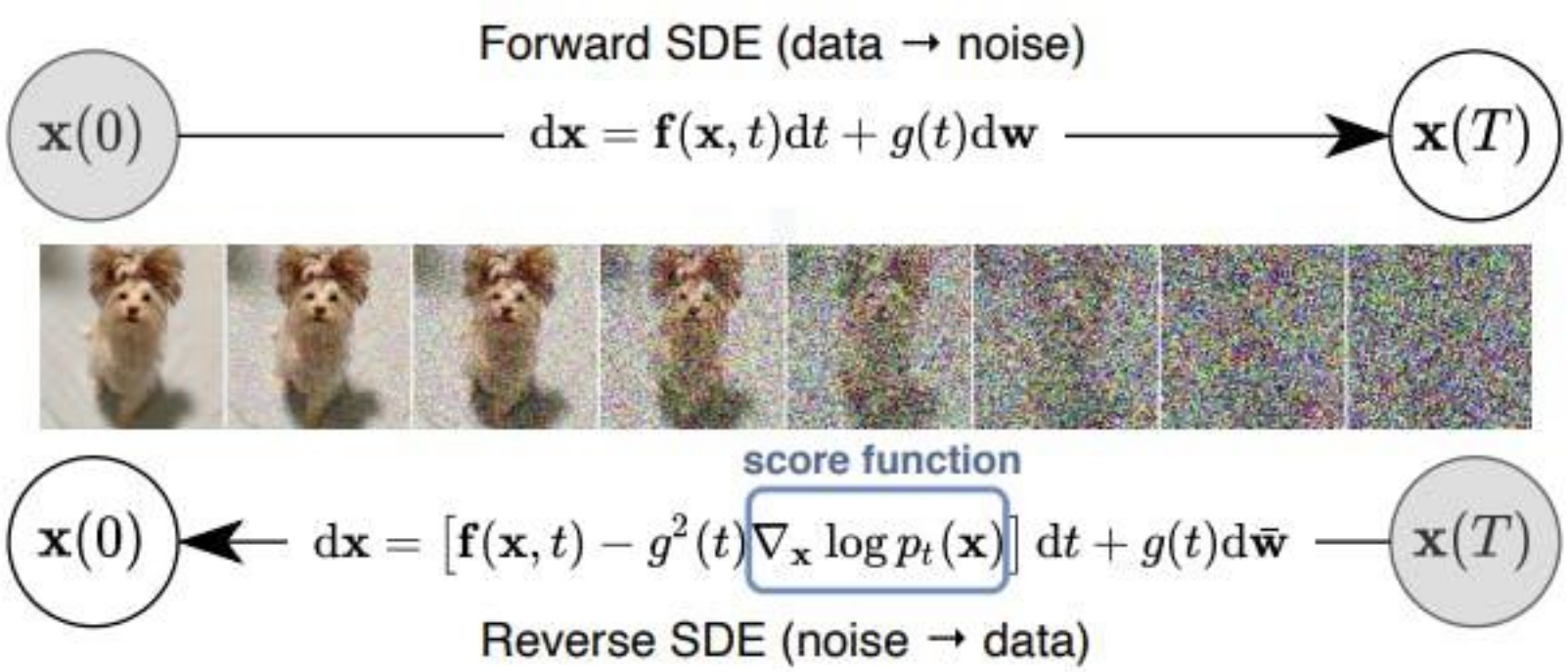
Prediction Crystal



Ligand RMSD: 0.95
Max tanimoto to train set: 0.53

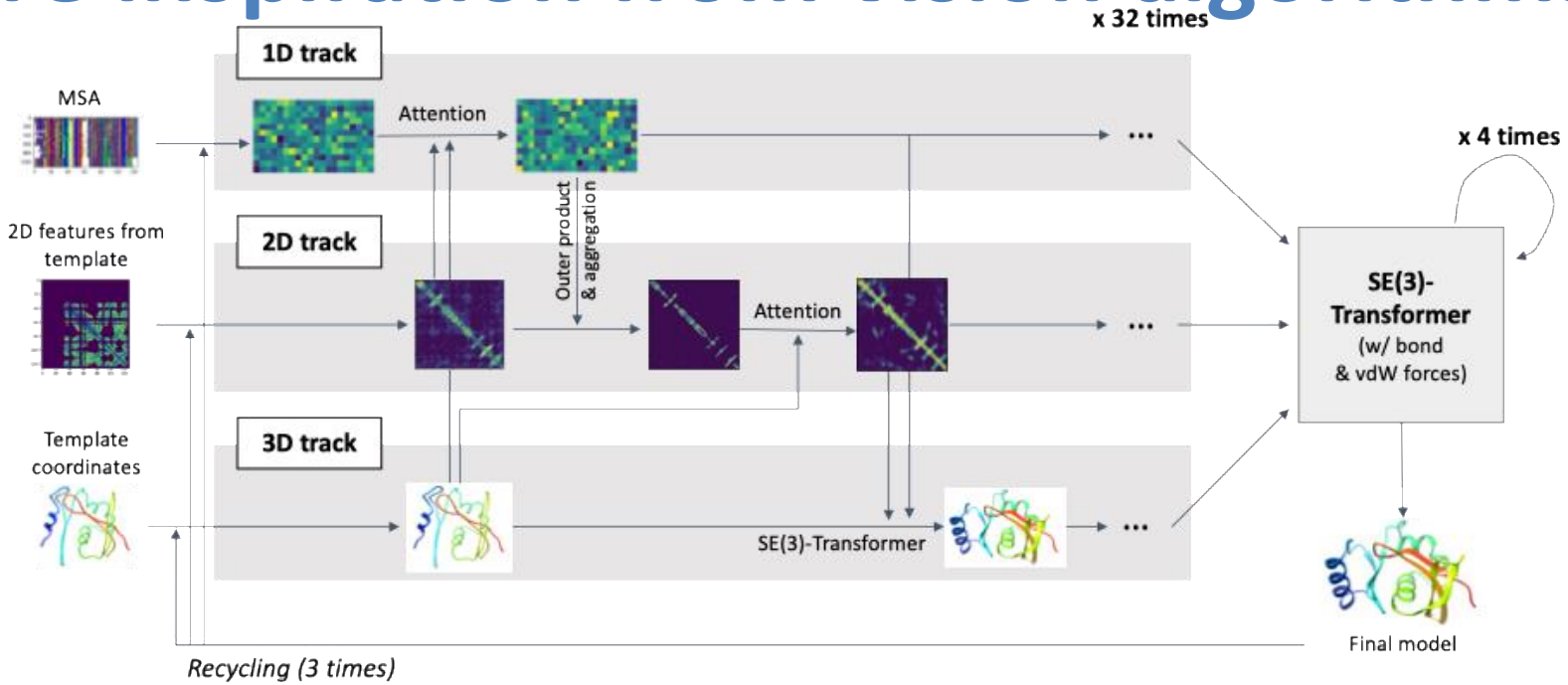
Ligand RMSD: 0.86
Max tanimoto to train set: 0.54

More inspiration from vision algorithms



More inspiration from vision algorithms

RF2:



AF3: