Structural Bioinformatics

GENOME 541

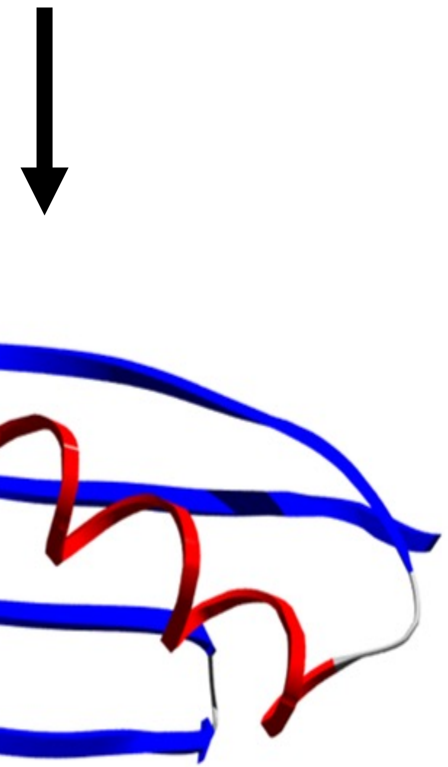
Spring 2023

Lecture 3: Protein Structure Prediction

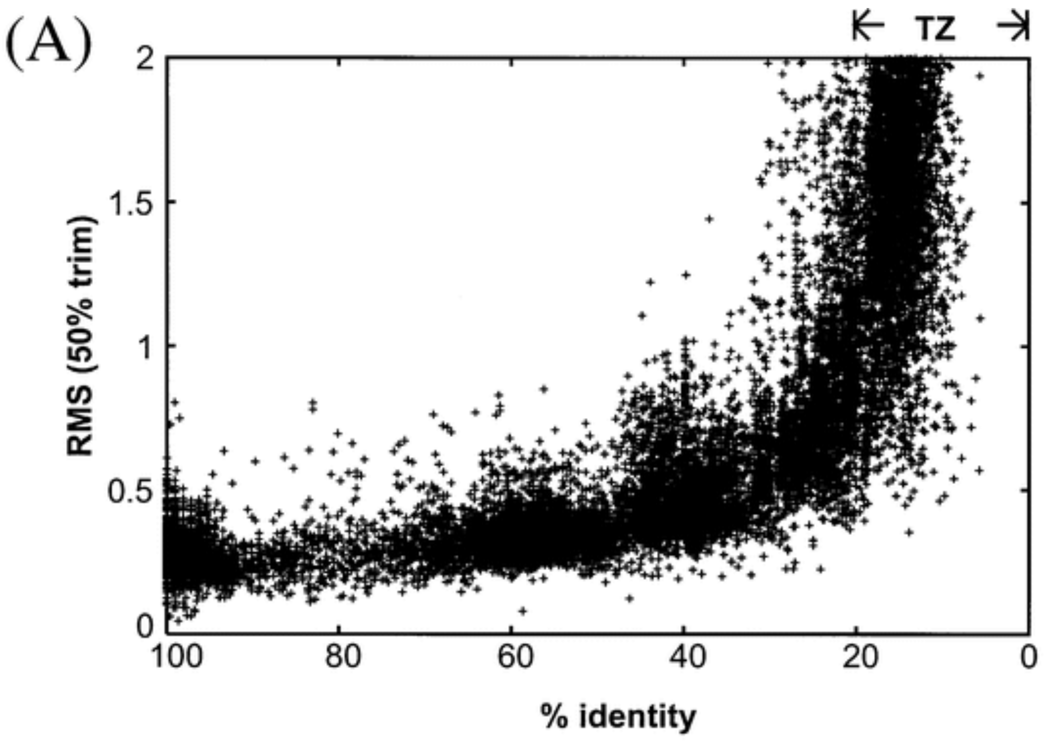
Frank DiMaio (dimaio@uw.edu)

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.

BLAST (Basic Local Alignment Search Tool)

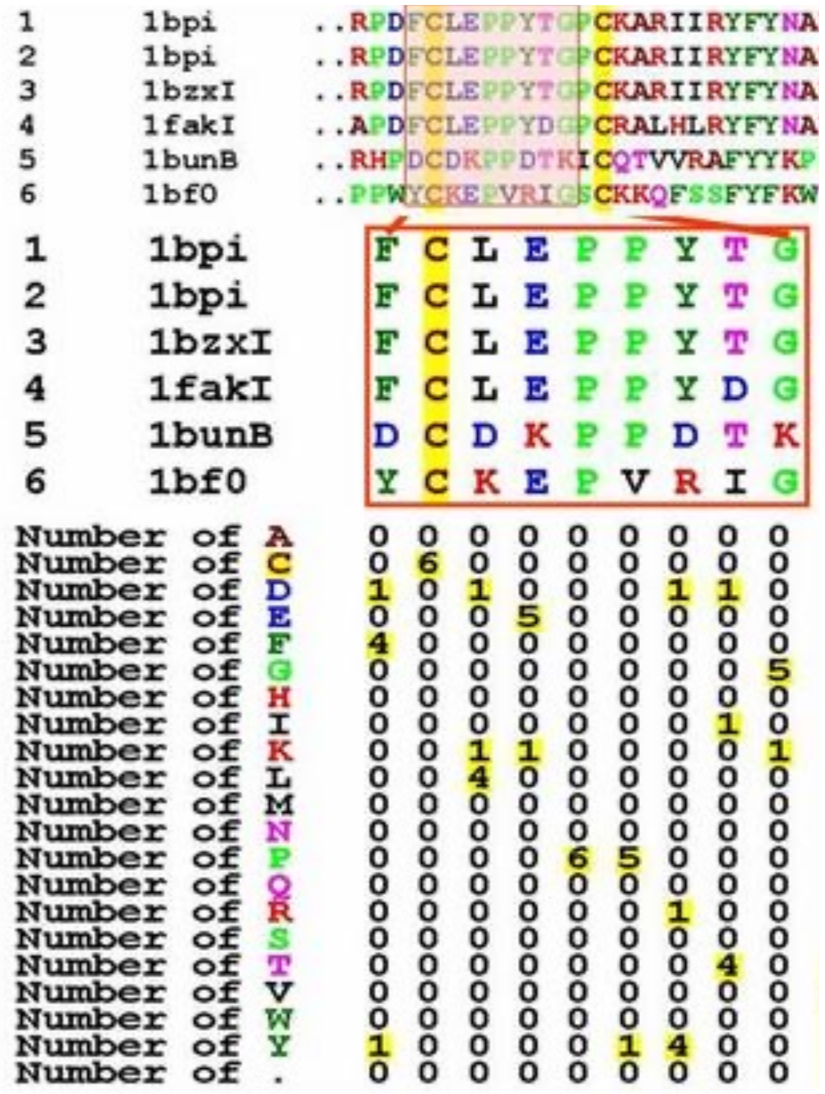
BLAST is a fast sequence alignment algorithm that identifies high-scoring local alignments by finding short exact matches (seeds) and extending outward. BLAST uses the BLOSUM62 aa substitution matrix by default.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

PSI-BLAST

- Position-Specific Iterated BLAST
- Allows more distantly related sequences to be identified
- Steps
  1. Use BLAST to identify related sequences
  2. Create a profile from related sequences
  3. Search for related sequences using this profile

Sequence Profile



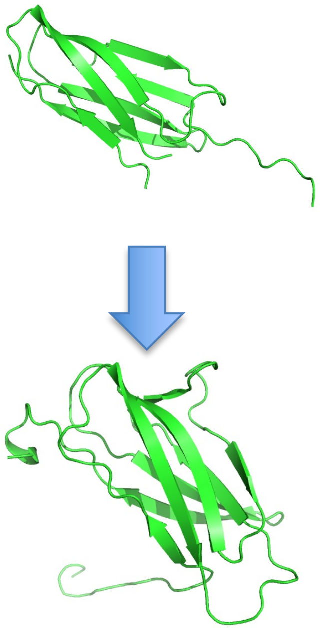
- For each column in a MSA count how often each amino acid occurs
- Combine with prior information about substitution frequencies (ie. BLOSUM62)
- Convert counts to log odds scores. End product is a Position-Specific Scoring Matrix (PSSM)

Homology Modeling

- Identify homologous protein sequences

MNDD--VDIQ---QSYP-FSI…
LTDSQLAQVAAFVNNYPNVEL…

- Build model by
  1. “Threading” residues in corresponding positions of homologous structure
  2. Sampling conformations of unaligned residues
  3. All-atom refinement



De novo protein structure prediction

MQIFVKTLTGKTIT
LEVEPSDTIENVKA
KIQDKEGIPPDQQR
LIFAGKQLEDGRTL
SDYNIQKESTLHLV
LRLRGG

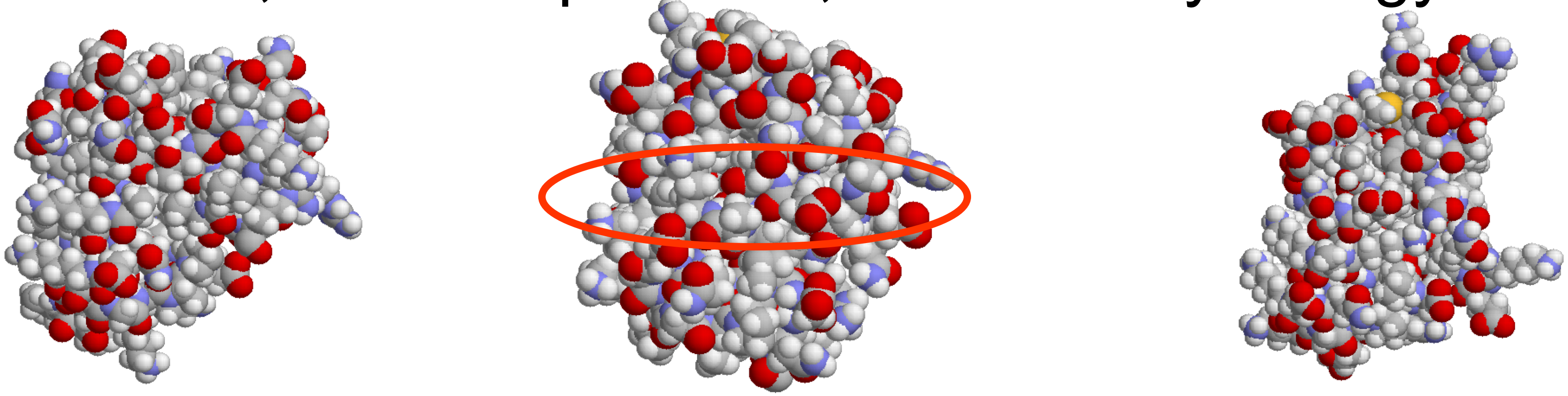
→

Thermodynamic hypothesis:
The native state is the lowest-energy conformation.

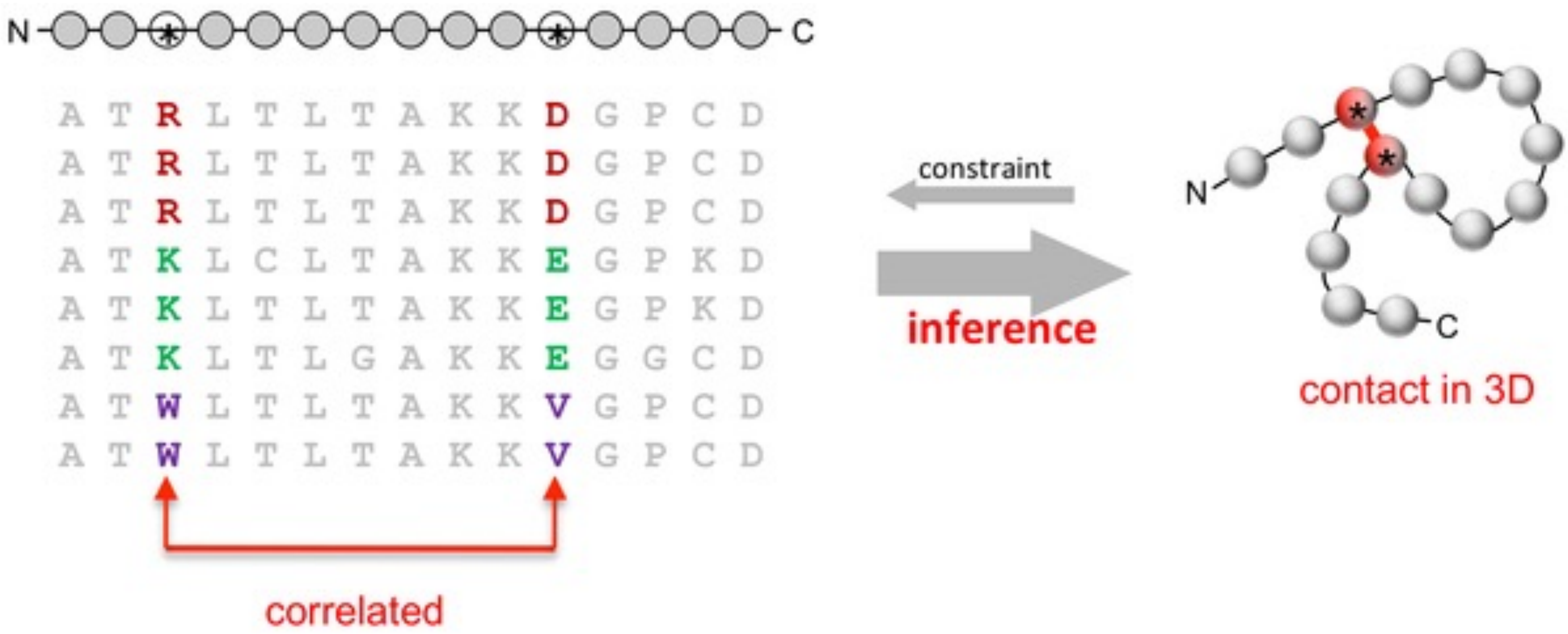
Structure Prediction Protocol

- Large-scale search of conformational space using a low-resolution potential

- Refinement of candidate models in a physically realistic, all-atom potential; selection by energy



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,

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

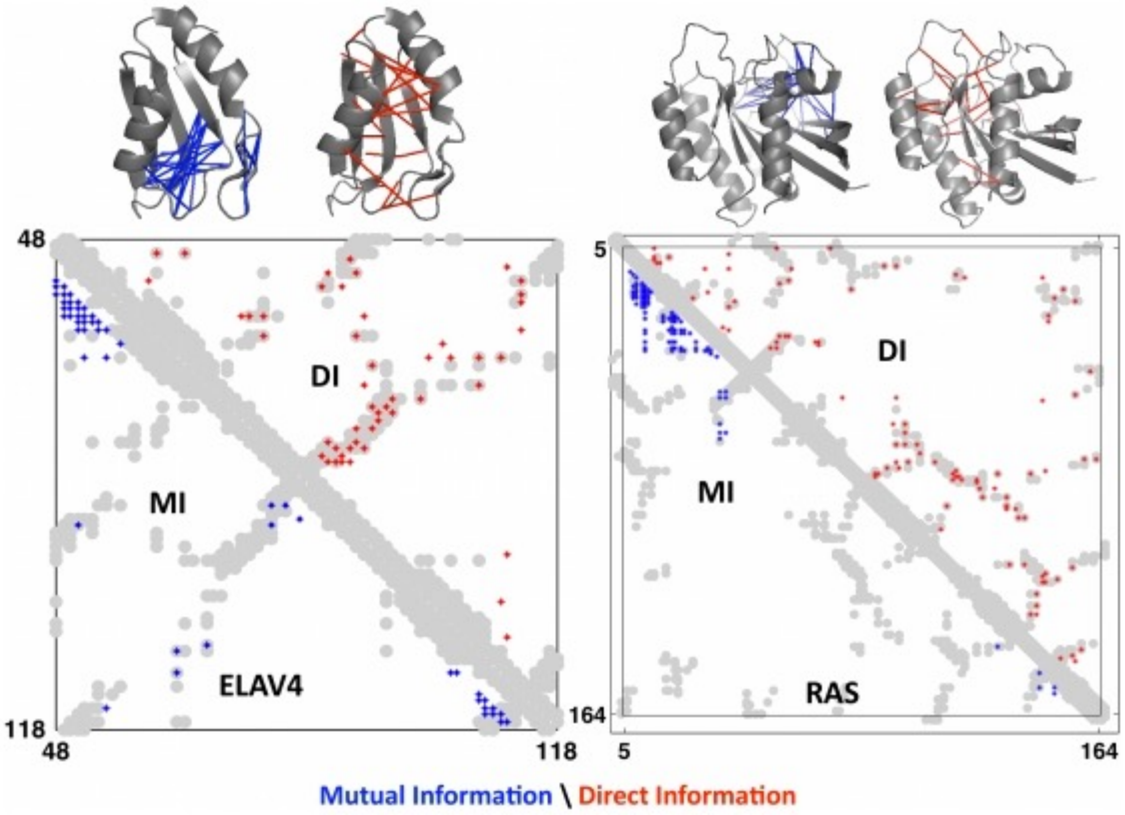
and to use the interaction parameters $\mathbf{J}_{ij}$ as predictions of spatial proximity among amino-acid pairs in the protein structure.

Problem: **Z** cannot be tractably computed

Solutions:

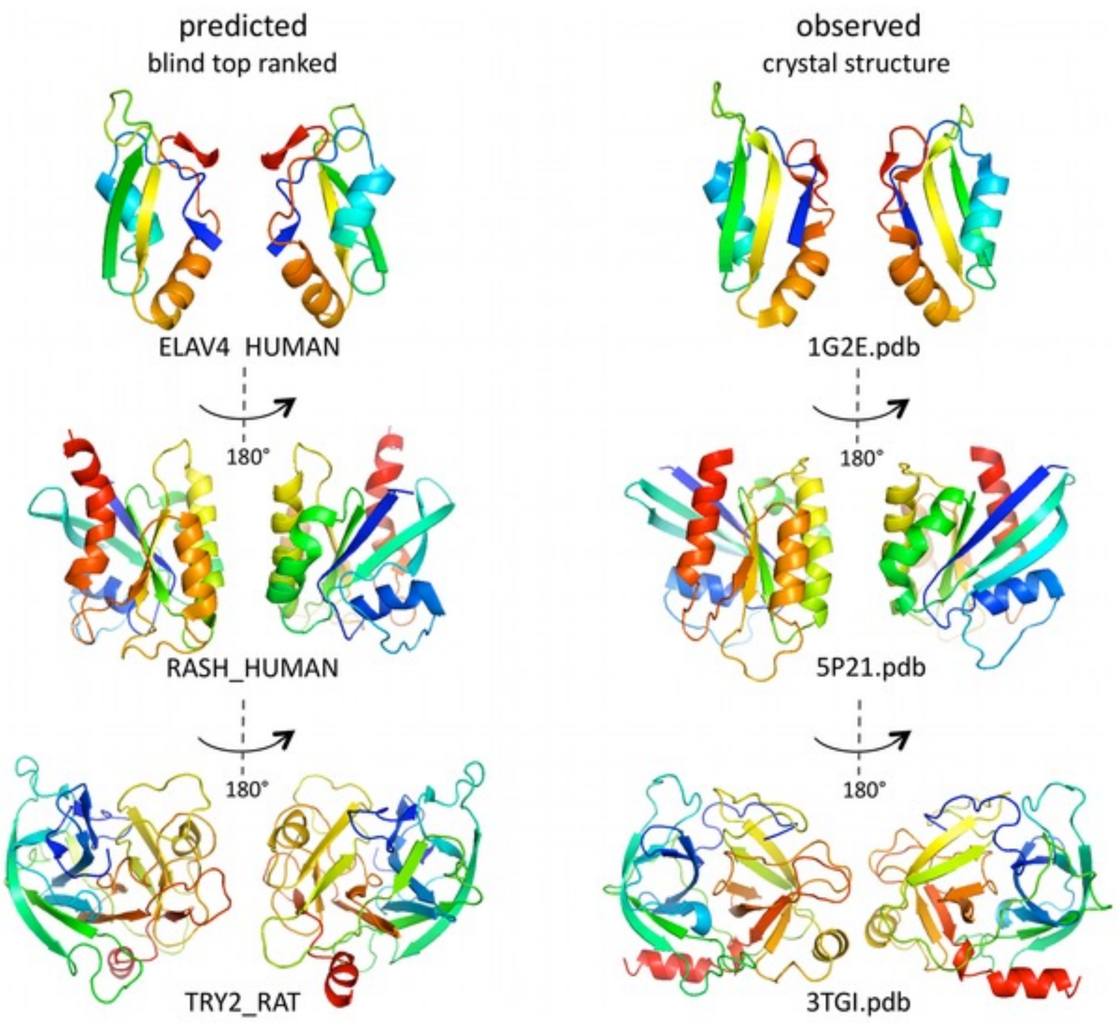
- Mean-field approach (mfDCA) (https://www.pnas.org/content/108/49/E1293)
- Pseudo-likelihood (plmDCA) (https://journals.aps.org/pre/abstract/10.1103/PhysRevE.87.012707)

Correlated mutations carry information about distance relationships in protein structure.



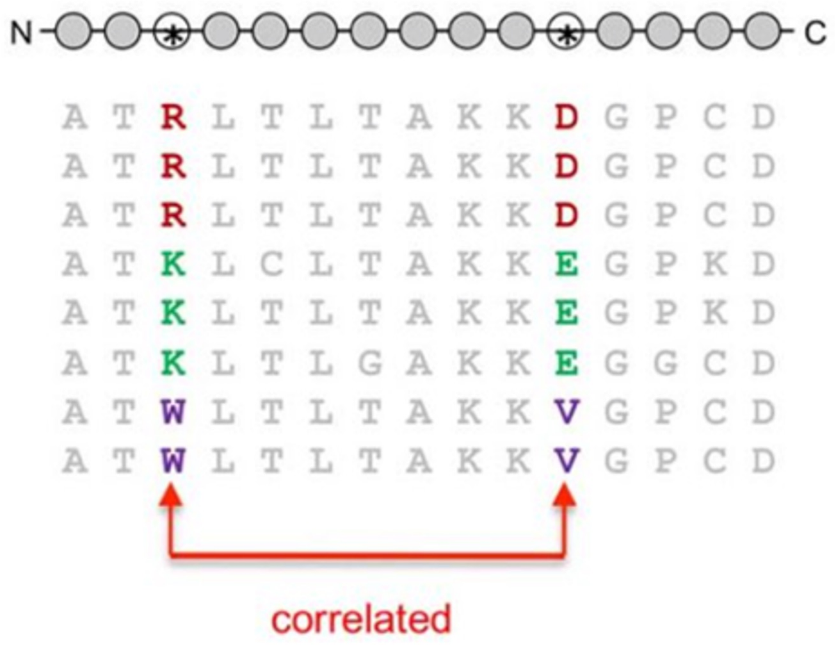
$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\exp\left(\sum_{i=1}^{L}\left[\mathbf{v}_i(x_i)+\sum_{j>i}^{L}\mathbf{w}_{i,j}(x_i,x_j)\right]\right).$$

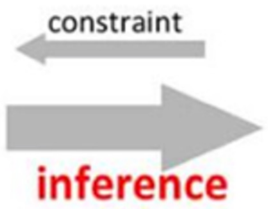
Predicted 3D structures for three representative proteins.



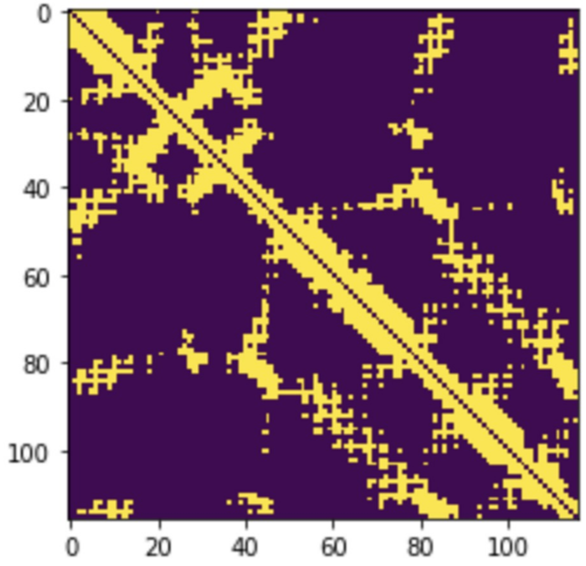
Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766. https://doi.org/10.1371/journal.pone.0028766
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

Coevolution guided modeling



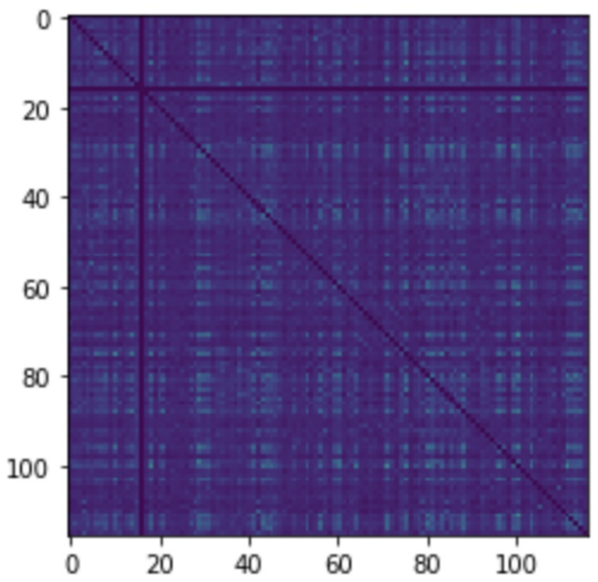


Native contact map



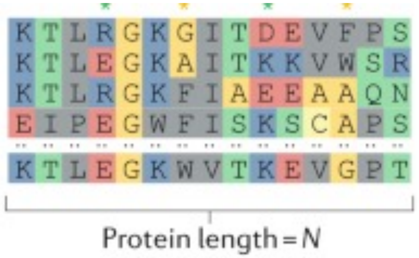
GREMLIN predictions on shallow MSAs

(Nseq=36, Nf=2.3)

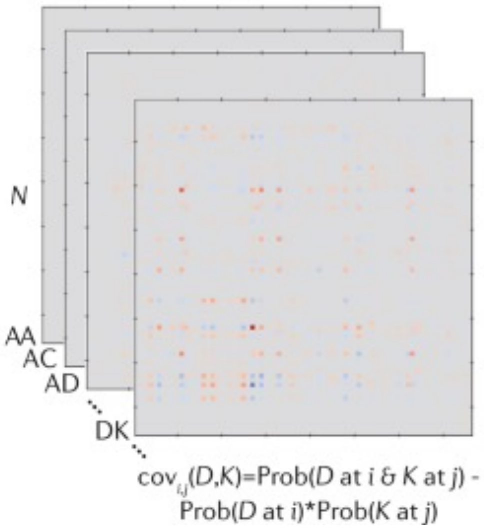


Contact maps = Computer Images?

MSA

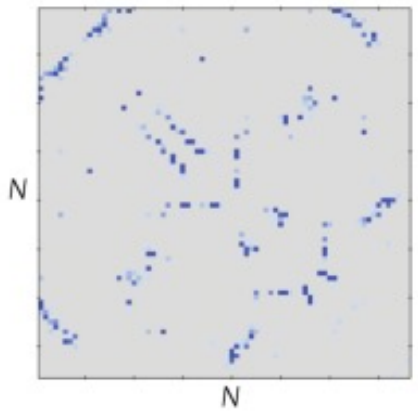


Covariances

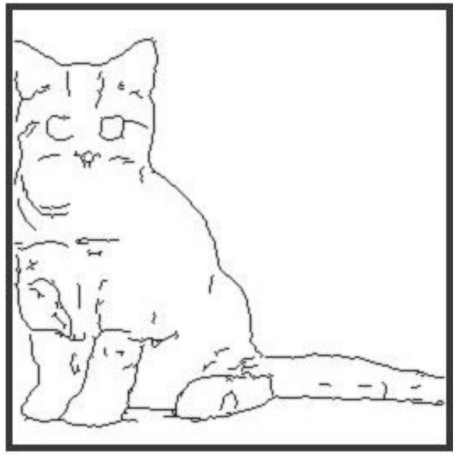


Contact prediction

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

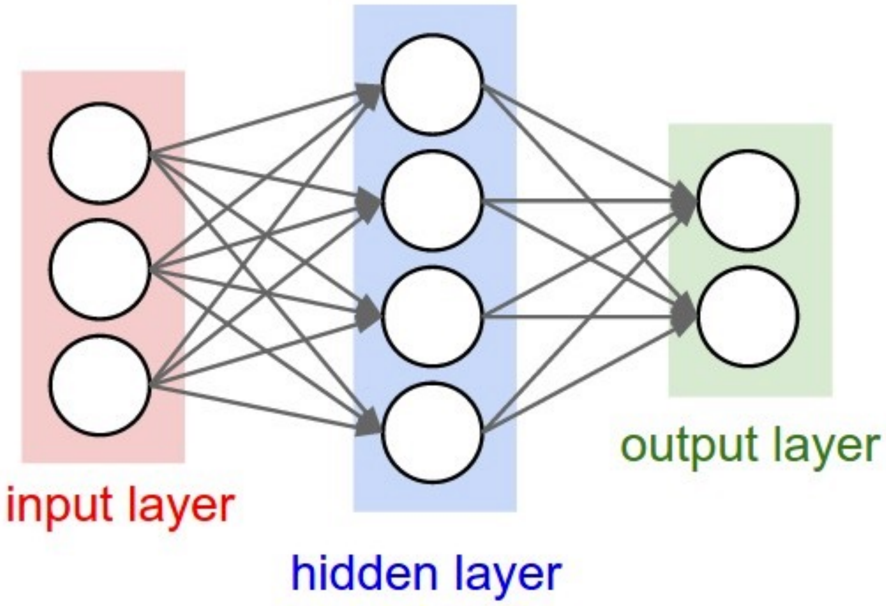
☯ These authors contributed equally to this work.
* jinboxu@gmail.com

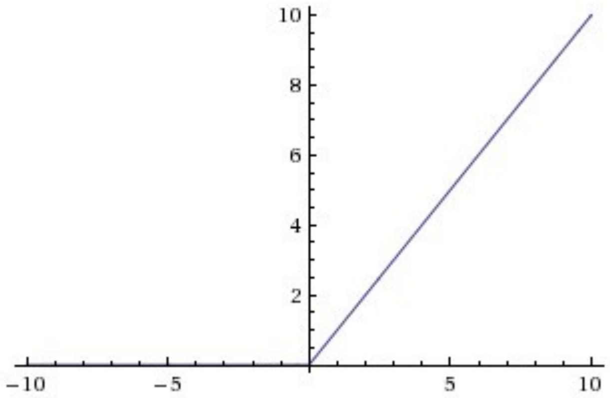
Abstract

Motivation

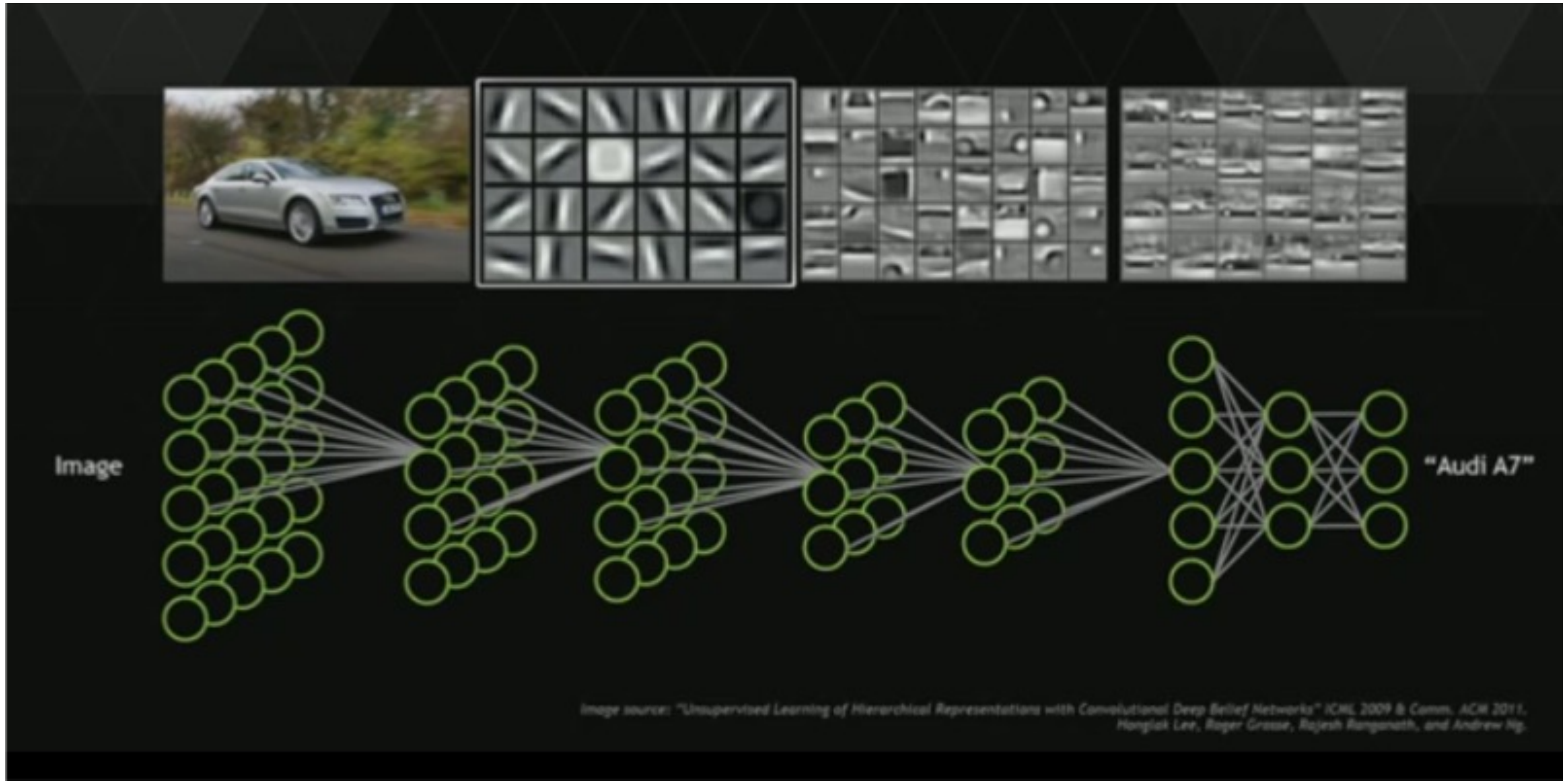
Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still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

Neural networks

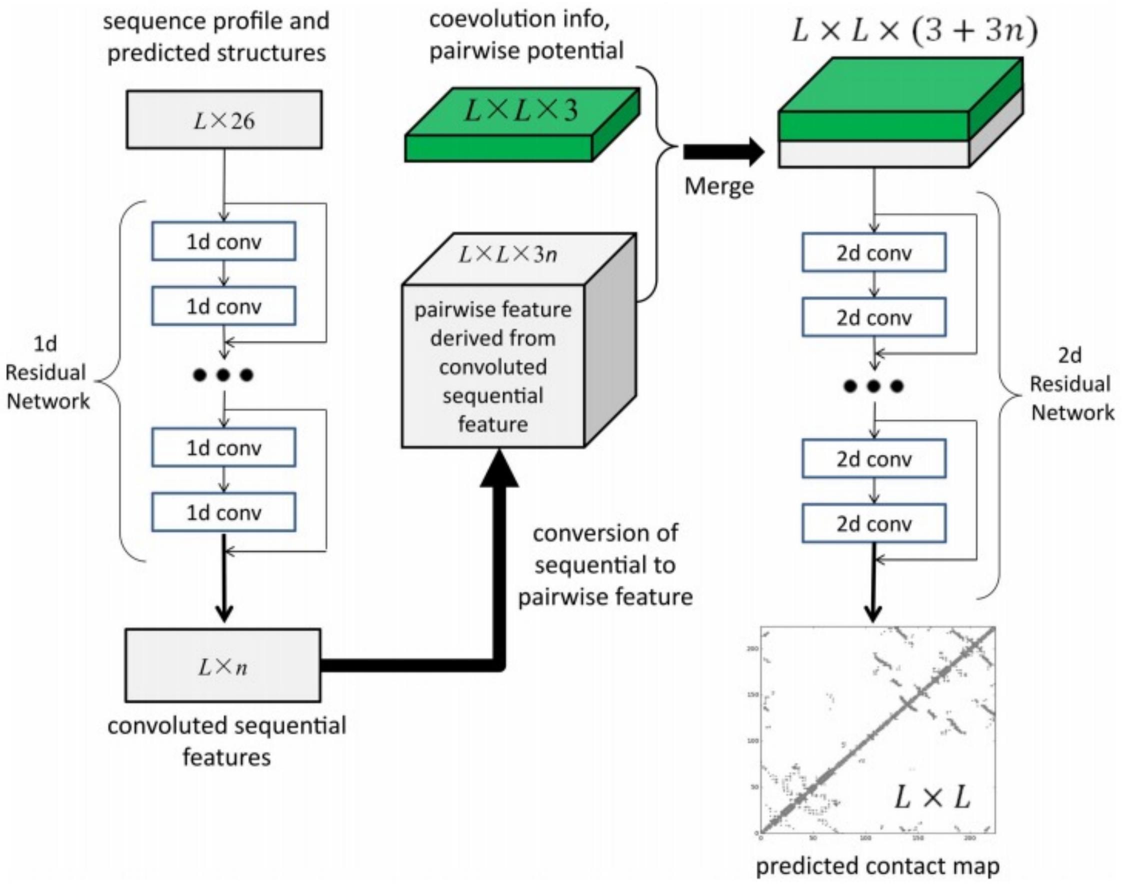




Convolutional neural networks



Learning a contact map from co-evolving residues



Inferring better contact maps (I)

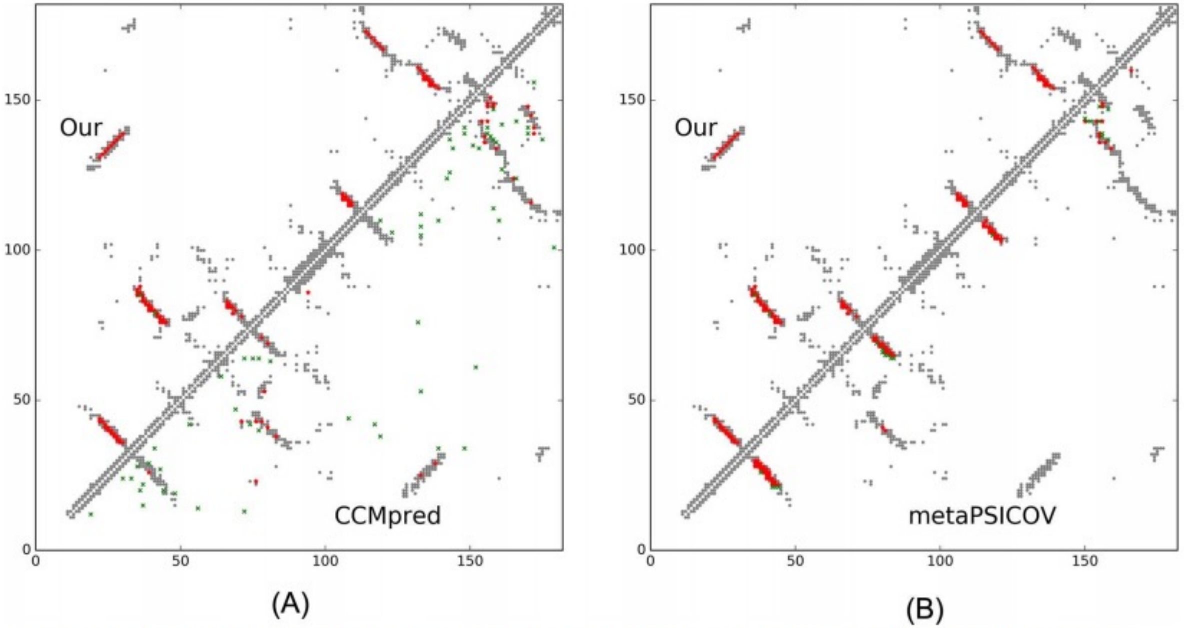
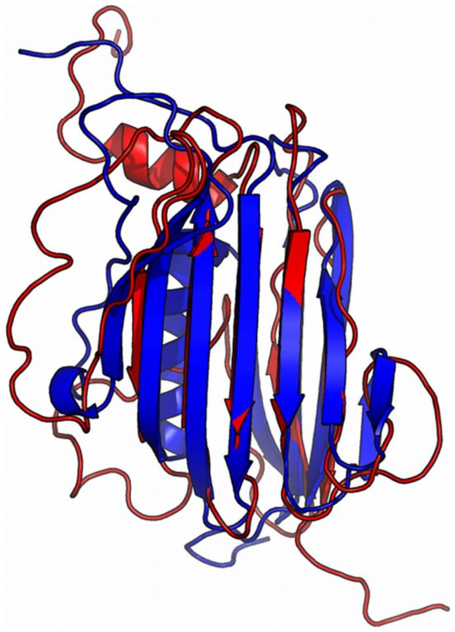


Fig 6. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 2nc8A. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The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



Inferring better contact maps (II)

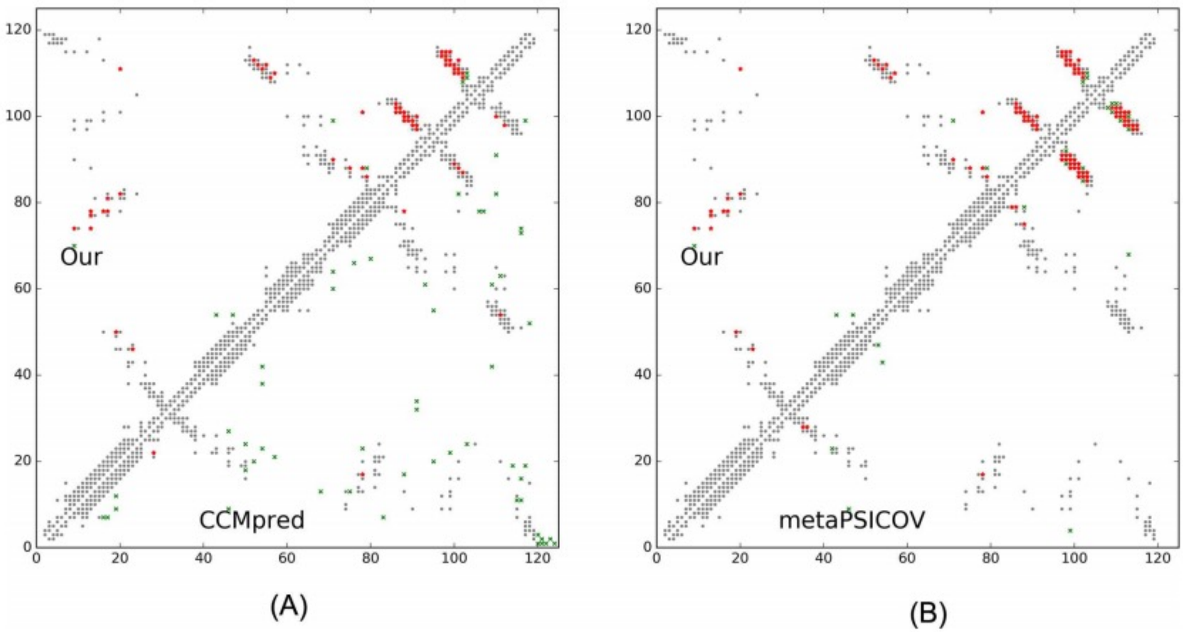
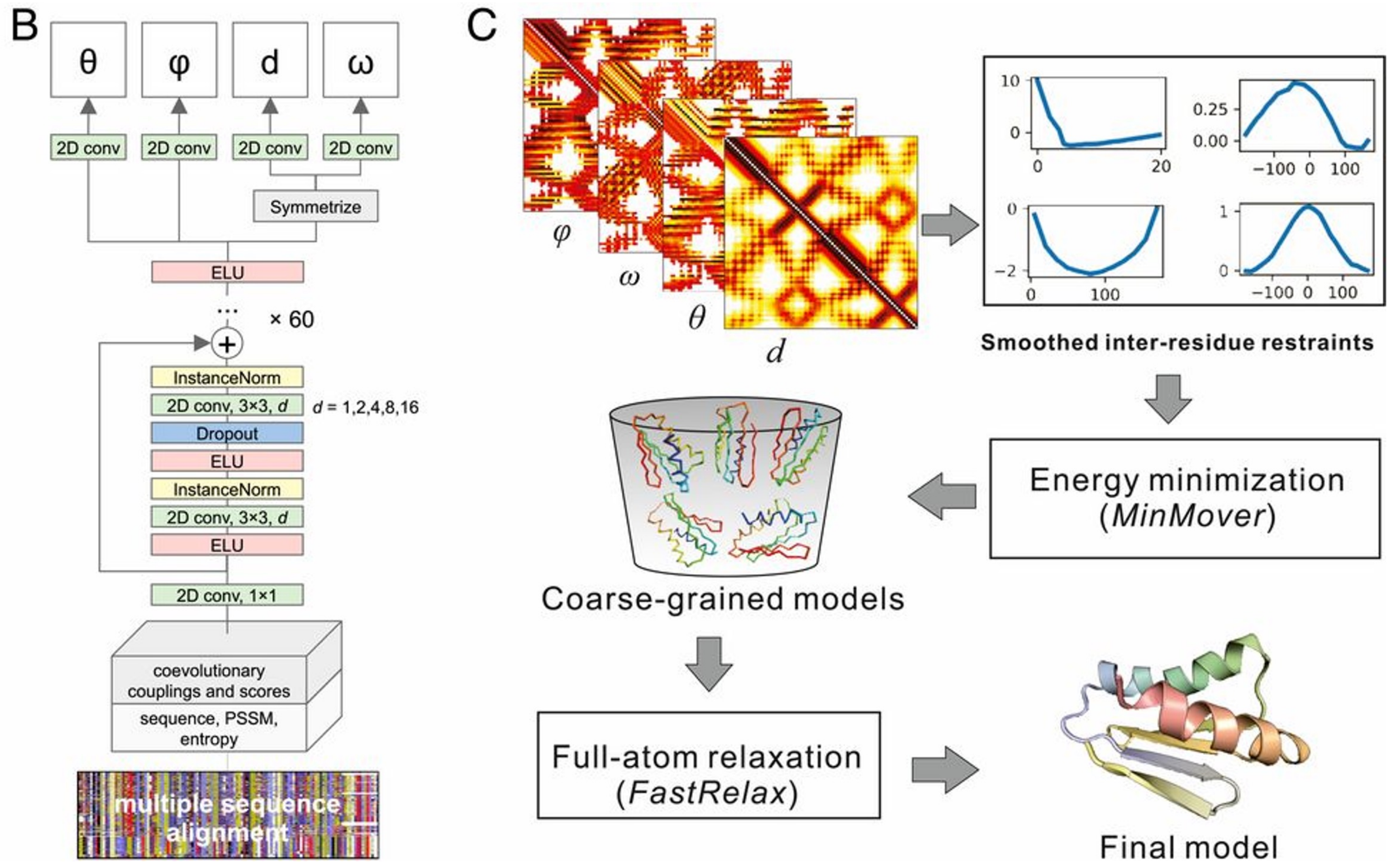


Fig 9. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 5dcjA. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The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

trRosetta



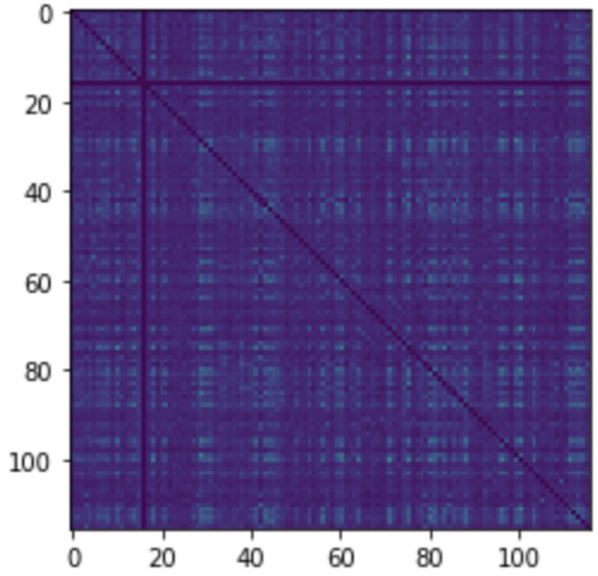
Improved protein structure prediction using predicted interresidue orientations

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

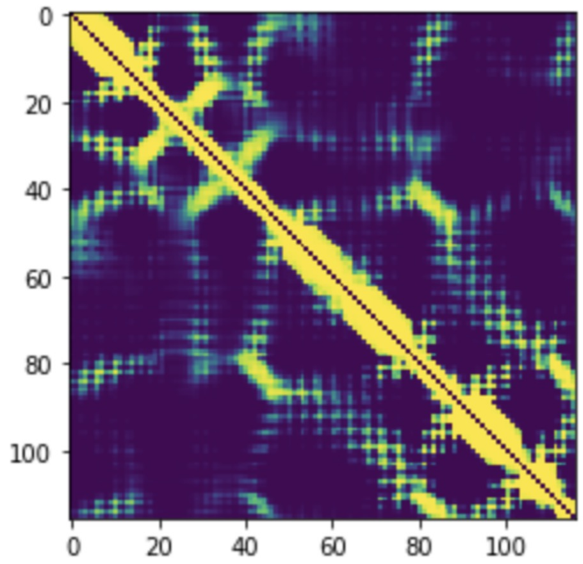
PNAS January 21, 2020 117 (3) 1496-1503; first published January 2, 2020 https://doi-org.offcampus.lib.washington.edu/10.1073/pnas.1914677117

Discovering hidden patterns with a learned model

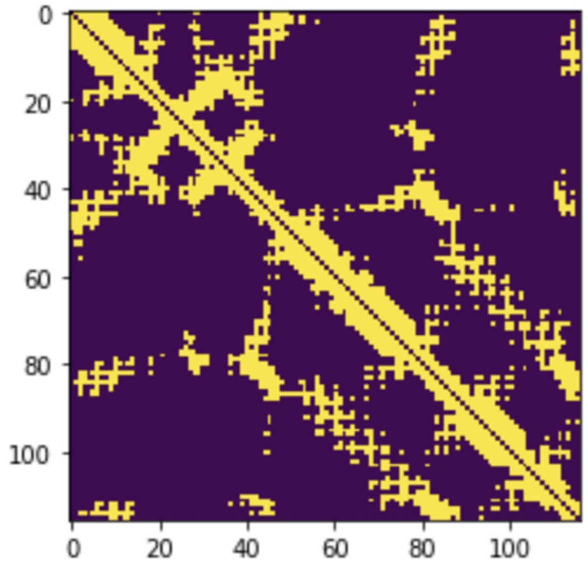
Gremlin predictions on shallow MSAs
(Nseq=36, Nf=2.3)



trRosetta predictions on shallow MSAs
(Nseq=36, Nf=2.3)



Native contact map



Improving protein structure prediction

Free modeling accuracy in CASP

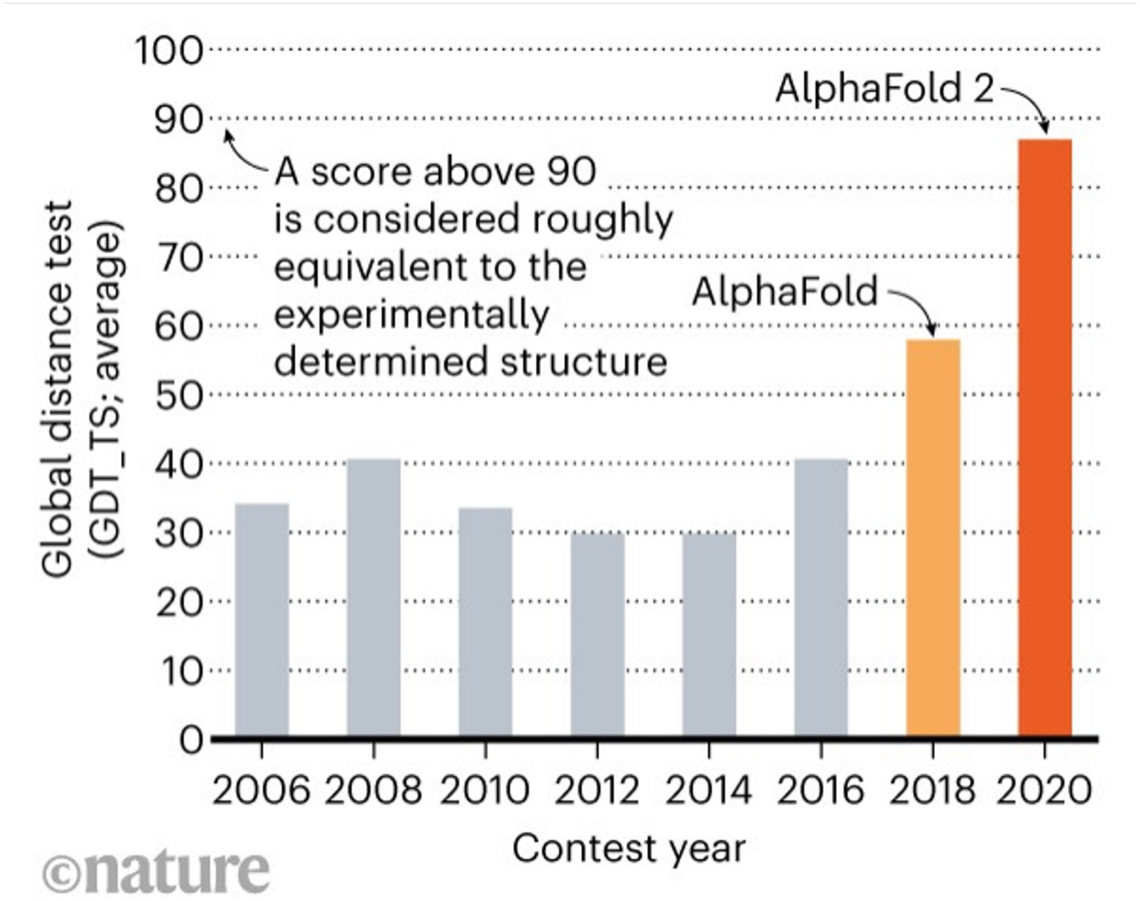
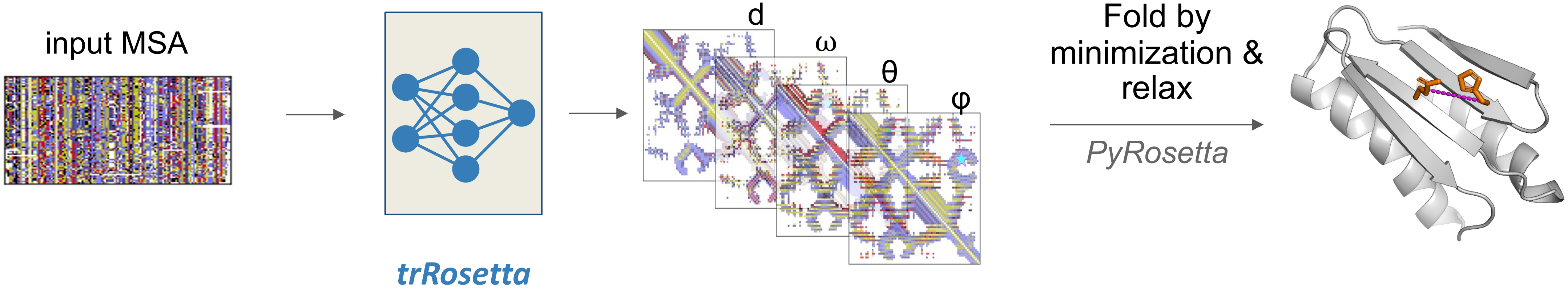


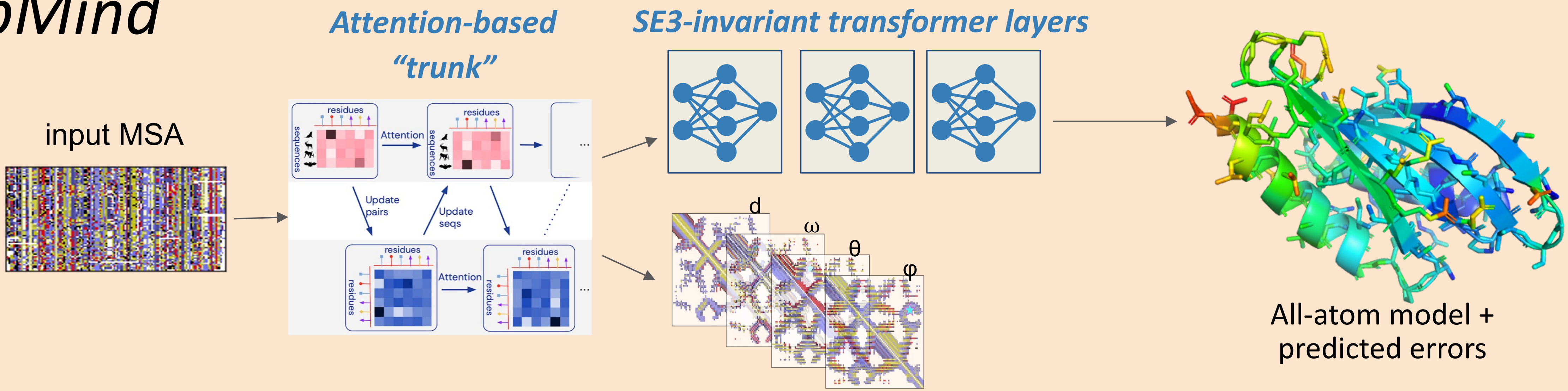
Figure from Callaway E. *Nature*, 2020

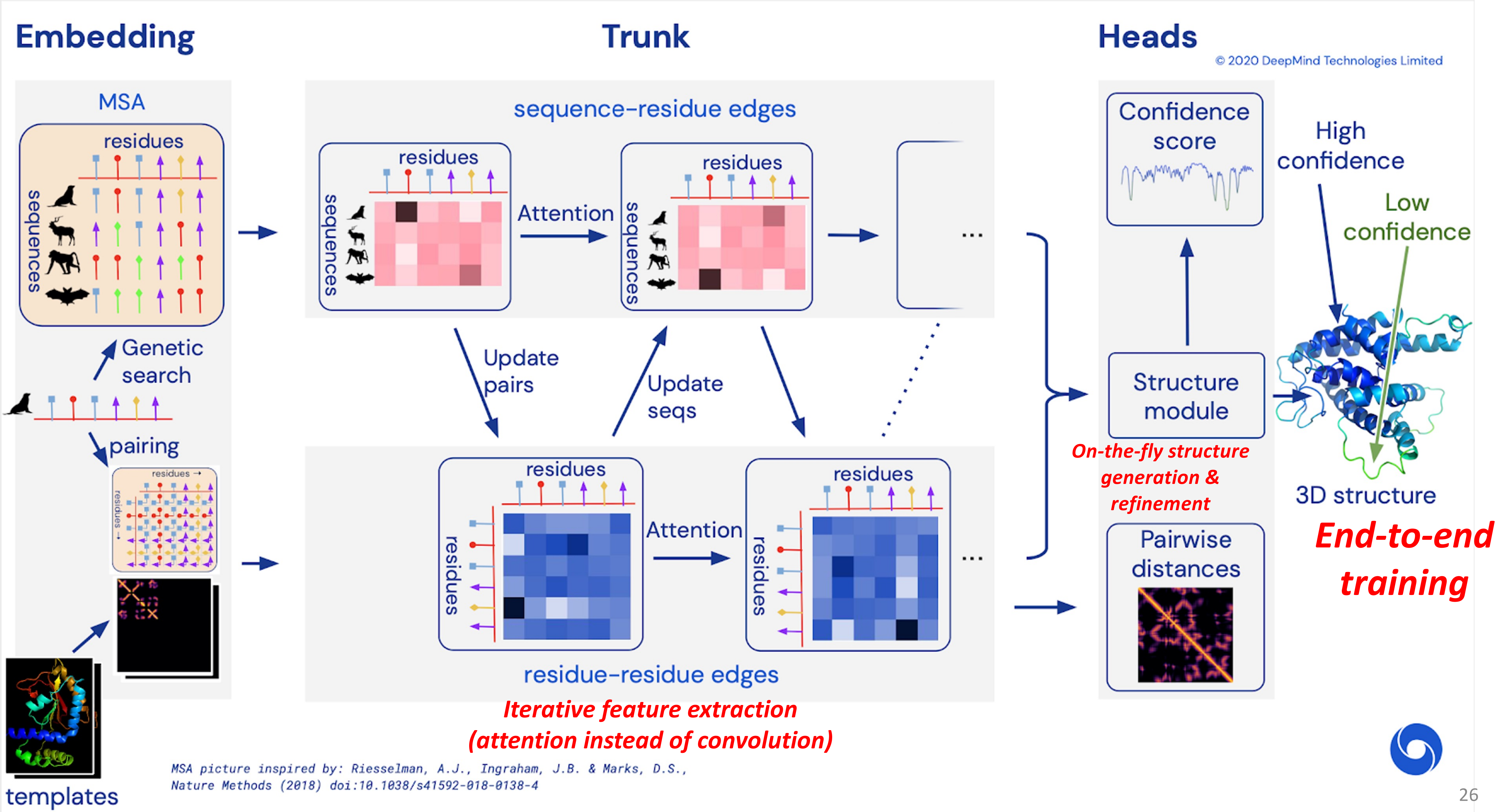
A differentiable end-to-end structure predictor

trRosetta



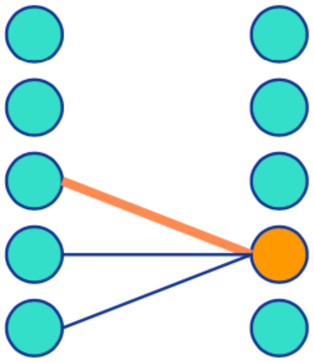
DeepMind





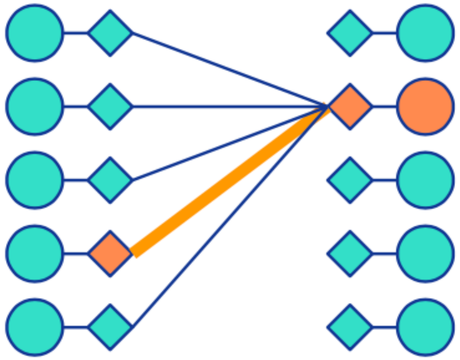
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

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



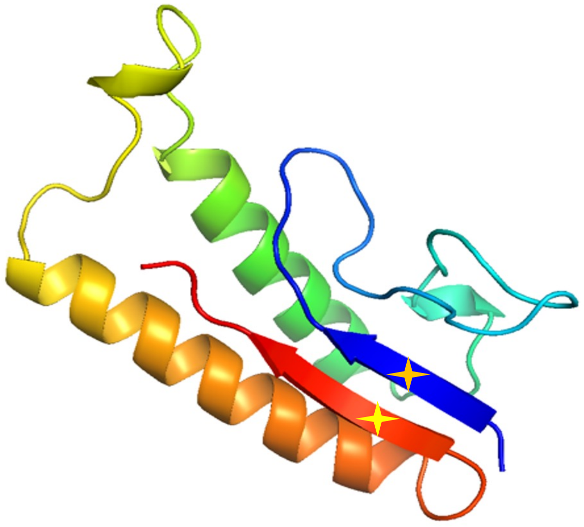
Convolutional Networks (e.g. computer vision)

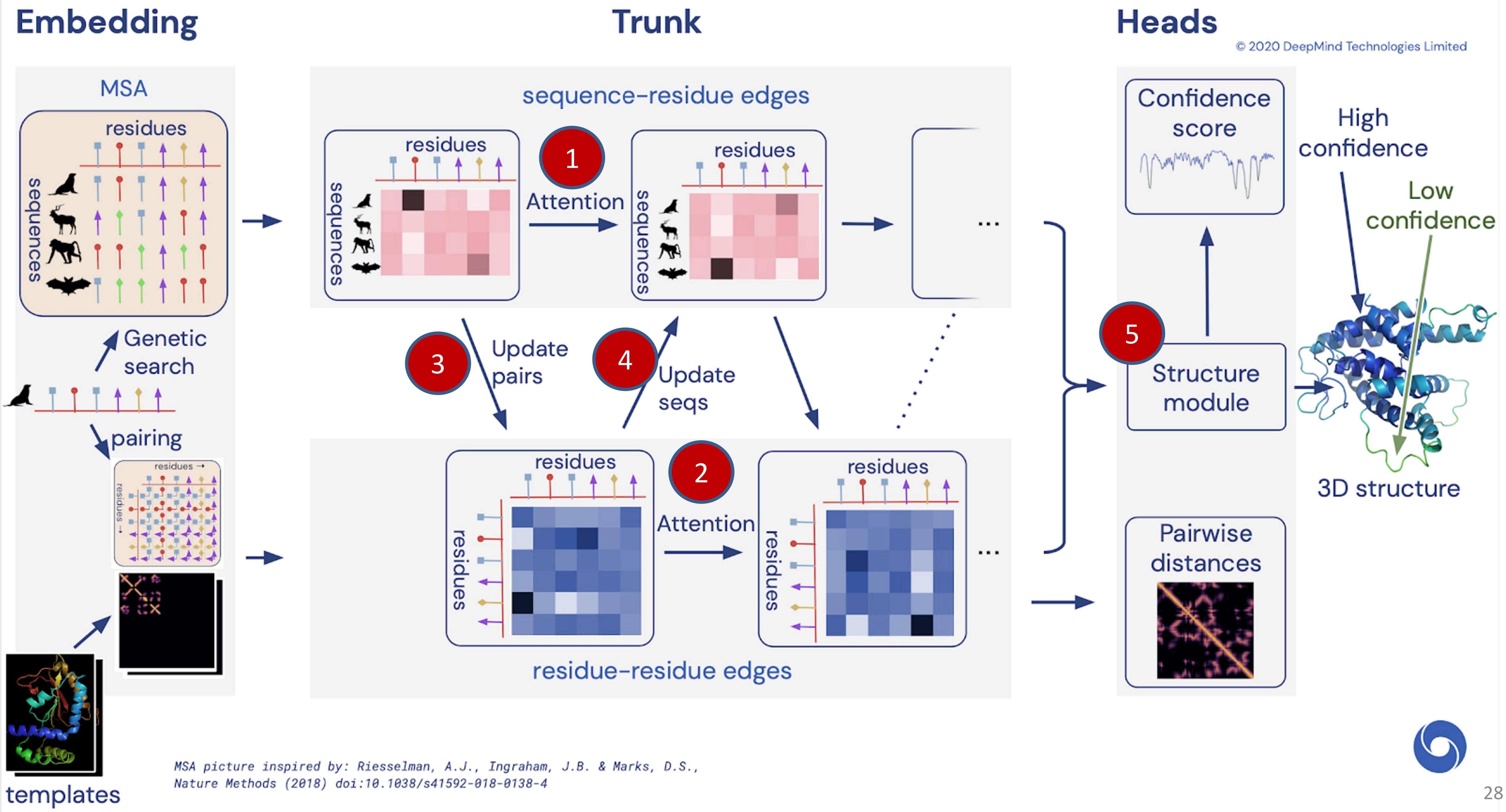
- data in regular grid
- information flow to local neighbours



Attention Module (e.g. language)

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

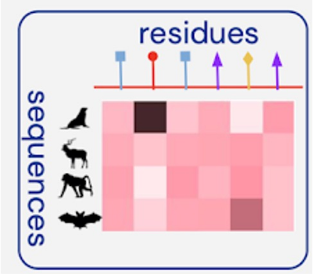




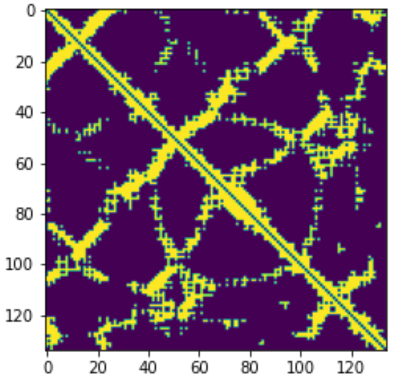
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

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)

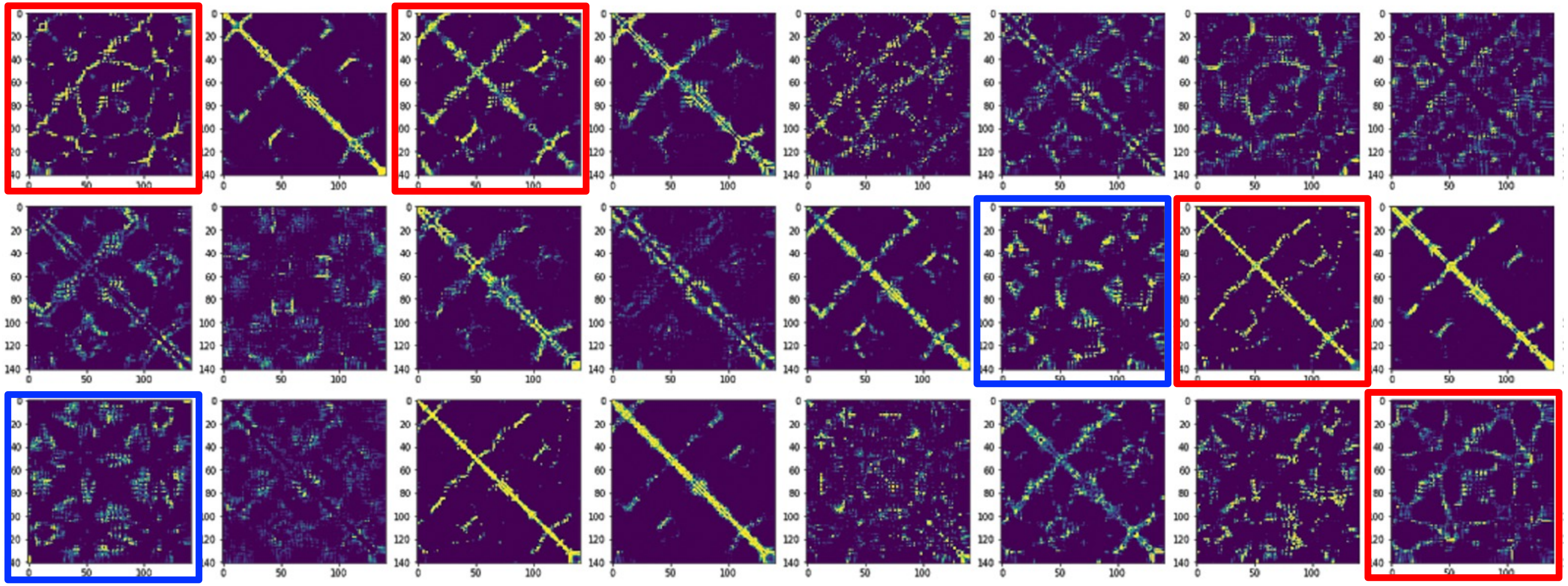


True contact map

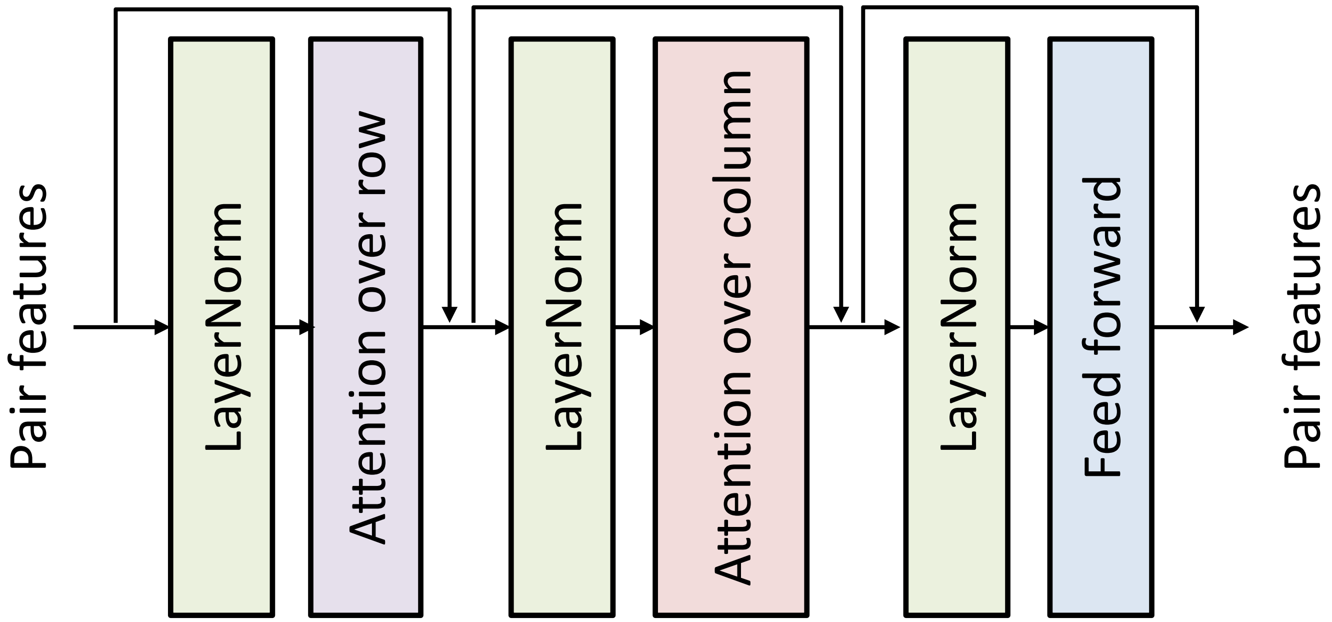


Attention heads

Last 3 blocks

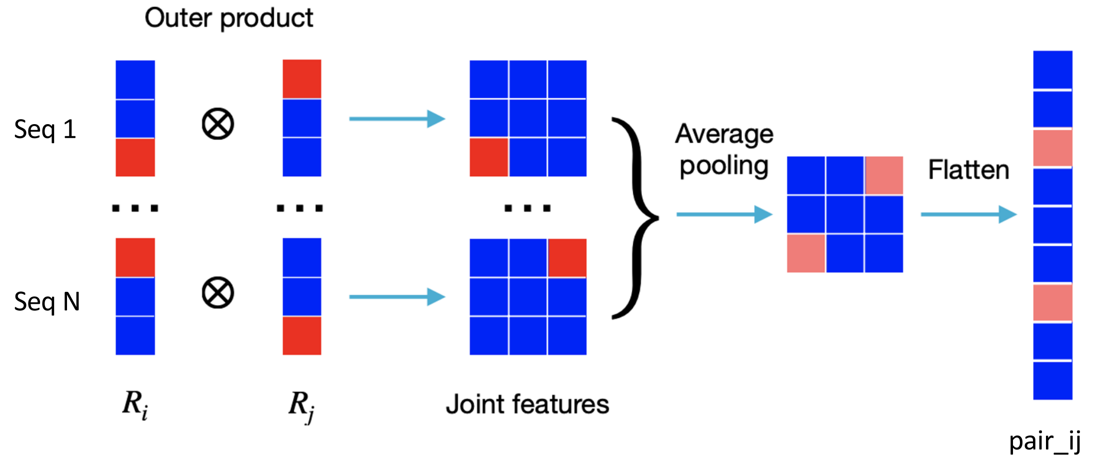


Component 2: Update pair features via self-attention

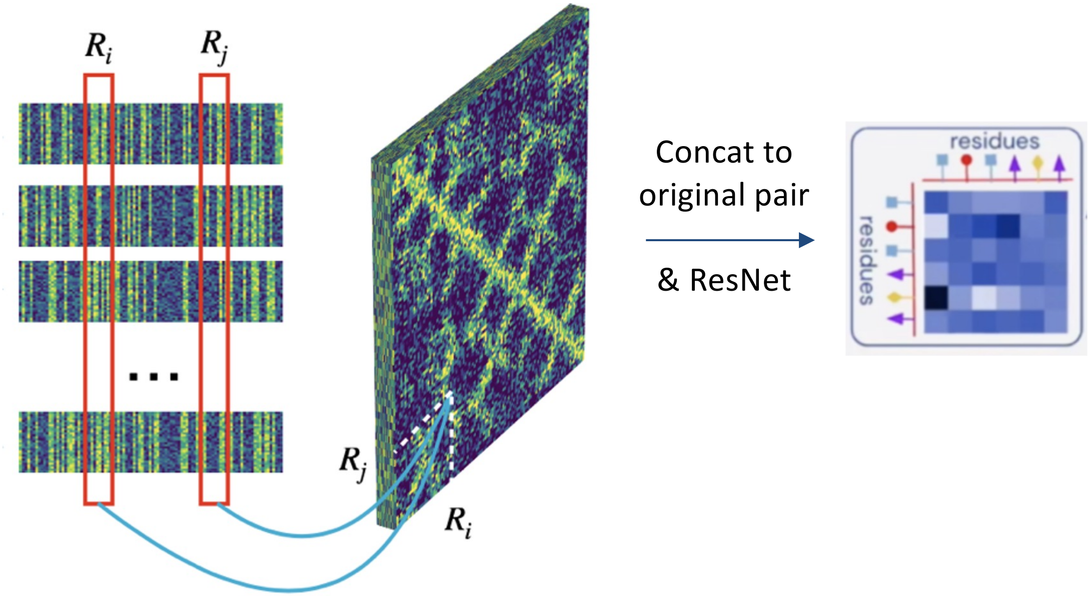


Axial Attention (attention over rows then columns) to reduce memory requirements & computation time

Component 3: Extract pair features from MSA



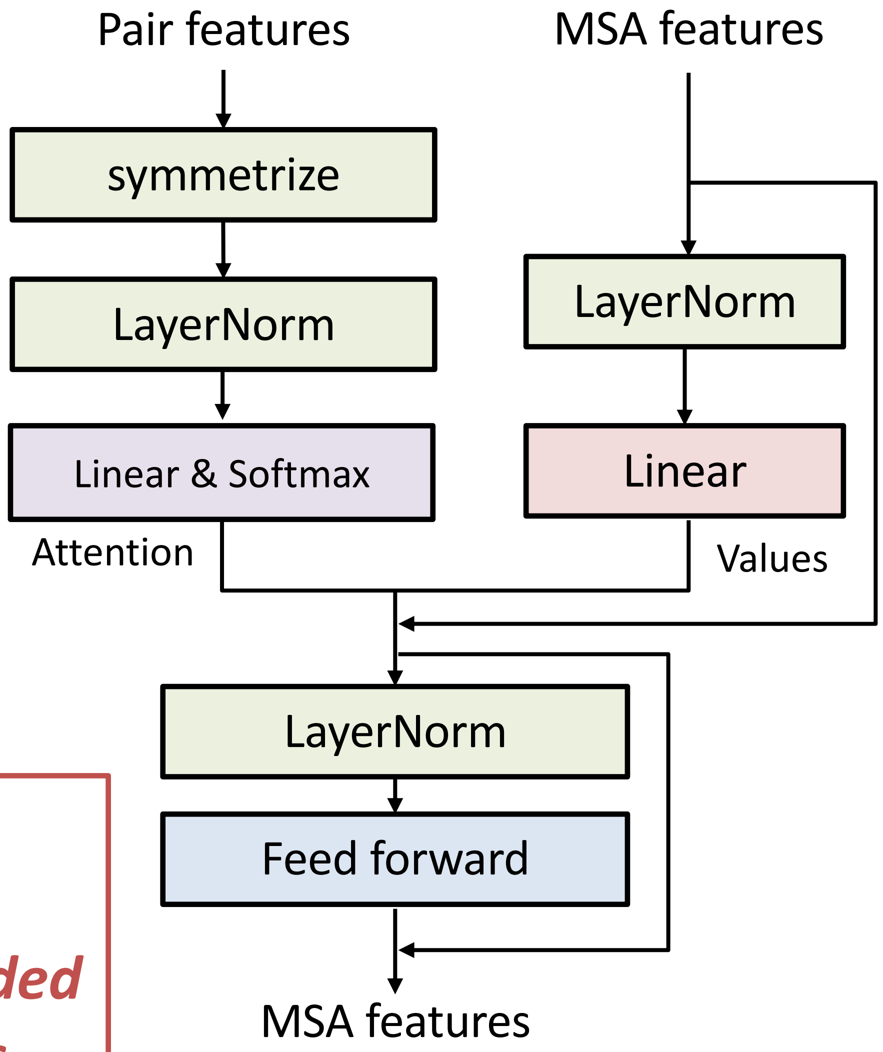
Non-interacting pairs → Broader distribution
Interacting pairs (co-mutating) → Sharper distribution



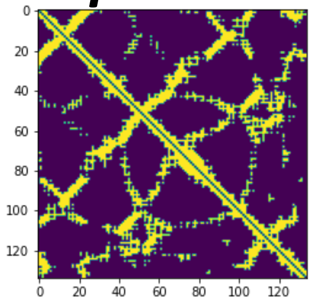
Outer product & aggregate

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

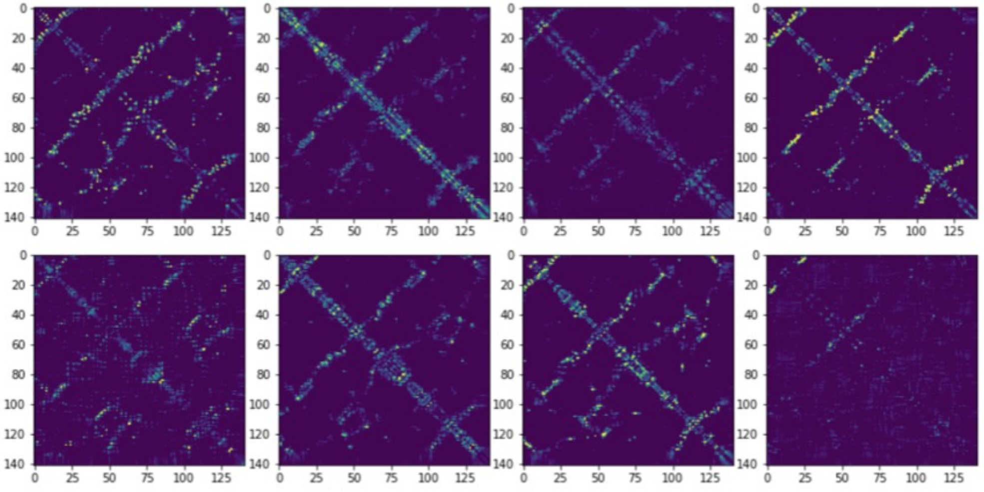


True contact map



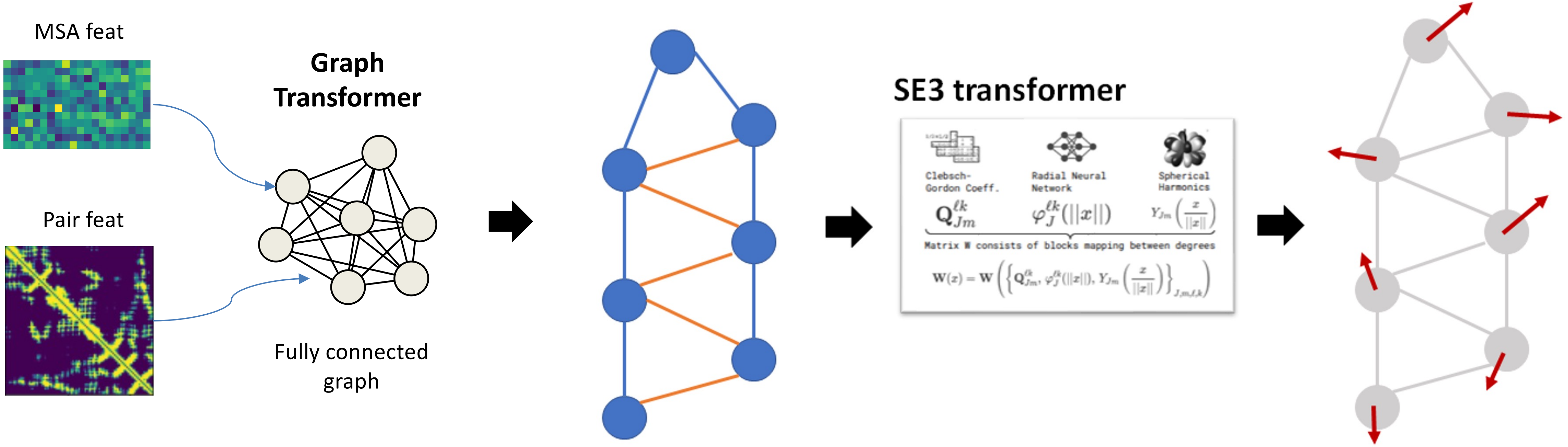
Attention heads

Last 2 blocks



Attention from structure information encoded in pair features

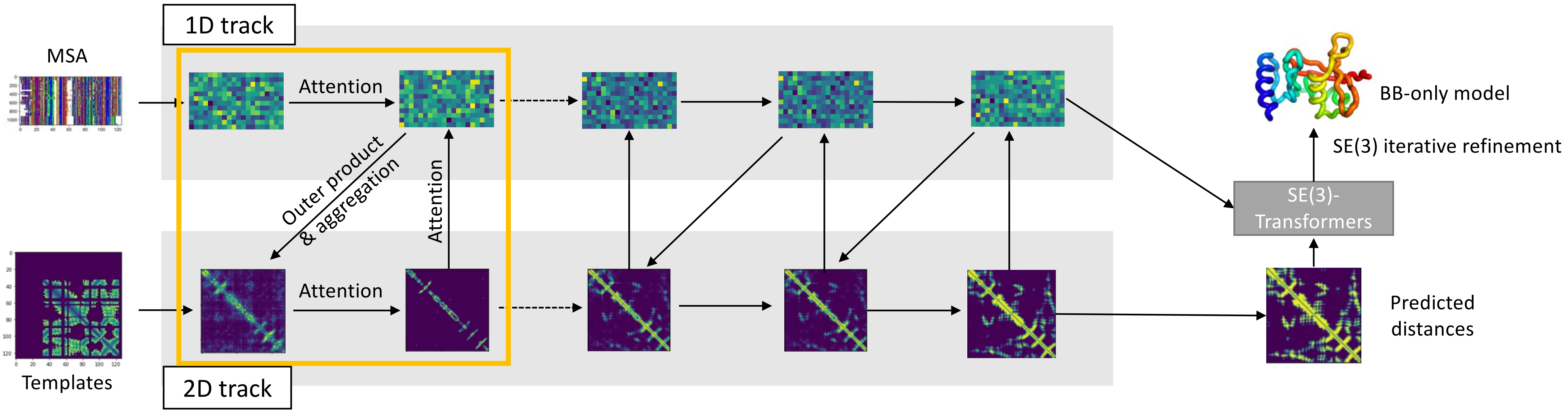
Component 5: SE(3)-Transformer for structure refinement



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

- Predicted offset vectors
- Predicted per-residue errors

RosettaFold 2-track model: Reproduce Alphafold 2 based on underlying principles



12 two-track blocks (orange box) + SE(3)-Transformer at the end
Trained on protein structures in PDB (clustered w/ seqID cutoff 30%)

What happens during iteration?