# CS 6824:
# Attention and Transformers for Prediction of Protein Structures and Interactions: RoseTTAFold

**Acknowledgement**:
Many of the images in the slides are derived from images.google.com or other publicly available sources.

# AlphaFold2 and RoseTTAFold

John Jumper[1,4 ✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4 ✉]

INSTITUTE FOR Protein Design
UNIVERSITY of WASHINGTON

About  Research  Labs  People  News  Contact  Search

RoseTTAFold: Accurate protein structure prediction accessible to all

July 15, 2021                                             Publication

## Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,21,7], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23*]

SARS-CoV-2 is evolving fast. What will it do next?

Effects on surface water from hydraulic fracturing

Baby bats babble like human babies

**Science** AAAS

20 AUGUST 2021

**PREDICTING STRUCTURES**
Deep learning accurately folds proteins

https://www.science.org/doi/10.1126/science.abj8754

https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/

# The RoseTTAFold System



https://www.science.org/doi/10.1126/science.abj8754

In this architecture, information flows back and forth between the 1D amino acid sequence information, the 2D distance map, and the 3D coordinates, allowing the network to collectively reason about relationships within and between sequences, distances, and coordinates. By contrast, reasoning about 3D atomic coordinates in the two-track AlphaFold2 architecture happens after processing of the 1D and 2D information is complete (although end-to-end training does link parameters to some extent).

To generate final models, combined and averaged the 1D features and 2D distance and orientation predictions produced for each of the crops and then used two approaches to generate final 3D structures. In the first, the predicted residue-residue distance and orientation distributions are fed into pyRosetta to generate all-atom models. In the second, the averaged 1D and 2D features are fed into a final SE(3)-equivariant layer, and, after end-to-end training from amino acid sequence to 3D coordinates,
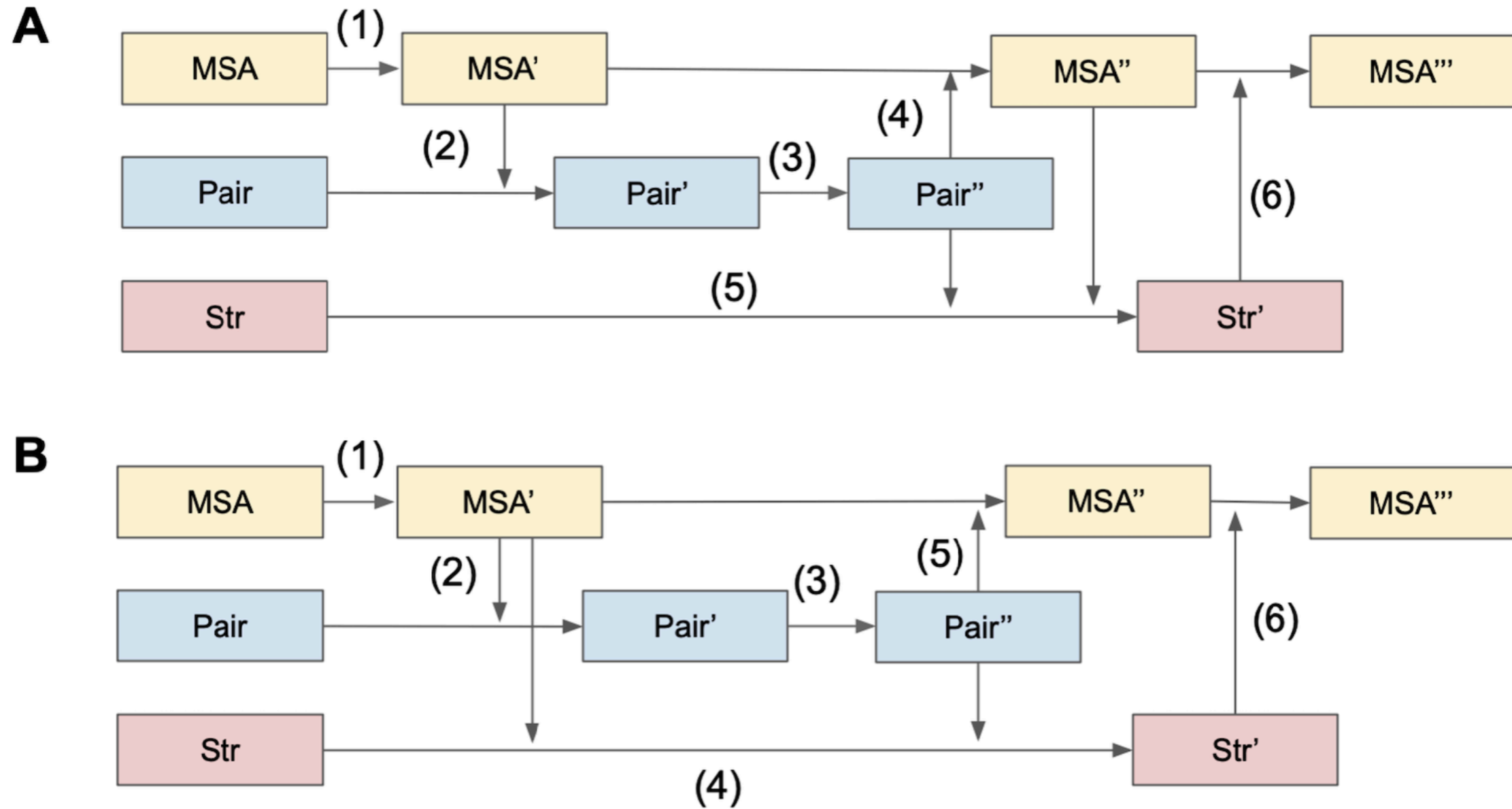
# 3-track Blocks



**Fig. S13. Two different 3-track block definitions. (A)** MSA and pair features are synchronized before structure updates. **(B)** The structure is updated based on unsynchronized MSA and pair features. The numbers in parentheses indicate the order of calculation.

# Performance of different model architectures

**Table S1**. Performance of different model architectures in terms of inter-residue geometry prediction loss (cross entropy), top L long-range contact accuracy and $C_\alpha$-lDDT.

| Architecture | Inter-residue geometry loss | Top L long-range contact accuracy | $C_\alpha$-lDDT |
|---|---|---|---|
| **Single Track (Sequential processing of MSA and pair feature)** | | | |
| Architecture 1) Hand-crafted features + 2D convolution | 5.56 | 54% | - |
| Architecture 2) MSA encoder + 2D convolution | 5.49 | 56% | - |
| Architecture 3) MSA encoder + Axial attention | 5.14 | 58% | - |
| **2-track (Parallel track for MSA and pair features)** | | | |
| Architecture 4) Untied + addition + cross | 5.54 | 54% | - |
| Architecture 5) Untied + addition + direct | 5.18 | 58% | - |
| Architecture 6) Untied + concat + direct | 5.01 | 60% | - |
| Architecture 7) Soft-tied + concat + direct | 4.84 | 62% | - |
| Architecture 8) architecture 7 + scale-up | 4.50 | 67% | - |
| Architecture 9) architecture 8 + SE(3) structure module | 4.54 | 67% | 0.70 |
| **3-track (Parallel track for MSA, pair, and 3D coordinates)** | | | |
| Architecture 10) Structure update w/ unsynchronized MSA and pair features (**Fig. S13B**) | 4.63 | 64% | 0.68 |
| Architecture 11) Structure update w/ synchronized MSA and pair features (**Fig. S13A**) | 4.36 | 69% | 0.72 |
| Architecture 12) architecture 11 + SE(3) structure module | 4.39 | 69% | 0.77 |

# SE(3)-Transformers

## SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks

**Fabian B. Fuchs**[*][†]
Bosch Center for Artificial Intelligence
A2I Lab, Oxford University
fabian@robots.ox.ac.uk

**Daniel E. Worrall**[*]
Amsterdam Machine Learning Lab, Philips Lab
University of Amsterdam
d.e.worrall@uva.nl

**Volker Fischer**
Bosch Center for Artificial Intelligence
volker.fischer@de.bosch.com

**Max Welling**
Amsterdam Machine Learning Lab
University of Amsterdam
m.welling@uva.nl

https://arxiv.org/pdf/2006.10503

SE(3)-Transformer is a variant of the self-attention module for 3D point clouds and graphs, which is *equivariant* under continuous 3D roto- translations

# SE(3)-Transformers

Attention mechanism

$$\text{Attn}\left(\mathbf{q}_i, \{\mathbf{k}_j\}, \{\mathbf{v}_j\}\right) = \sum_{j=1}^{n} \alpha_{ij} \mathbf{v}_j, \qquad \alpha_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{j'=1}^{n} \exp(\mathbf{q}_i^\top \mathbf{k}_{j'})}$$

In the case of *self-attention* the query, key, and value vectors are embeddings of the input features, so

$$\mathbf{q} = h_Q(\mathbf{f}), \qquad\qquad \mathbf{k} = h_K(\mathbf{f}), \qquad\qquad \mathbf{v} = h_V(\mathbf{f}),$$

where $\{h_Q, h_K, h_V\}$ are, in the most general case, neural networks. The query $q_i$ is associated with a point i in the input, which has a geometric location $x_i$. Thus if we have n points, we have n possible queries. For query $q_i$, we say that node i *attends* to all other nodes j/= i.

A key property of self-attention is *permutation equivariance*. Permutations of point labels 1, ..., n lead to permutations of the self-attention output. This guarantees the attention output does not depend arbitrarily on input point ordering.

The SE(3)-transformer is a special case of this attention mechanism, inheriting permutation equivariance. However, it limits the space of learnable functions to rotation and translation equivariant ones.
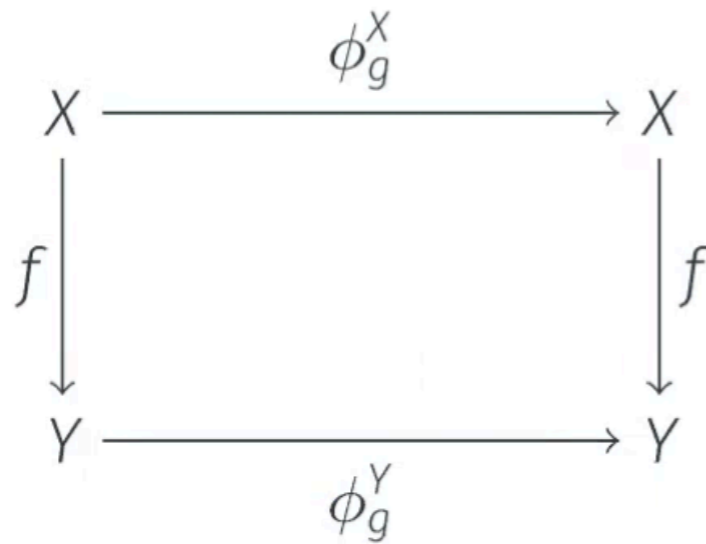
https://arxiv.org/pdf/2006.10503

# Equivariance

let $f : X \to Y$ be a function and $\phi^X$ and $\phi^Y$ be transformations on $X$ and $Y$
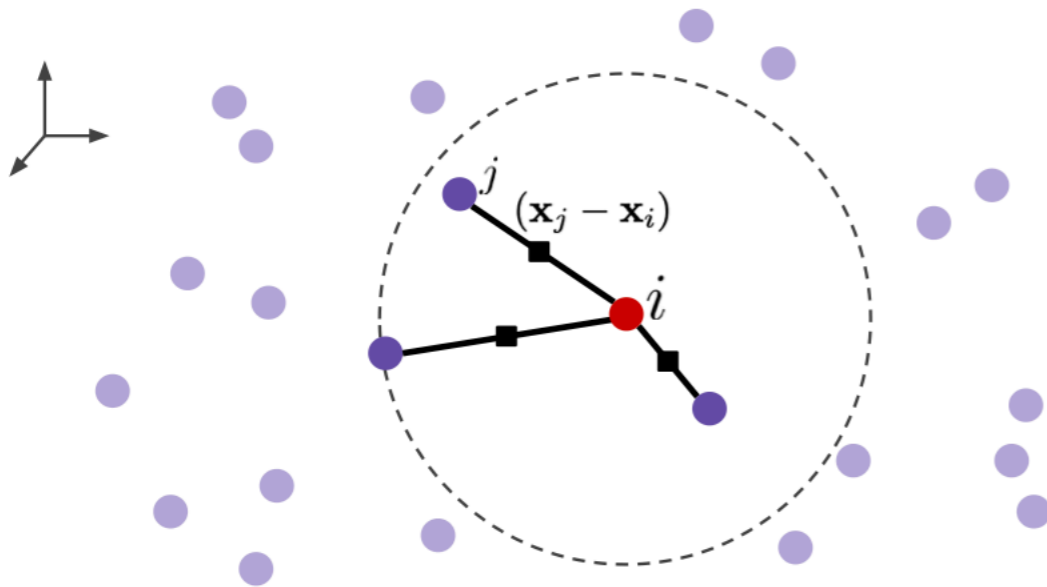
then $f$ is said to be *equivariant* iff

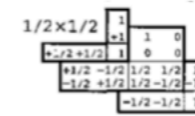$$f \circ \phi_g^X = \phi_g^Y \circ f$$
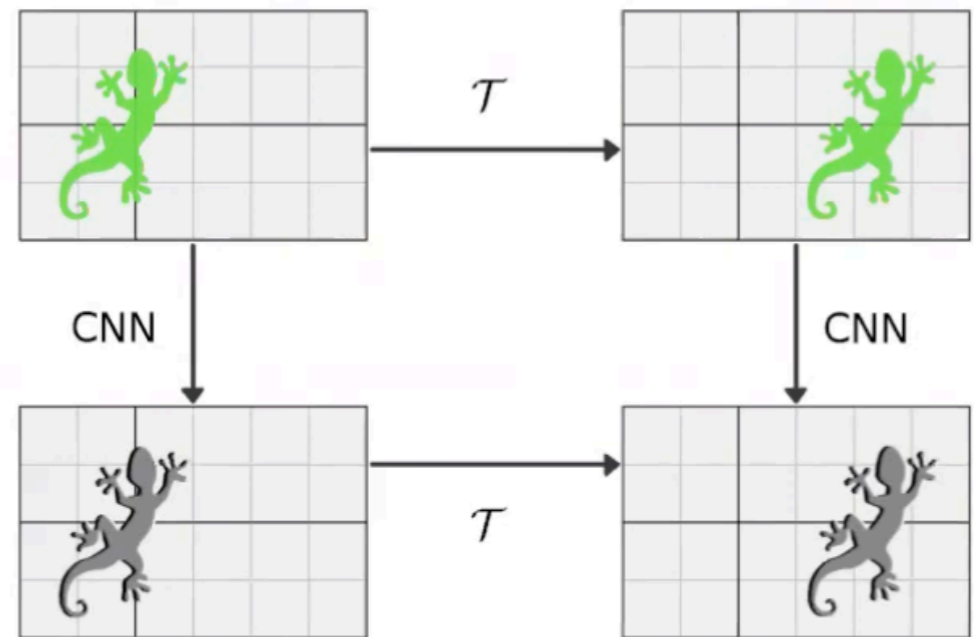
# SE(3)-Transformers



**Step 1**: Get nearest neighbours and relative positions

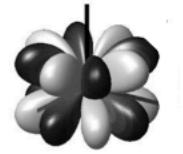**Step 2**: Get SO(3)-equivariant weight matrices

Clebsch-Gordon Coeff. $\mathbf{Q}^{\ell k}_{Jm}$
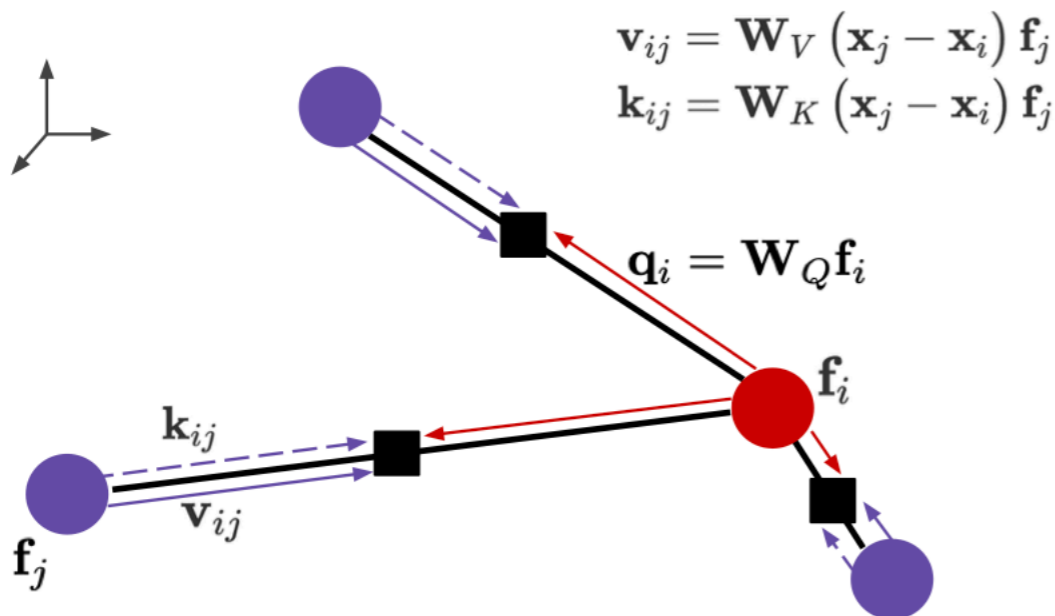
Radial Neural Network $\varphi^{\ell k}_J(||x||)$

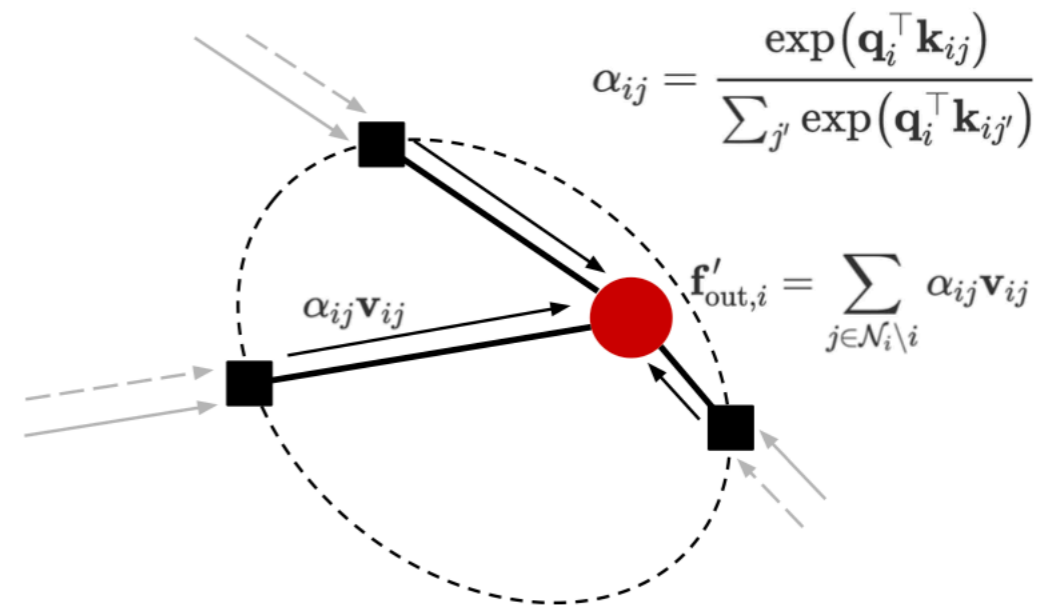Spherical Harmonics $Y_{Jm}\left(\dfrac{x}{||x||}\right)$

Matrix W consists of blocks mapping between degrees

$$\mathbf{W}(x) = \mathbf{W}\left(\left\{\mathbf{Q}^{\ell k}_{Jm}, \varphi^{\ell k}_J(||x||), Y_{Jm}\left(\dfrac{x}{||x||}\right)\right\}_{J,m,\ell,k}\right)$$

**Step 3**: Propagate queries, keys, and values to edges

$$\mathbf{v}_{ij} = \mathbf{W}_V\left(\mathbf{x}_j - \mathbf{x}_i\right)\mathbf{f}_j$$
$$\mathbf{k}_{ij} = \mathbf{W}_K\left(\mathbf{x}_j - \mathbf{x}_i\right)\mathbf{f}_j$$
$$\mathbf{q}_i = \mathbf{W}_Q\mathbf{f}_i$$

**Step 4**: Compute attention and aggregate

$$\alpha_{ij} = \dfrac{\exp\left(\mathbf{q}_i^\top \mathbf{k}_{ij}\right)}{\sum_{j'}\exp\left(\mathbf{q}_i^\top \mathbf{k}_{ij'}\right)}$$

$$\mathbf{f}'_{\text{out},i} = \sum_{j\in\mathcal{N}_i\backslash i}\alpha_{ij}\mathbf{v}_{ij}$$

https://arxiv.org/pdf/2006.10503
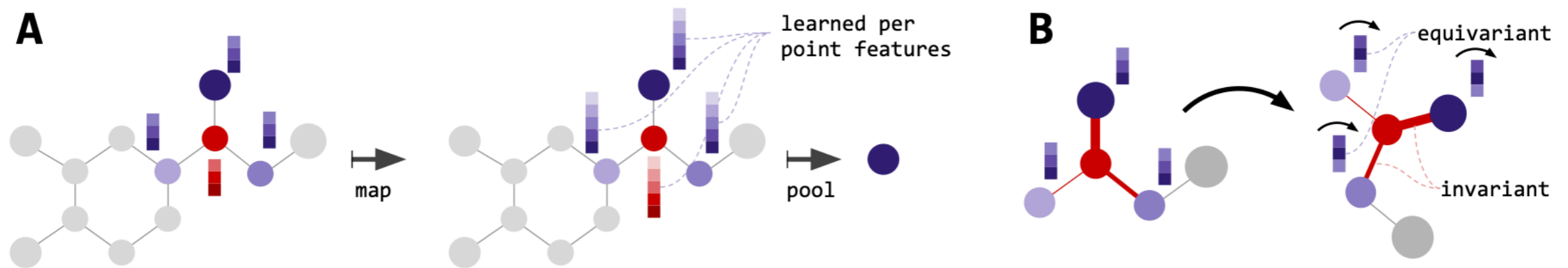
# SE(3)-Transformers



A) Each layer of the SE(3)-Transformer maps from a point cloud to a point cloud (or graph to graph) while guaranteeing equivariance. For classification, this is followed by an invariant pooling layer and an MLP. B) In each layer, for each node, attention is performed. Here, the red node attends to its neighbours. Attention weights (indicated by line thickness) are invariant w.r.t. input rotation.

# RoseTTAFold 1D and 2D Tracks
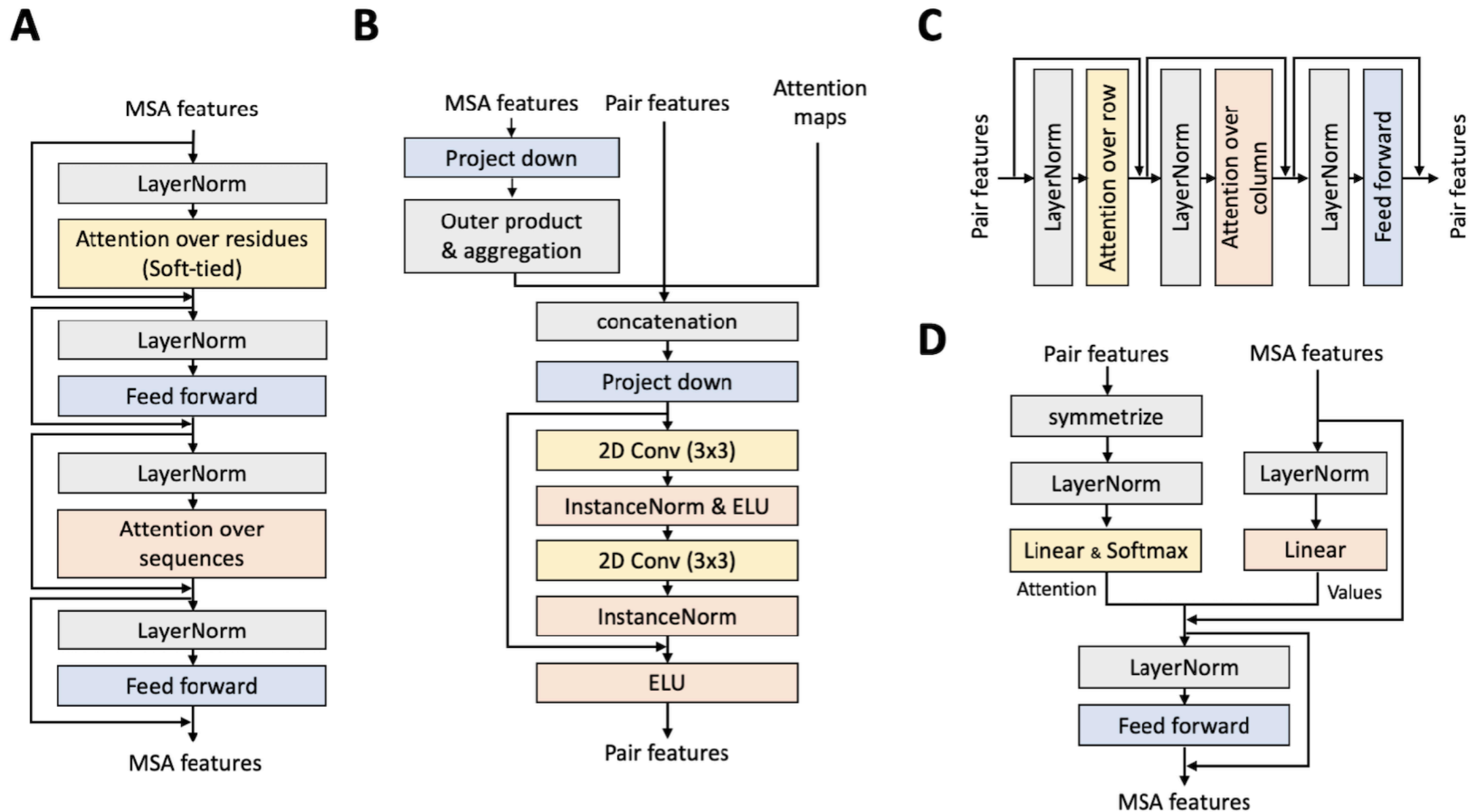


(A) MSA updates via self-attention on MSA features. The attention maps over residues are softly tied. (B) Pair feature updates based on co-evolution signals derived from MSA features by taking outer-products and weighted averages. (C) Pair feature refinement through axial attention. (D) MSA feature updates based on attention maps derived from given pair features.

# RoseTTAFold 3D Track
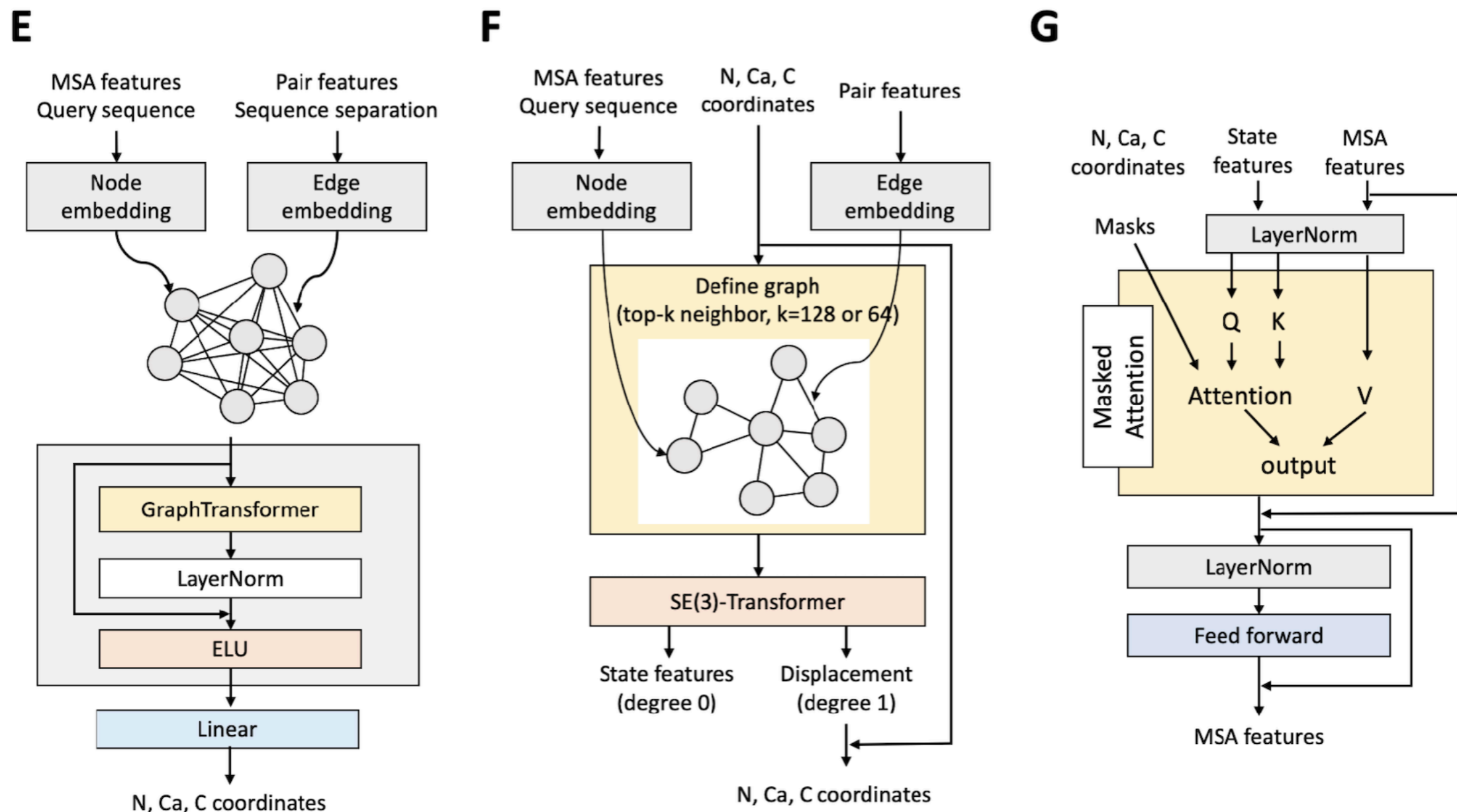


(E) Initial N, C$_\alpha$, C coordinate generation using Graph Transformer architecture. (F) 3D coordinate refinements with SE(3)-Transformer. degree 0 node features (scaler node features, called state features here) are used to calculate attention maps for structure-based MSA. degree 1 node features (vector node features) encode the positions of N and C atoms by including displacement vectors to the corresponding C$_\alpha$ atoms.(G) MSA feature updates based on given 3D structures using masked attention maps.

# RoseTTAFold Architecture

1. Update pair features with coevolution signal derived from MSA features

2. Refine pair features via row and column-wise self-attention

3. Update MSA features based on structure information encoded in pair features

4. Initial 3D structure prediction using Graph Transformer-based architecture

Input is defined as a fully connected graph with nodes representing the residues in the protein. The node and edge embeddings are learned from the averaged MSA features combined with a one-hot encoded query sequence and the pair features along with sequence separation, respectively. The backbone coordinates are estimated using a stack of four Graph Transformer layers followed by a simple linear transformation to predict Cartesian coordinates of N, $C_\alpha$, C atoms for each residue node.

5. Structure updates through SE(3)-Transformer

6. Update MSA features based on a 3D structure

Similar to the MSA updates based on pair features in the 2-track model, MSA features are updated based on attention maps derived from the current 3D structures. Four attention maps are calculated based on the state features, and they are masked based on the $C_\alpha$ distances with four different cutoffs (8, 12, 16, and 20 Å) so that it only attends to the neighbors in 3D space.

7. Residue pairwise distance and orientation prediction

8. Additional structure module for iterative refinement through the network

# RoseTTAFold Results
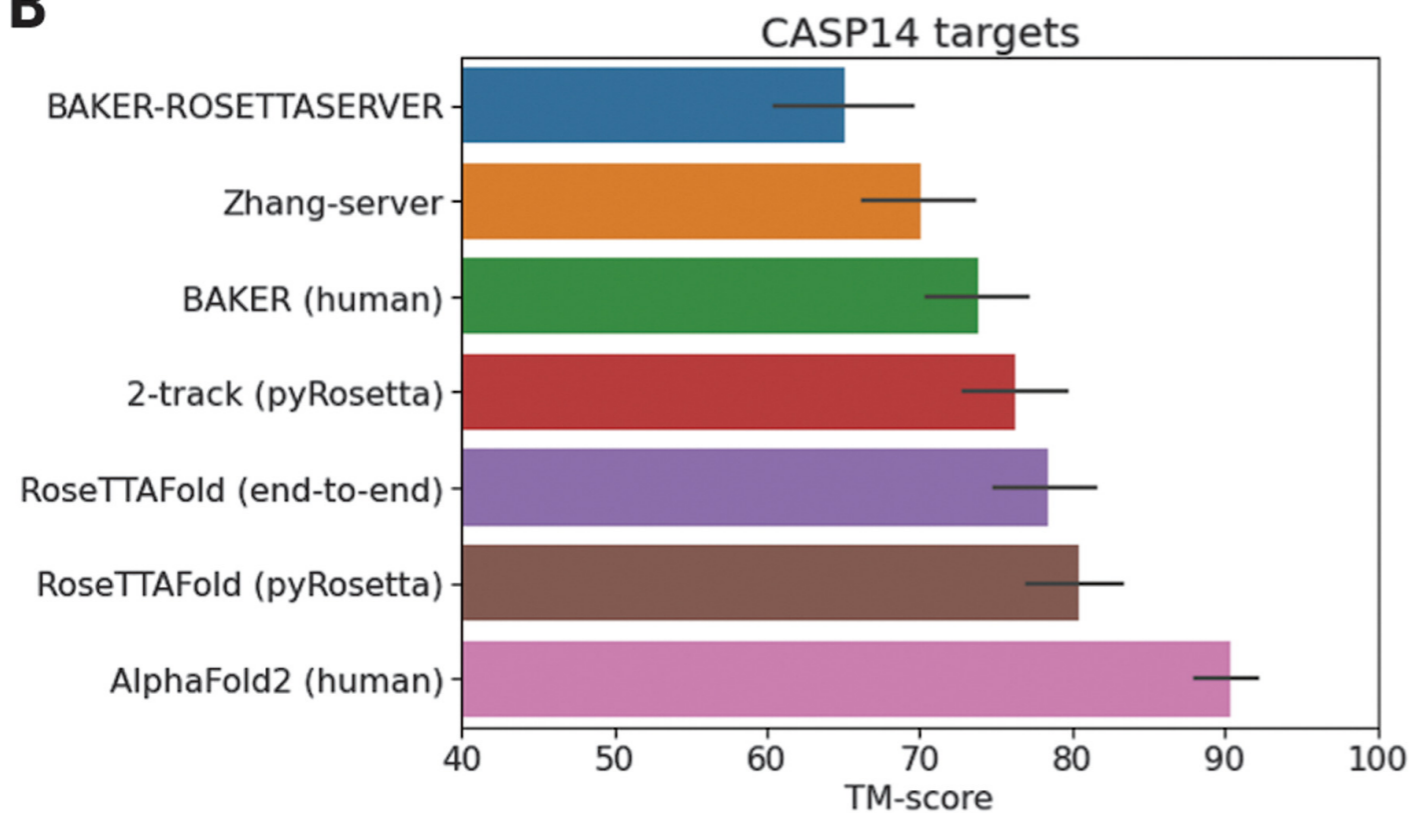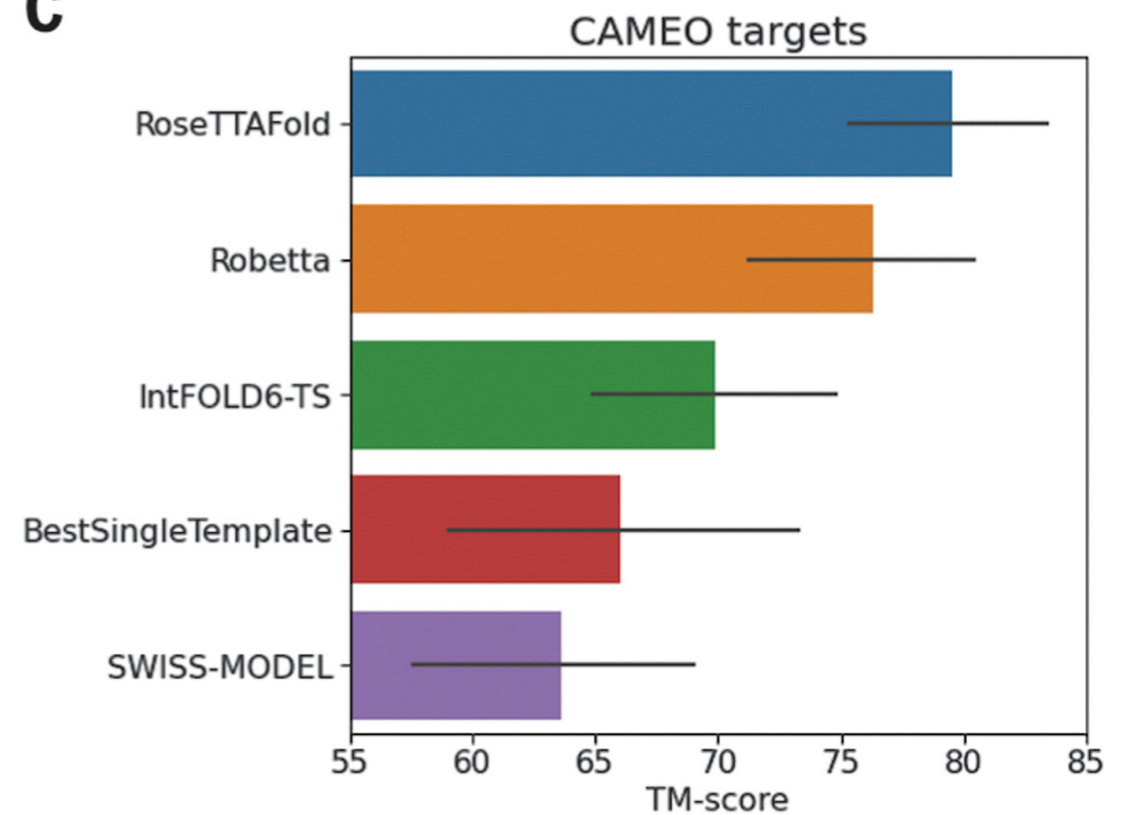


(**B**) Average TM-score of prediction methods on the CASP14 targets. Zhang-server and BAKER-ROSETTASERVER were the top two server groups, whereas AlphaFold2 and BAKER were the top two human groups in CASP14; BAKER-ROSETTASERVER and BAKER predictions were based on trRosetta. Predictions with the two-track model and RoseTTAFold (both end-to-end and pyRosetta version) were completely automated. (**C**) Blind benchmark results on CAMEO medium and hard targets; model accuracies are TM-score values from the CAMEO website (https://cameo3d.org/). In (B) and (C), the error bars represent a 95% confidence interval.

# RoseTTAFold models provide insights into function
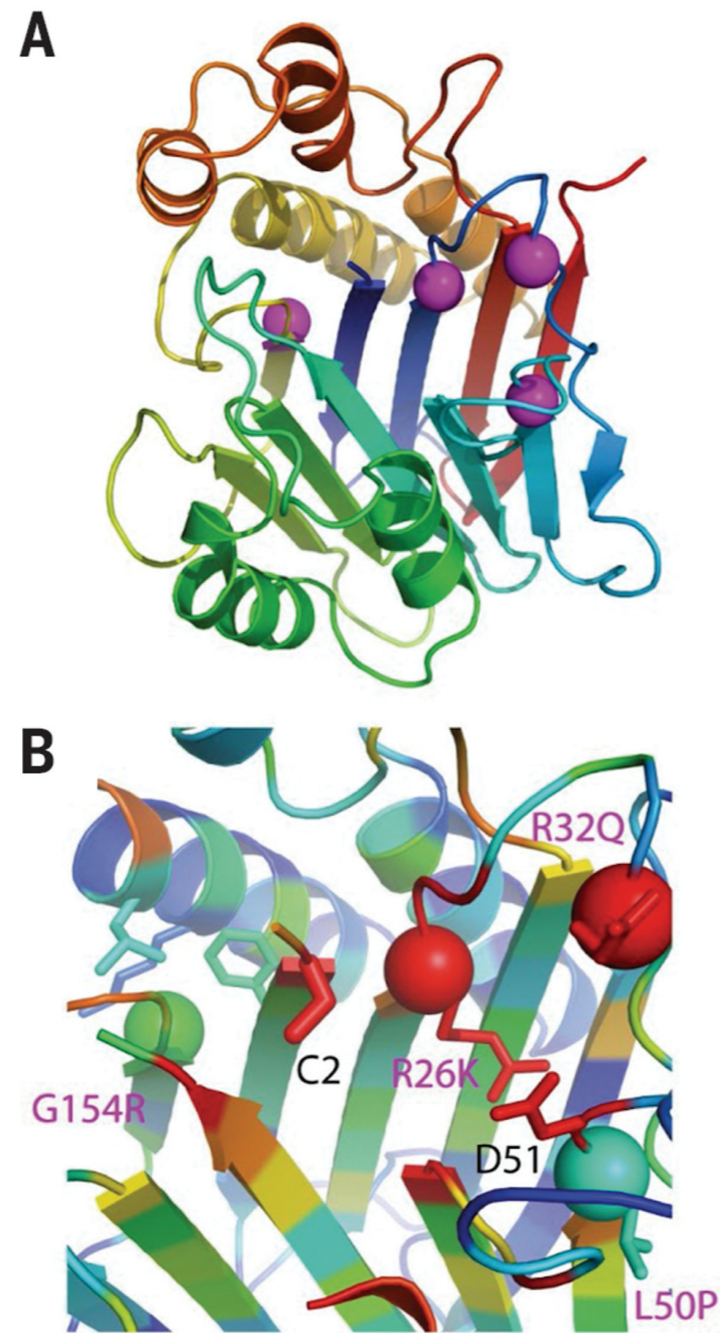


Deficiencies in TANGO2 (transport and Golgi organization protein 2) lead to metabolic disorders, and the protein plays an unknown role in Golgi membrane redistribution into the endoplasmic reticulum. The RoseTTAFold model of TANGO2 adopts an N-terminal nucleophile aminohydrolase (Ntn) fold (Fig. 3A) with well-aligned active-site residues that are conserved in TANGO2 orthologs (Fig. 3B).
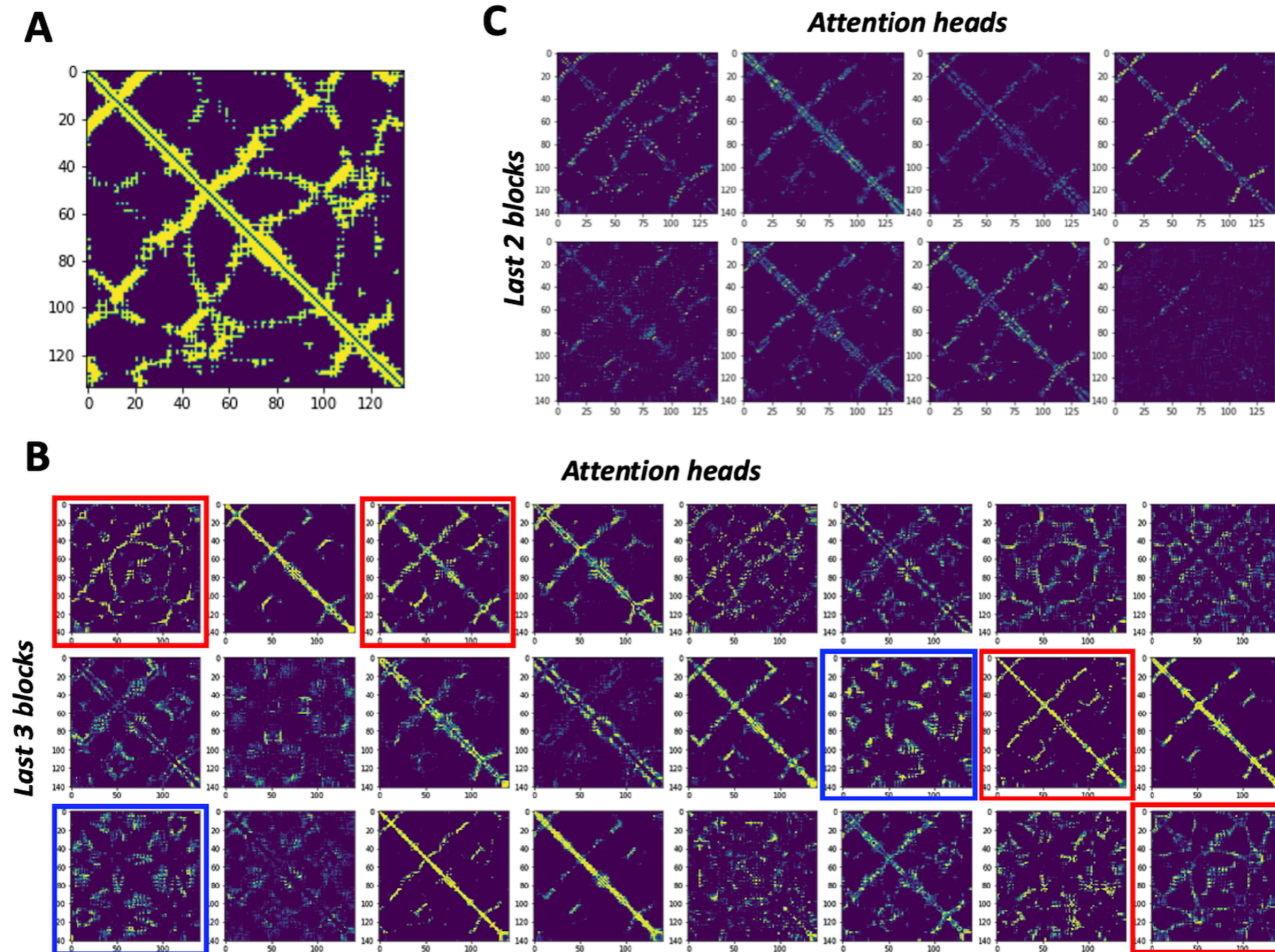
# Attention maps used to update MSA



Fig. S12. Examples of attention maps used to update MSA. (A) True contact map of CASP14 target T1049. (B) Attention maps from self-attention on MSA features for the last three blocks of the 2-track model (76M parameter model). Some of the attention heads (red boxes) resemble a true contact map. Some cases (blue boxes) only attend to the positions not making the direct contacts. (C) Attention maps derived from pair features used to update MSA features. It also shows a similar pattern to the true contact map. The attention maps shown in this figure are symmetrized for clear visualization.

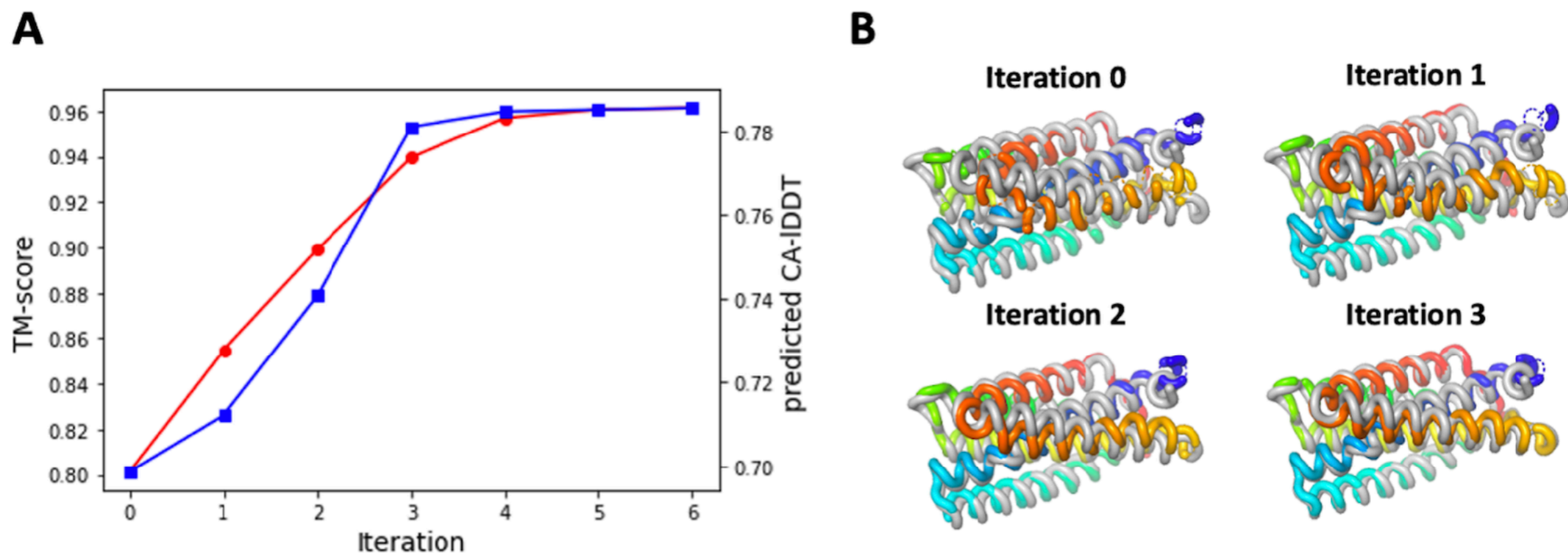# Iterative refinement using SE(3)-Transformers



**Fig. S15. An example (T1024-D1 from CASP14 targets) of Iterative refinement using SE(3)-Transformers. (A)** Model accuracy (TM-score) is improved with iterative refinement. Predicted $C_\alpha$-lDDT from the network shows a good correlation to the actual model accuracy. **(B)** The model structure at each iteration is shown. The RoseTTAFold models are colored in a rainbow (blue; N-terminal, red; C-terminal), and the native structures are colored in gray.

# Discussions

**AlphaFold 2**

Strengths:

1.



2.



3.



Weaknesses:

1.



2.



3.

**RoseTTAFold**

Strengths:

1.



2.



3.



Weaknesses:

1.



2.



3.

AI-powered Molecular Modeling | Virginia Tech