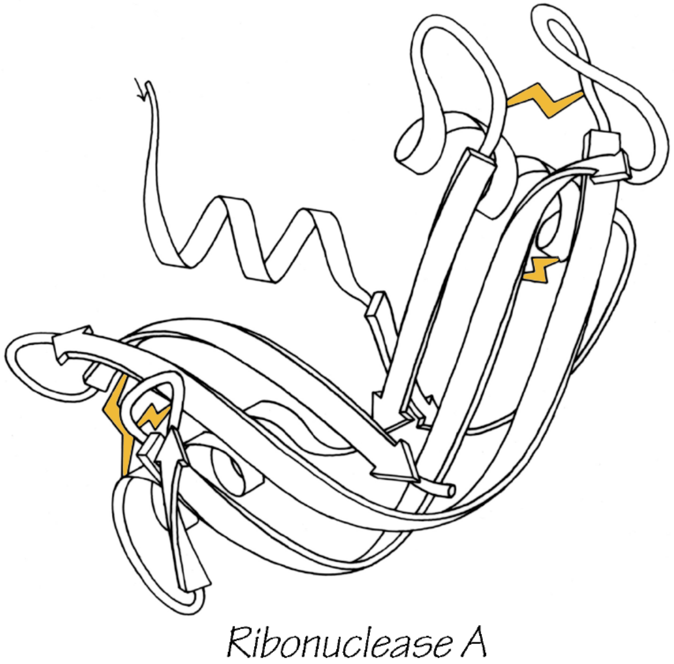
Highly accurate protein structure prediction with AlphaFold

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]

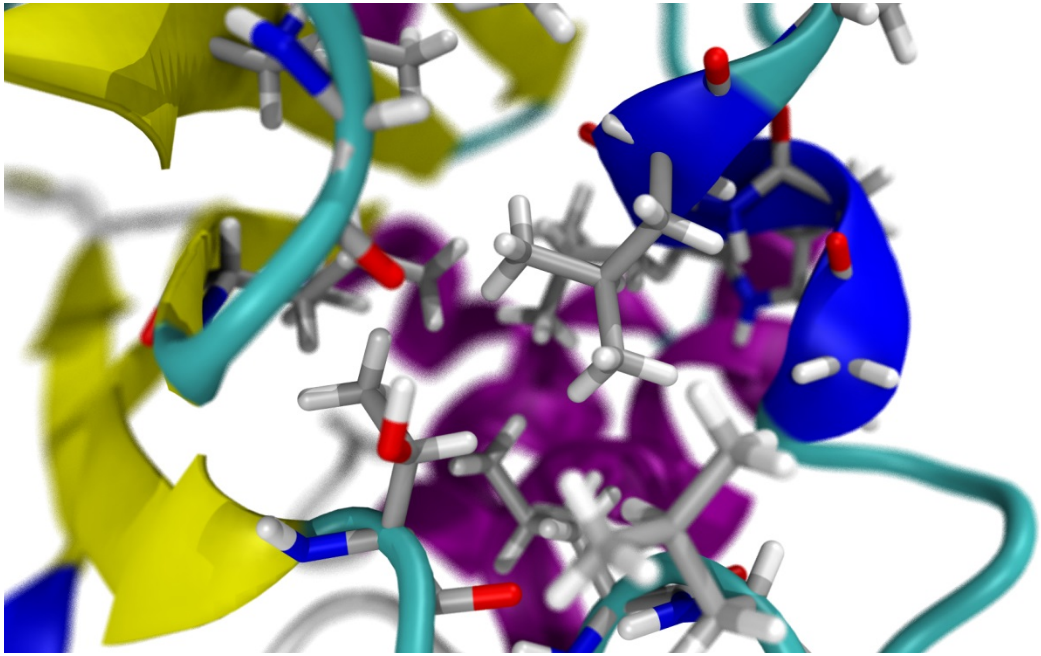
A fifty-year problem



Refolding experiment of Ribonuclease A.

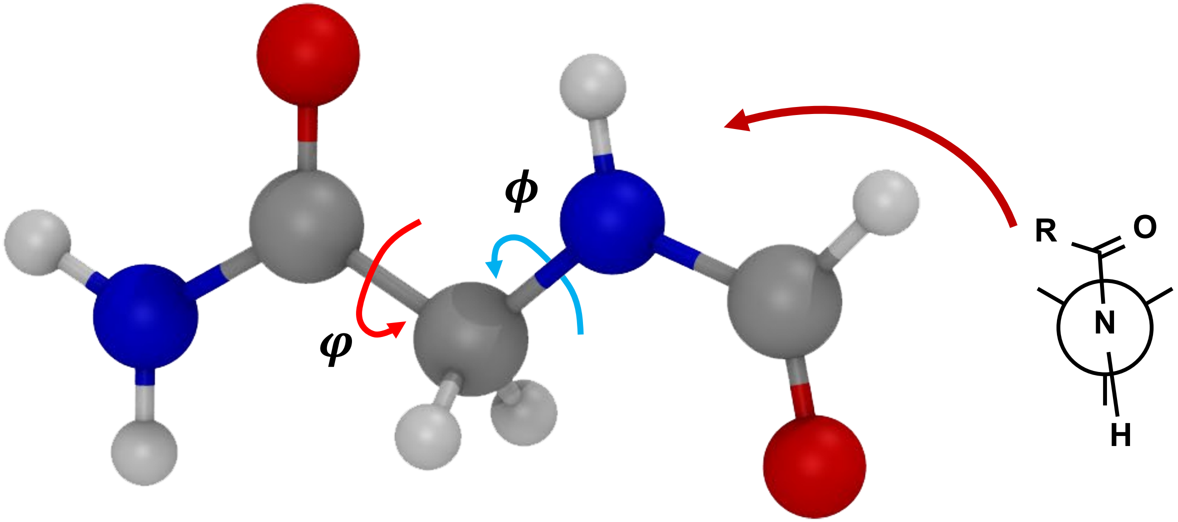
Christian B. Anfinsen
3.26.1916 – 5.14.1995

The solutions:



Could we optimize the protein by physics law?

The solutions:

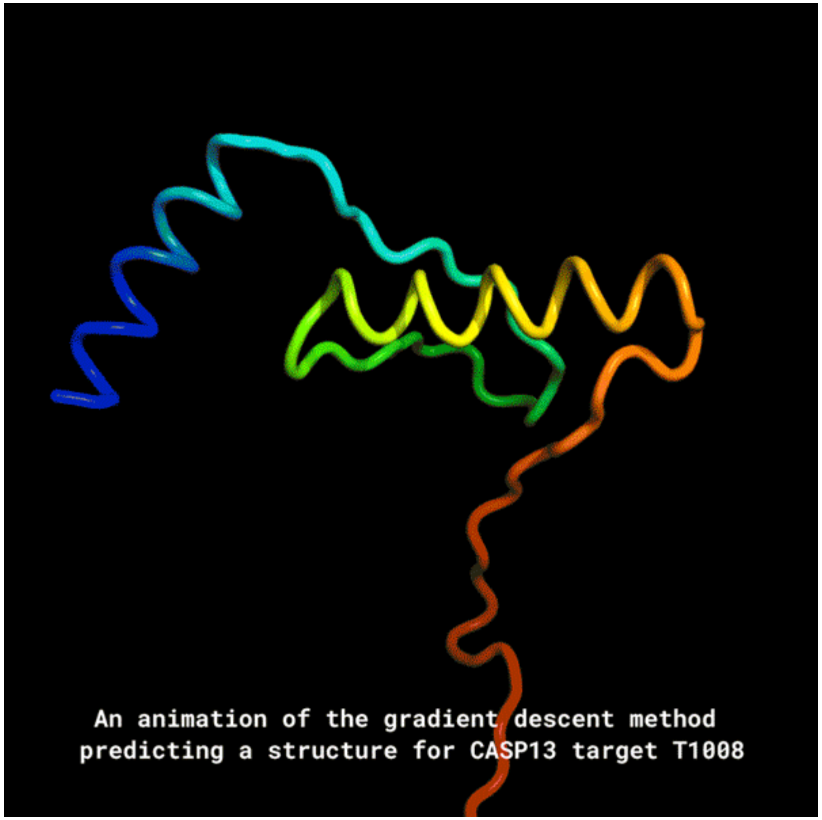


Only consider backbond: 3 conformers for φ and 3 conformers for ϕ

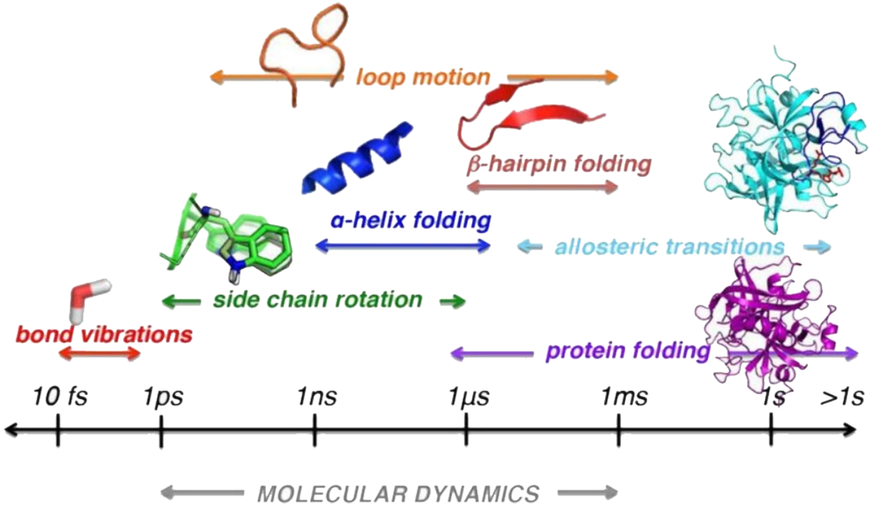
For a 200aa protein: $3^{400} \approx 10^{191}$

Using empirical constrains to reduce the searching space.

The solutions:

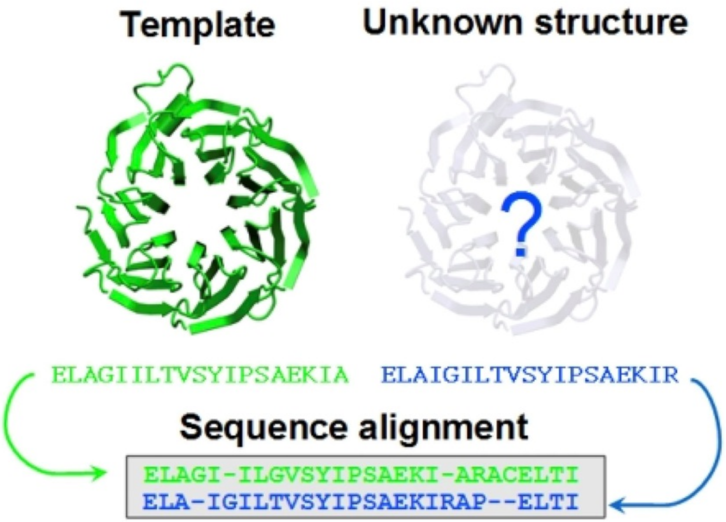


An animation of the gradient descent method predicting a structure for CASP13 target T1008



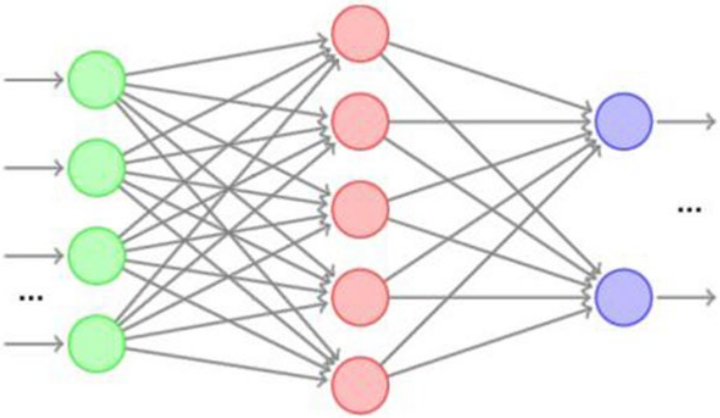
Our computation resource could not afford full atom folding simulation to microsecond level.

Templet Modeling



ROSETTA's compare modeling could predict high quality model by templets.

DeepMind



Sequence

Structure

Parameterize by template

Anfinsen suggest the mapping relationship from sequence to structure.

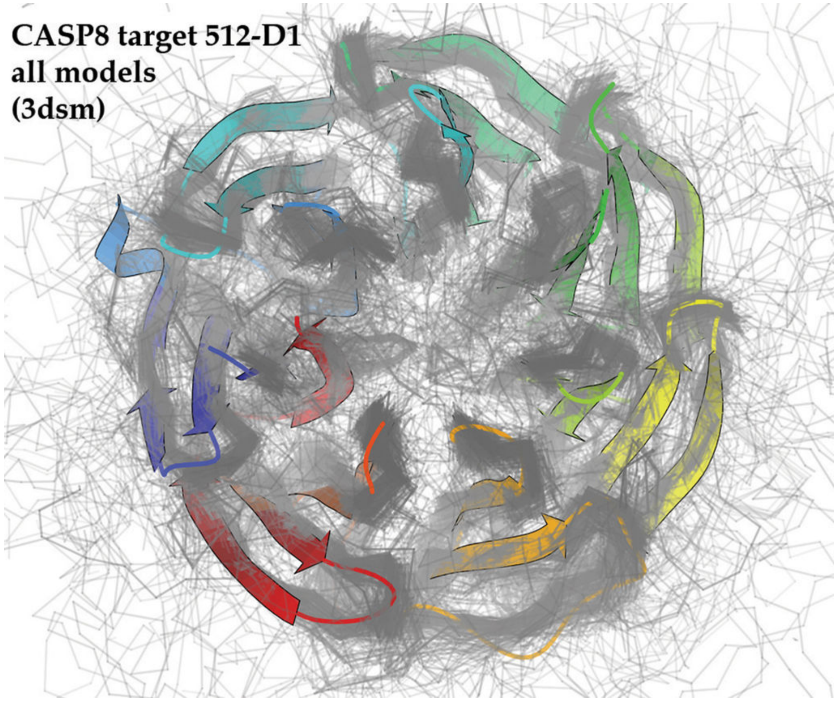
The development of AI approaches allow us training model to solve the problem directly.

Critical Assessment of protein Structure Prediction (CASP)

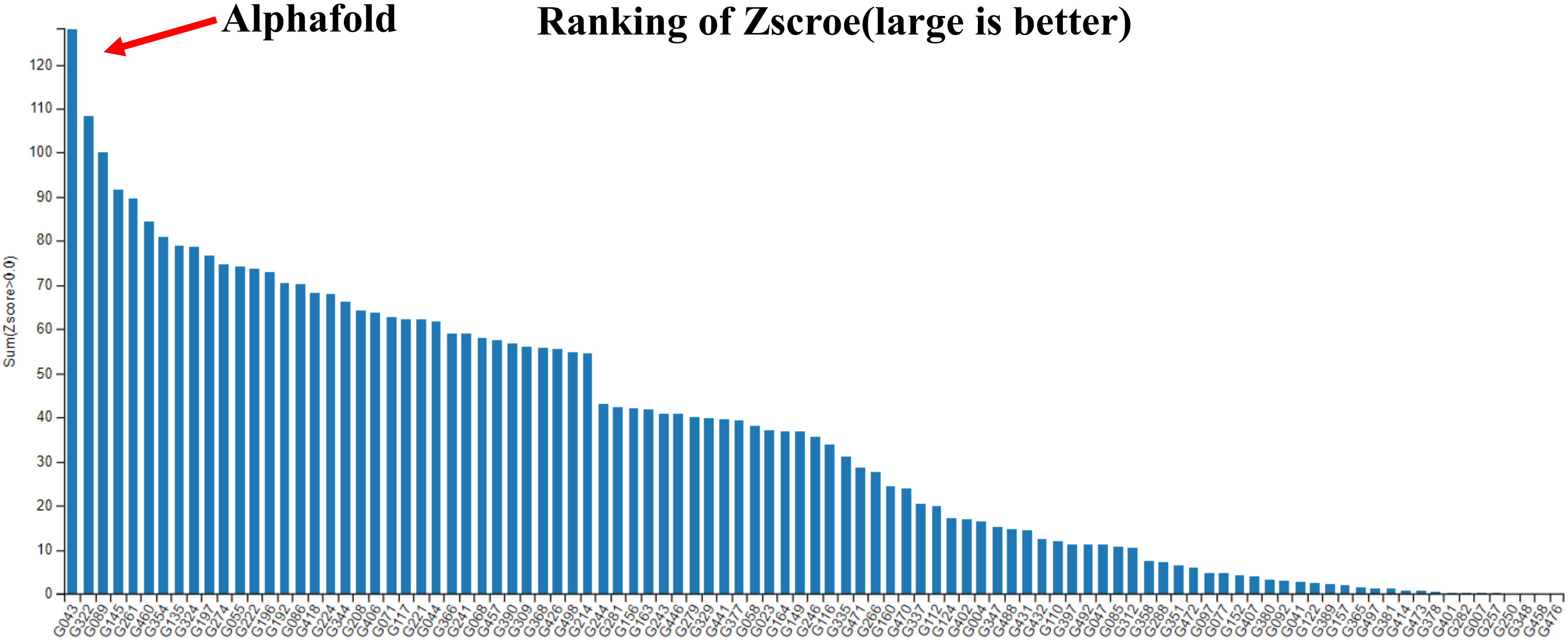
Taking place every two years since 1994

Neither predictors nor the organizers and assessors know the structures of the target proteins at the time when predictions are made.

Targets for structure prediction are either structures soon-to-be solved, or structures that have just been solved and are kept on hold by the Protein Data Bank.



AI Breakthrough



Alphafold

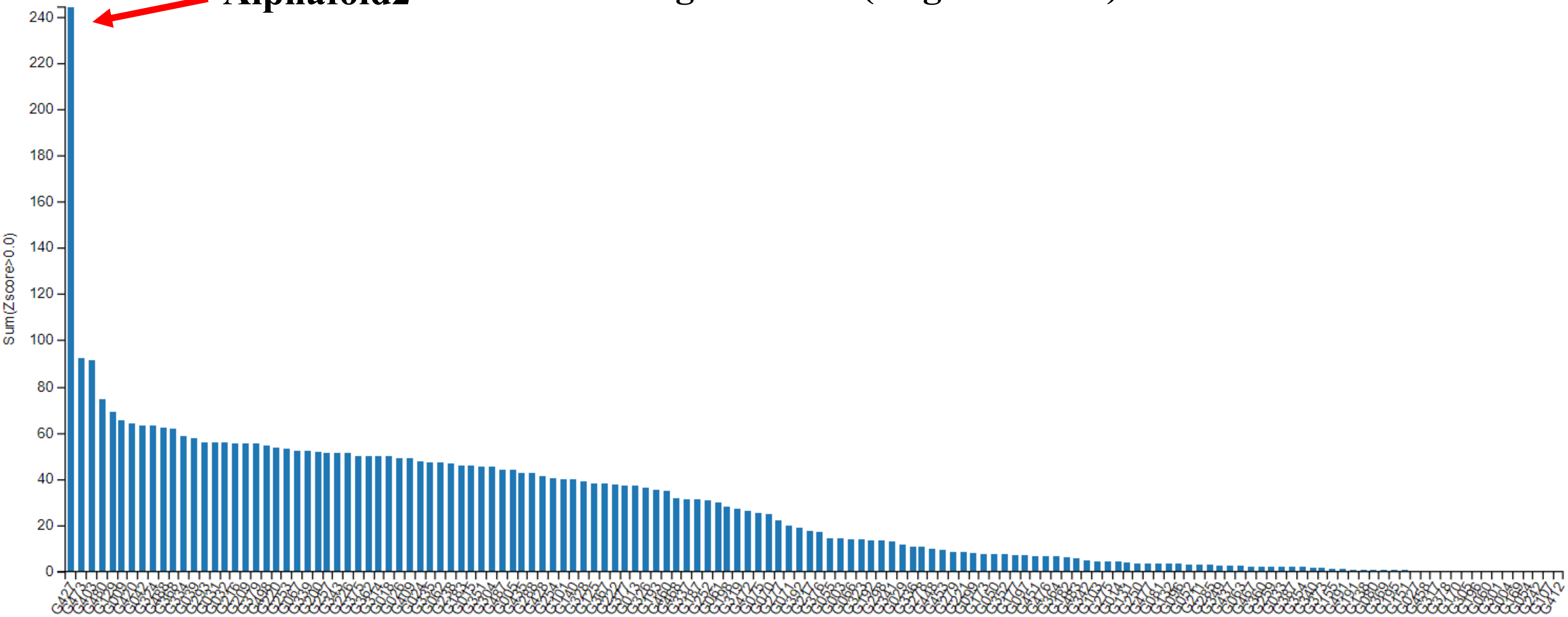
Ranking of Zscroe(large is better)

CASP13 in 2018

AI Breakthrough

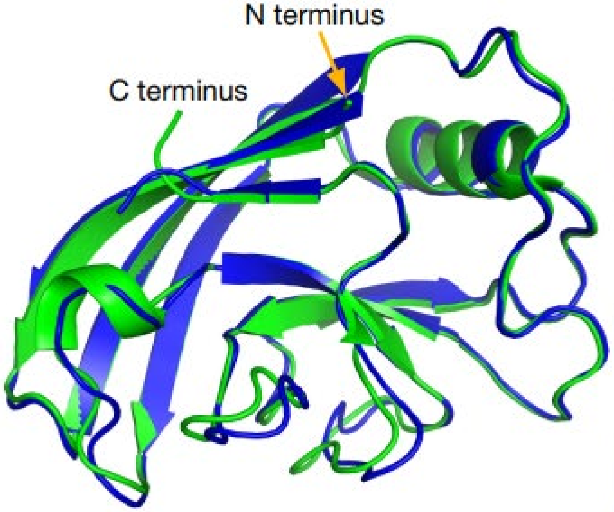
Alphafold2

Ranking of Zscroe(large is better)



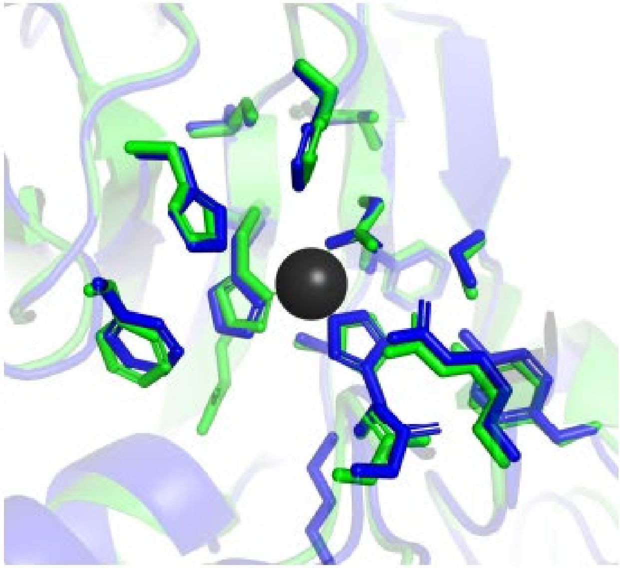
CASP14 in 2020

AIphafold2



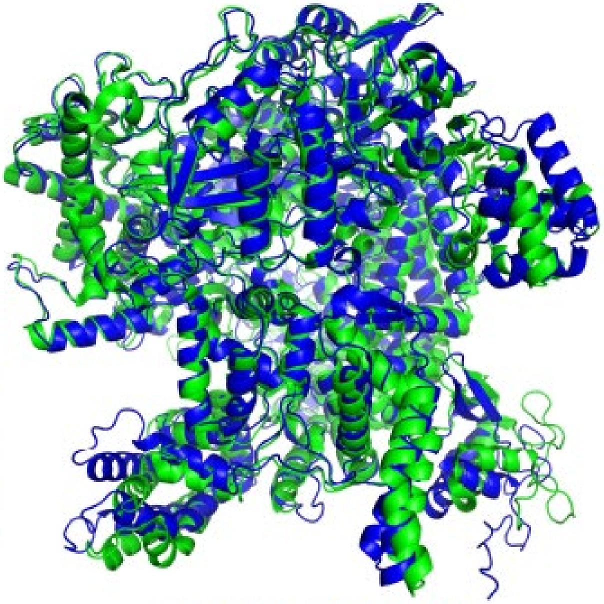
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

Backbone accuracy



r.m.s.d. = 0.59 Å within 8 Å of Zn

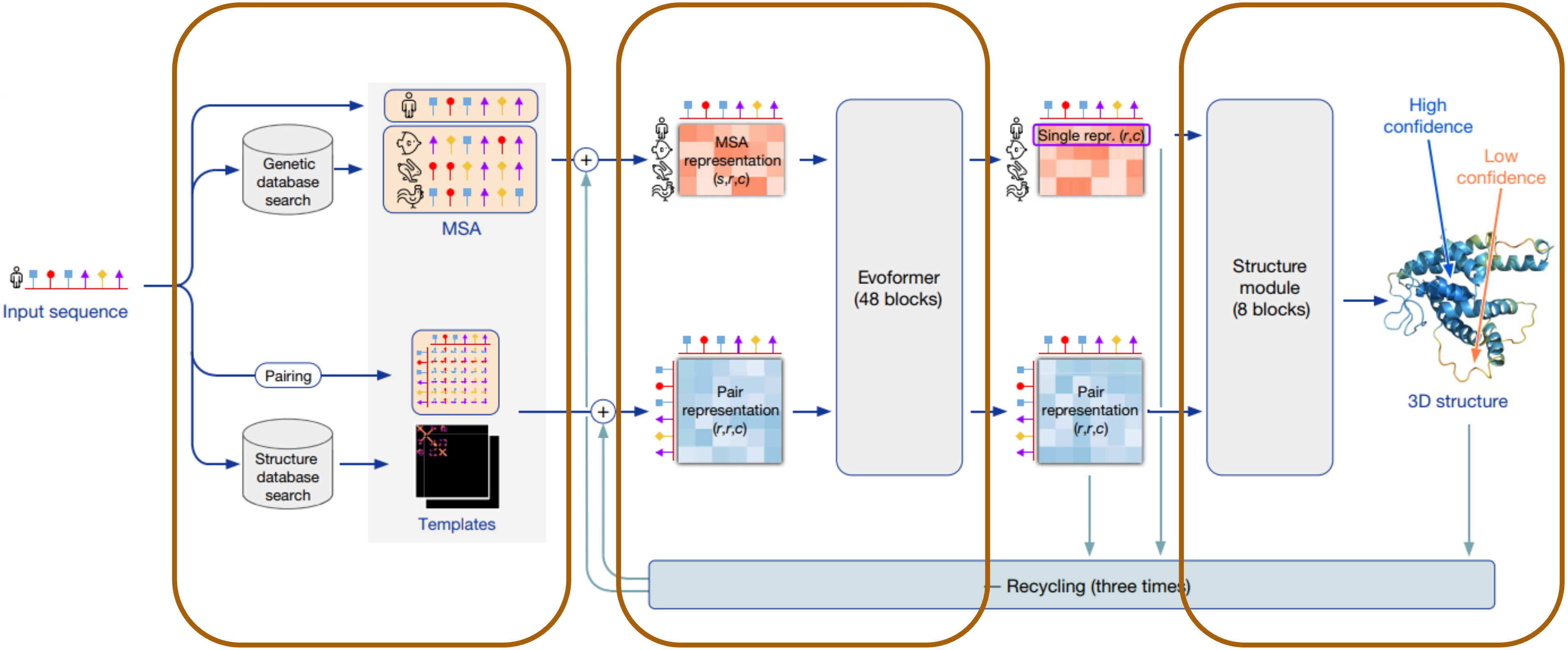
Side-chain accuracy



r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

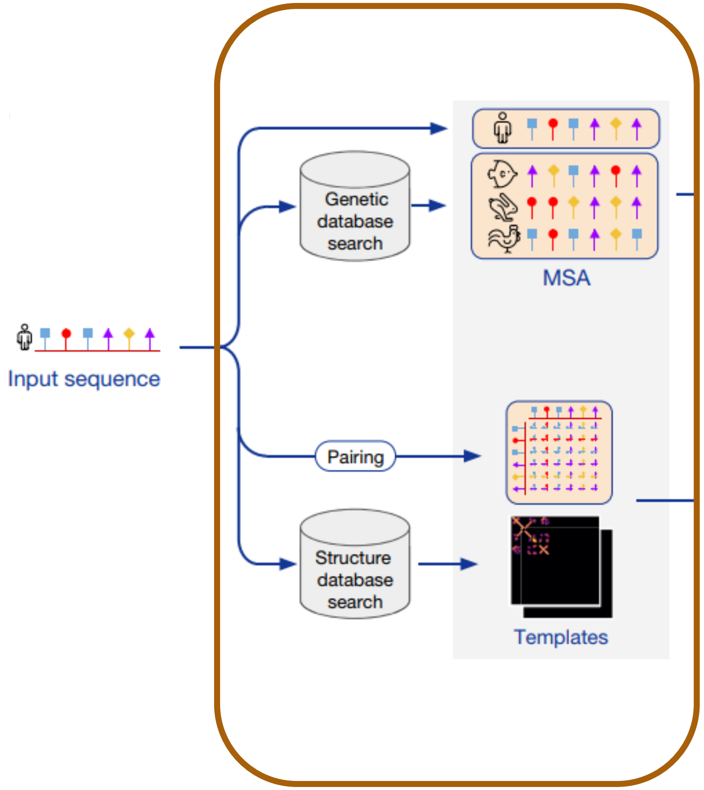
Domain packing accuracy

AIphafold2 Model Architecture

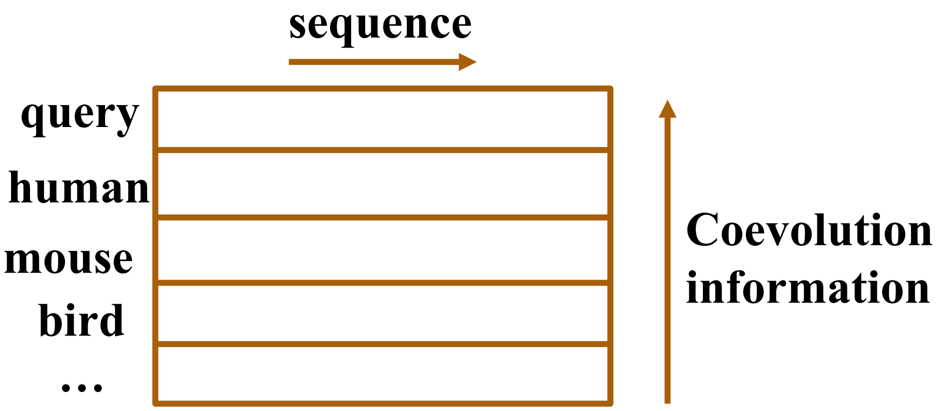


Data processing **Evoformer** **Structure module**

Data processing



Data processing


sequence

query

human

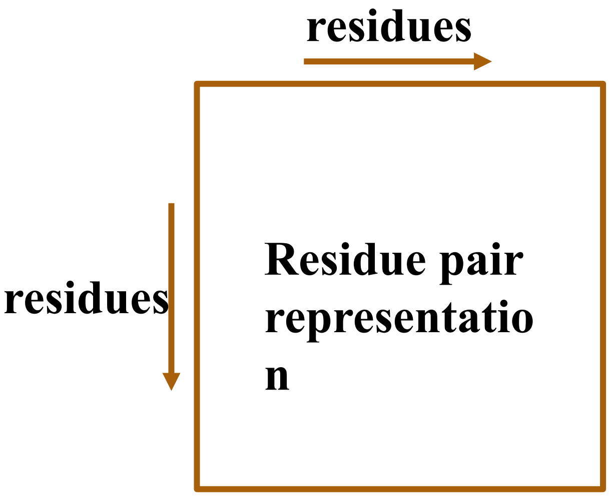
mouse

bird

…

Coevolution information

MSA (Multisequence alignment)


residues

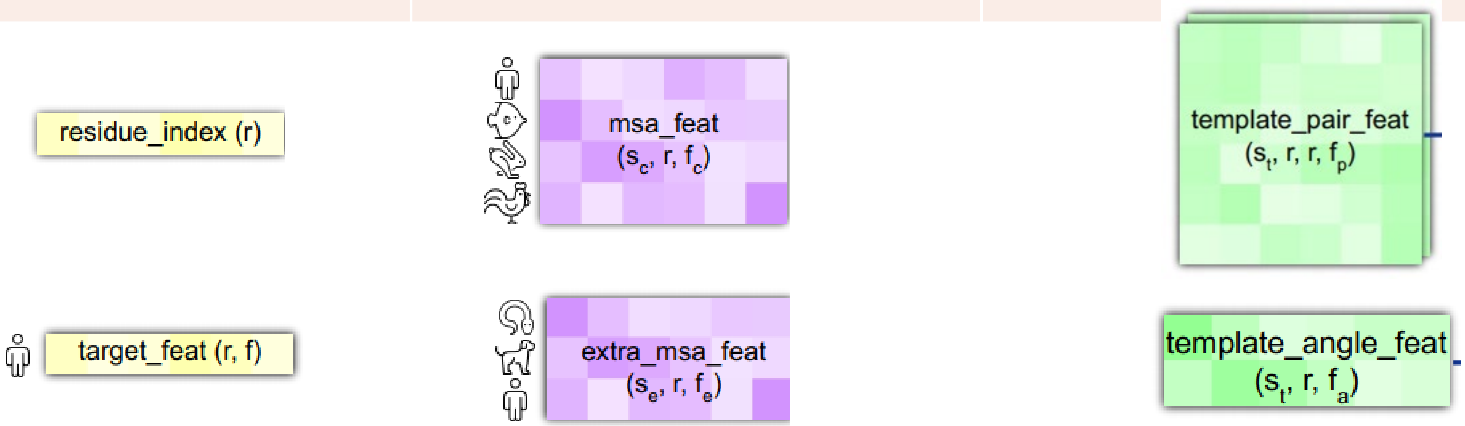
residues

Residue pair representation

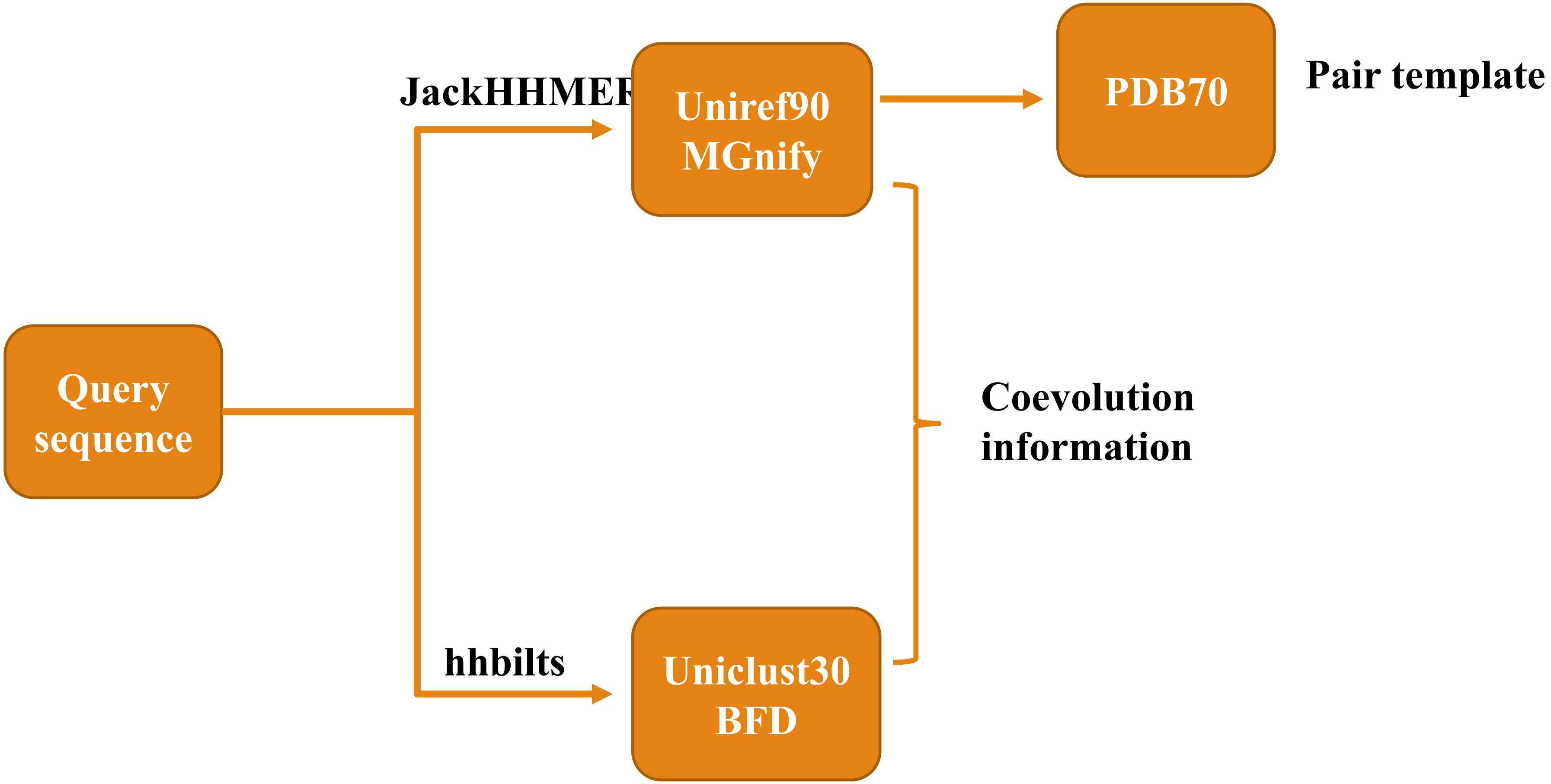
pair representation

Data processing

Feature	shape	description
residue_index	[Nres]	residue index
target_feat	[Nres,21]	residue index and aa type
msa_feat	[Nclust, Nres,49]	Profile of MSA cluster
extra_msa_feat	[Nextra,Nres, 25]	Profile of extra MSA
template_pair_feat	[Ntempl; Nres; Nres; 88]	Dis. aaType info of tpl
template_angle_feat	[Ntempl; Nres; 51]	Torsions, aaType info of tpl



Data processing



JackHHMER

Uniref90
MGnify

PDB70

Pair template

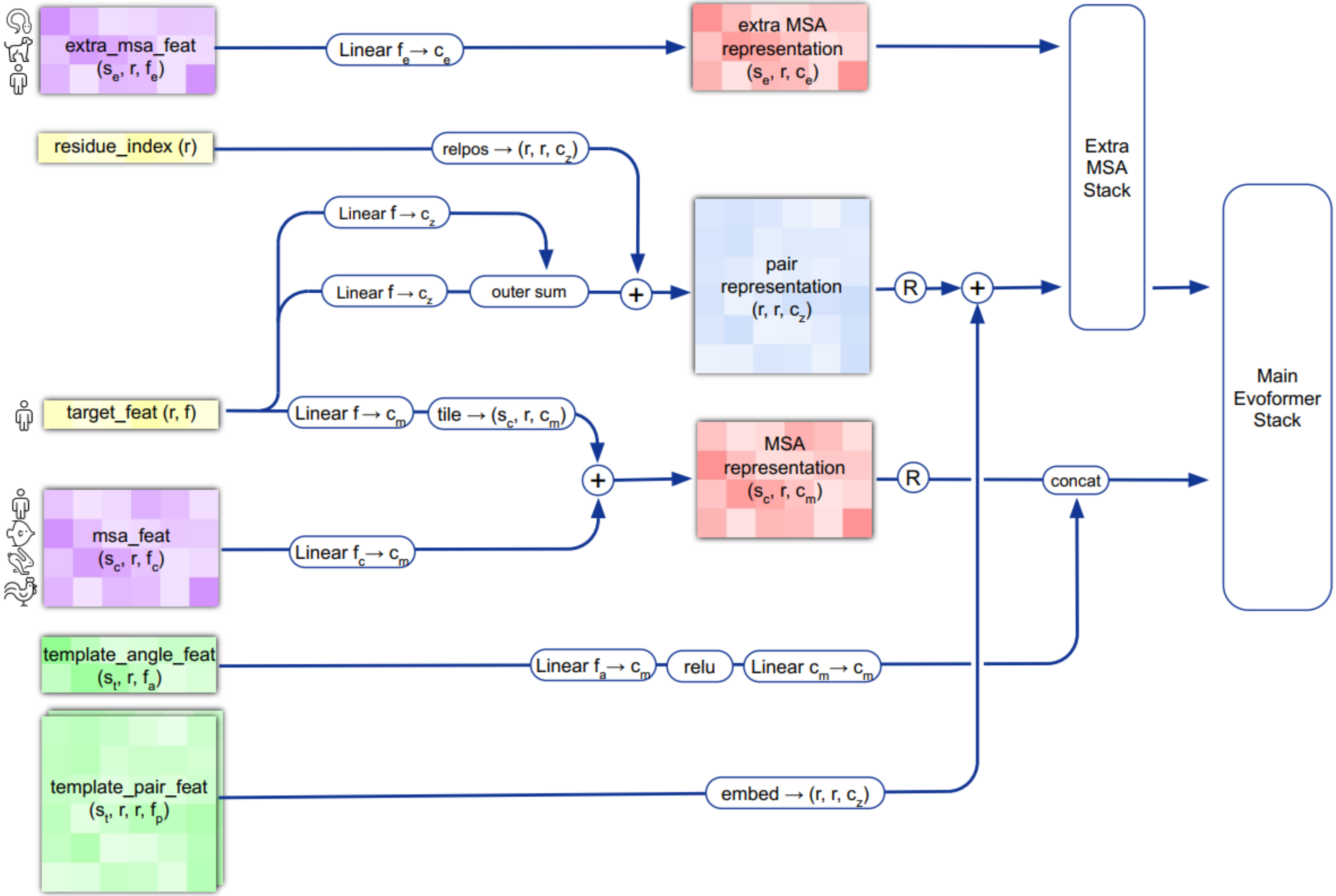
Query
sequence

Coevolution
information

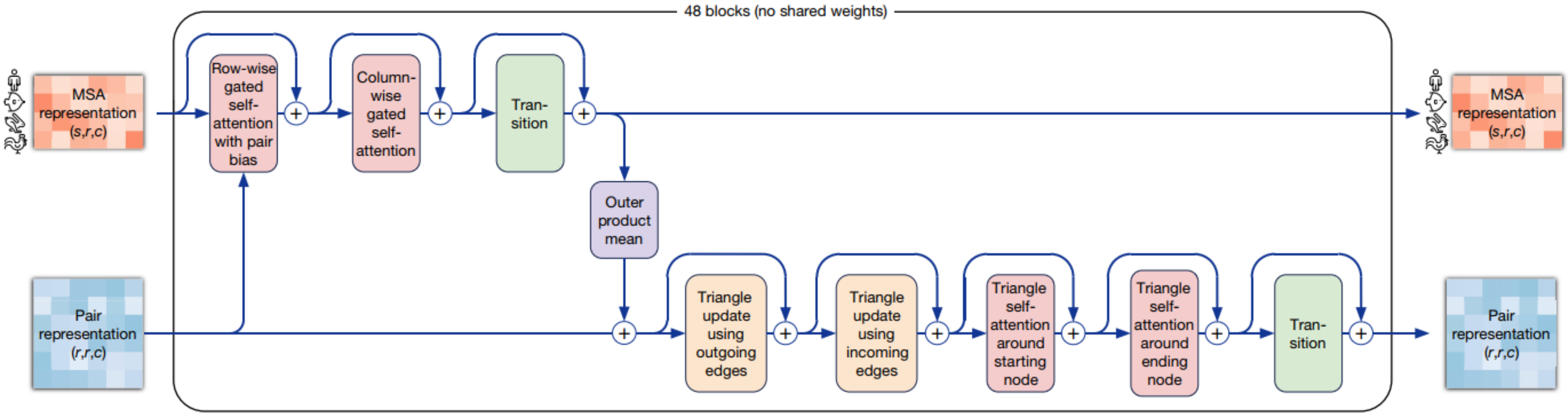
hhbilts

Uniclust30
BFD

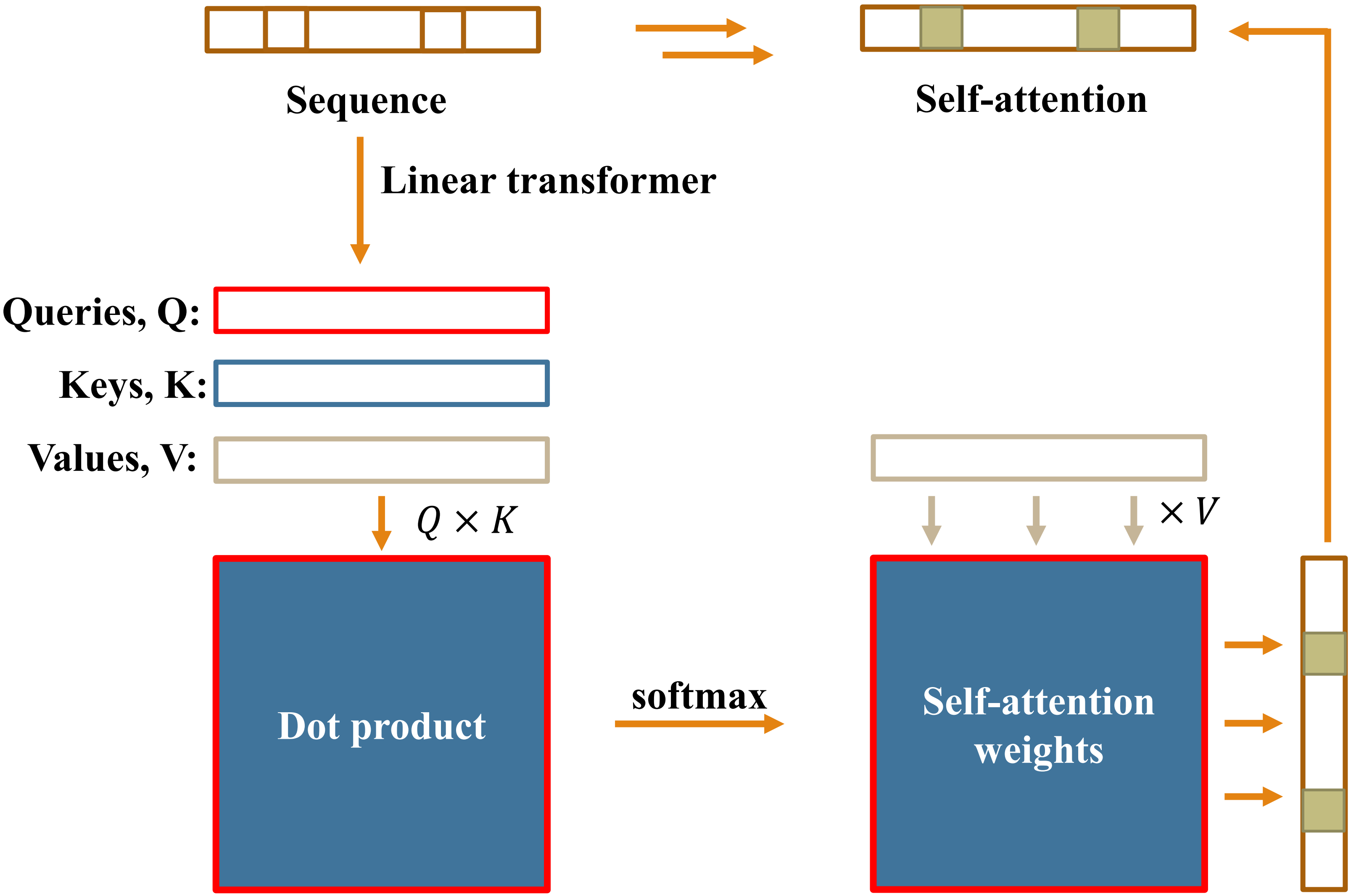
Data processing



Evoformer


48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
MSA representation (s,r,c)
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
Pair representation (r,r,c)

Self-attention



Sequence

Self-attention

Linear transformer

Queries, Q:

Keys, K:

Values, V:

$Q \times K$

$\times V$

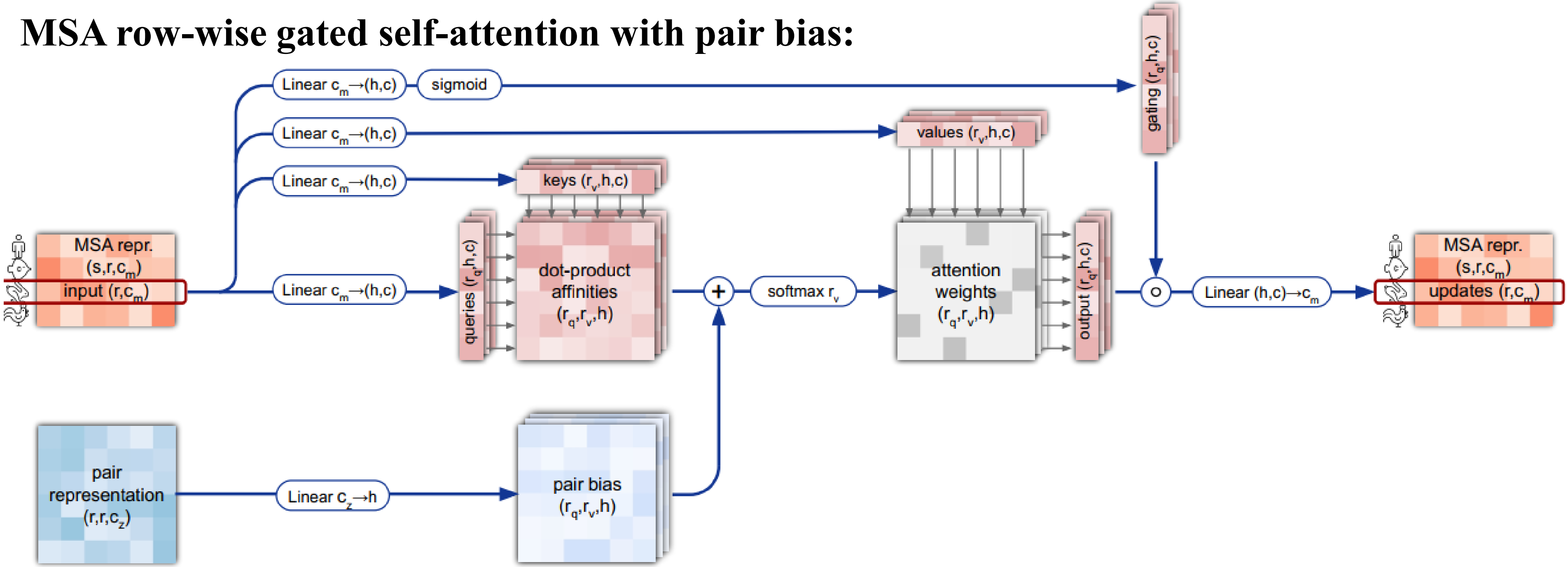
Dot product

softmax

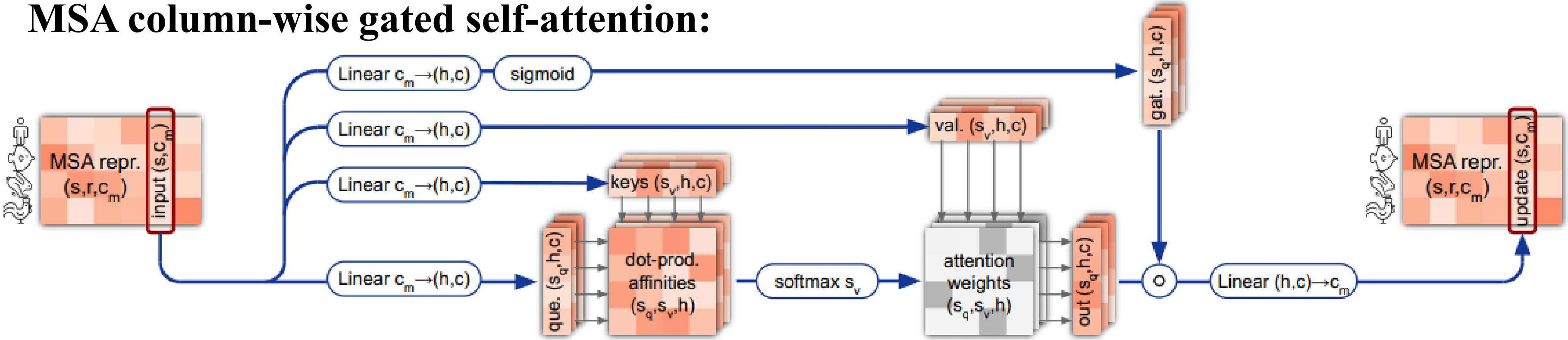
Self-attention weights

Evoformer

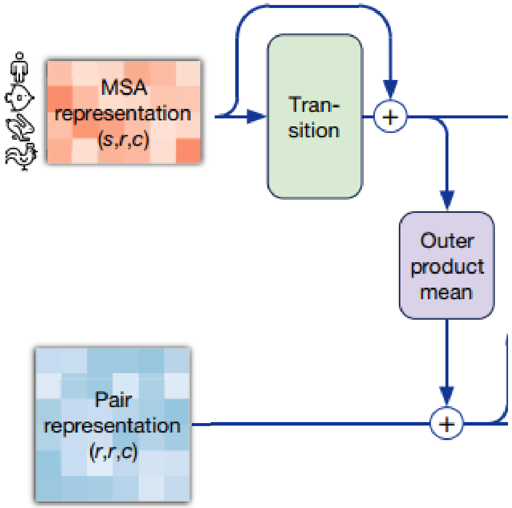
MSA row-wise gated self-attention with pair bias:



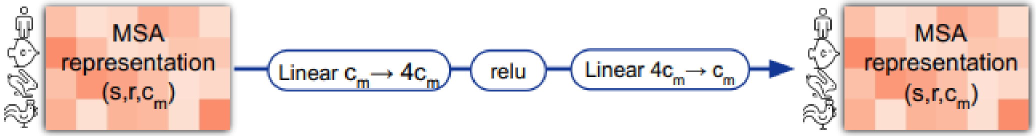
MSA column-wise gated self-attention:



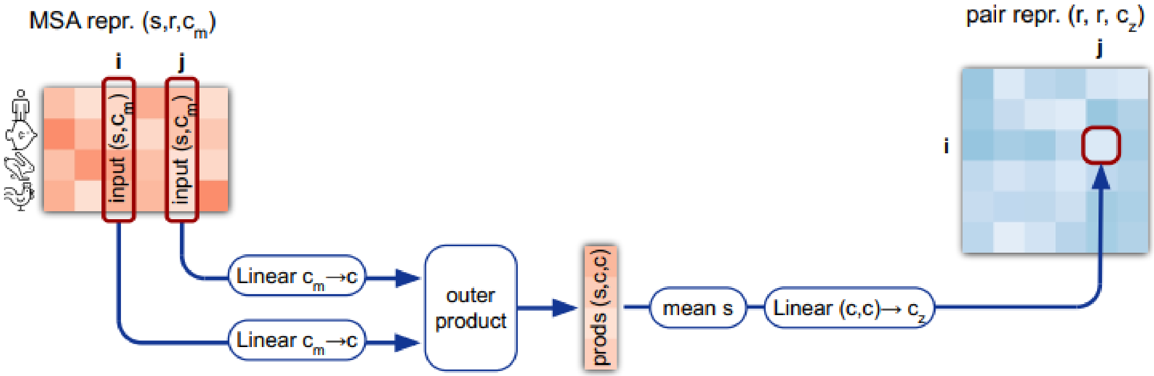
Information exchange



MSA transition layer:

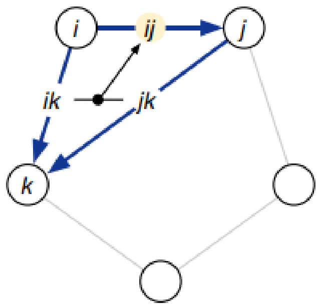


Outer product mean:

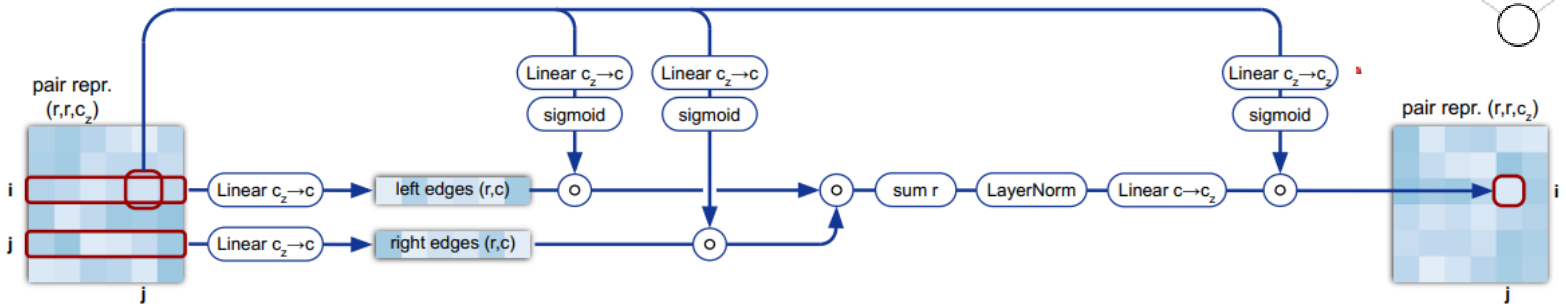


Edges update

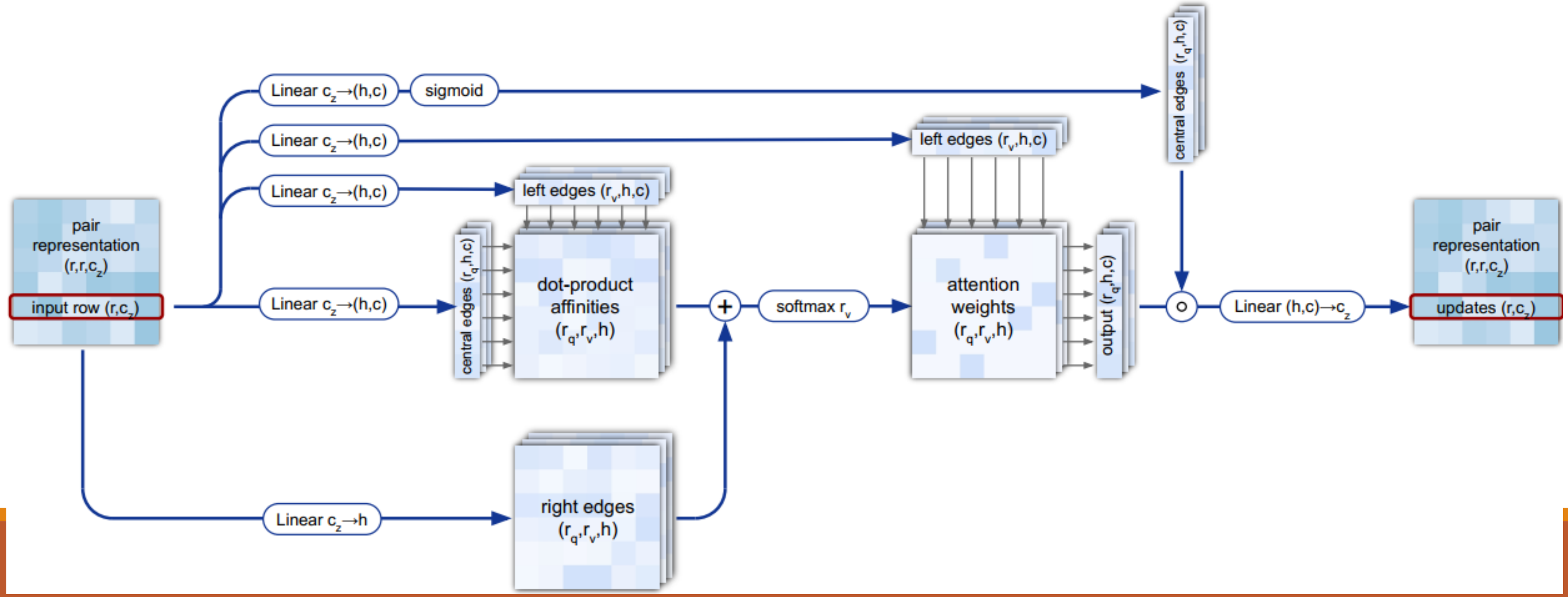
Triangle multiplicative update using ‘outgoing’ edges



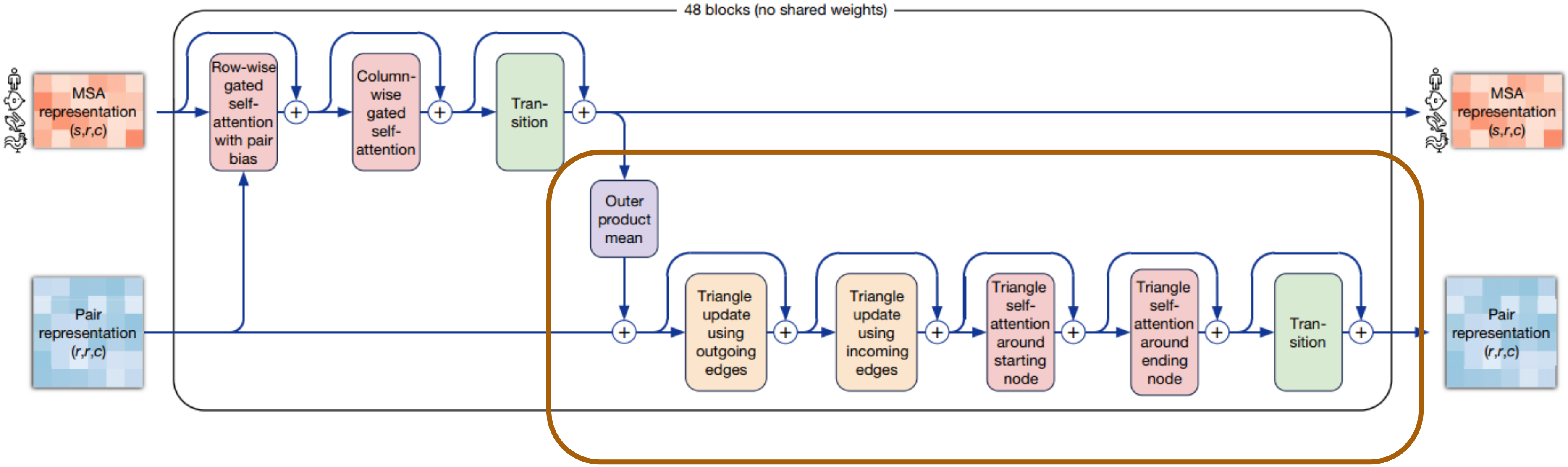
Triangular multiplicative update using “outgoing” edges:



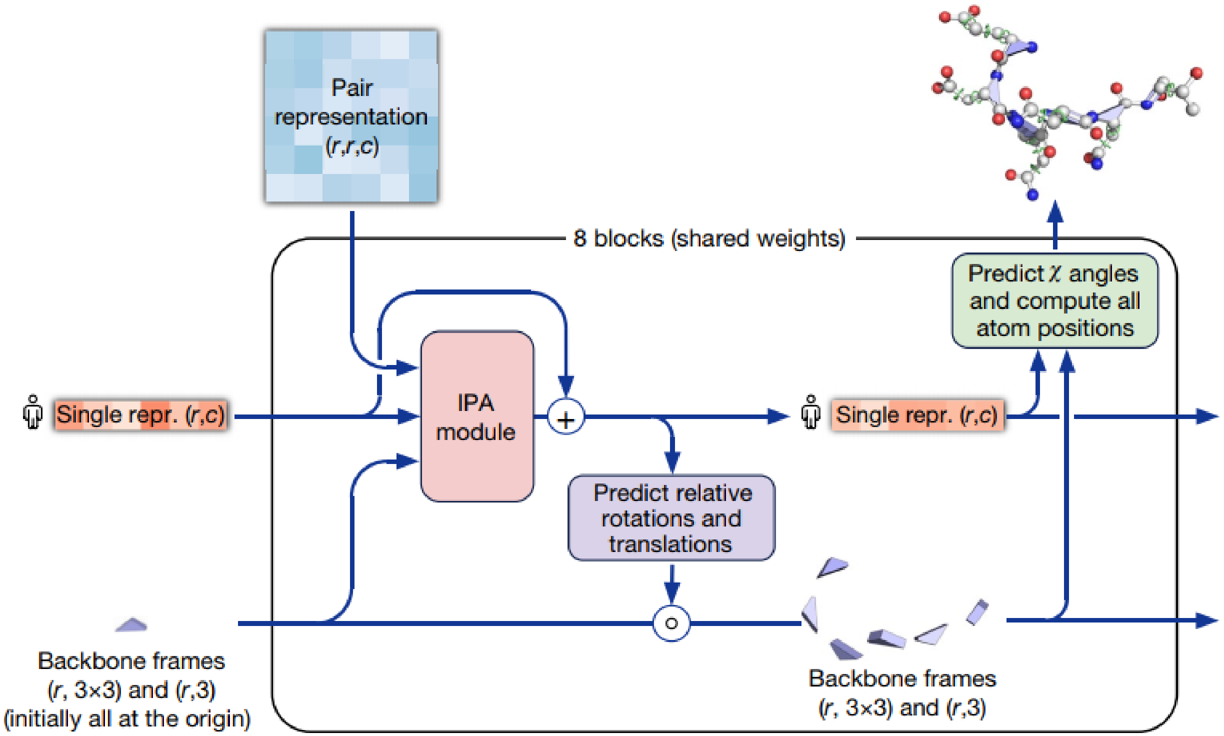
Triangular self-attention around starting node:



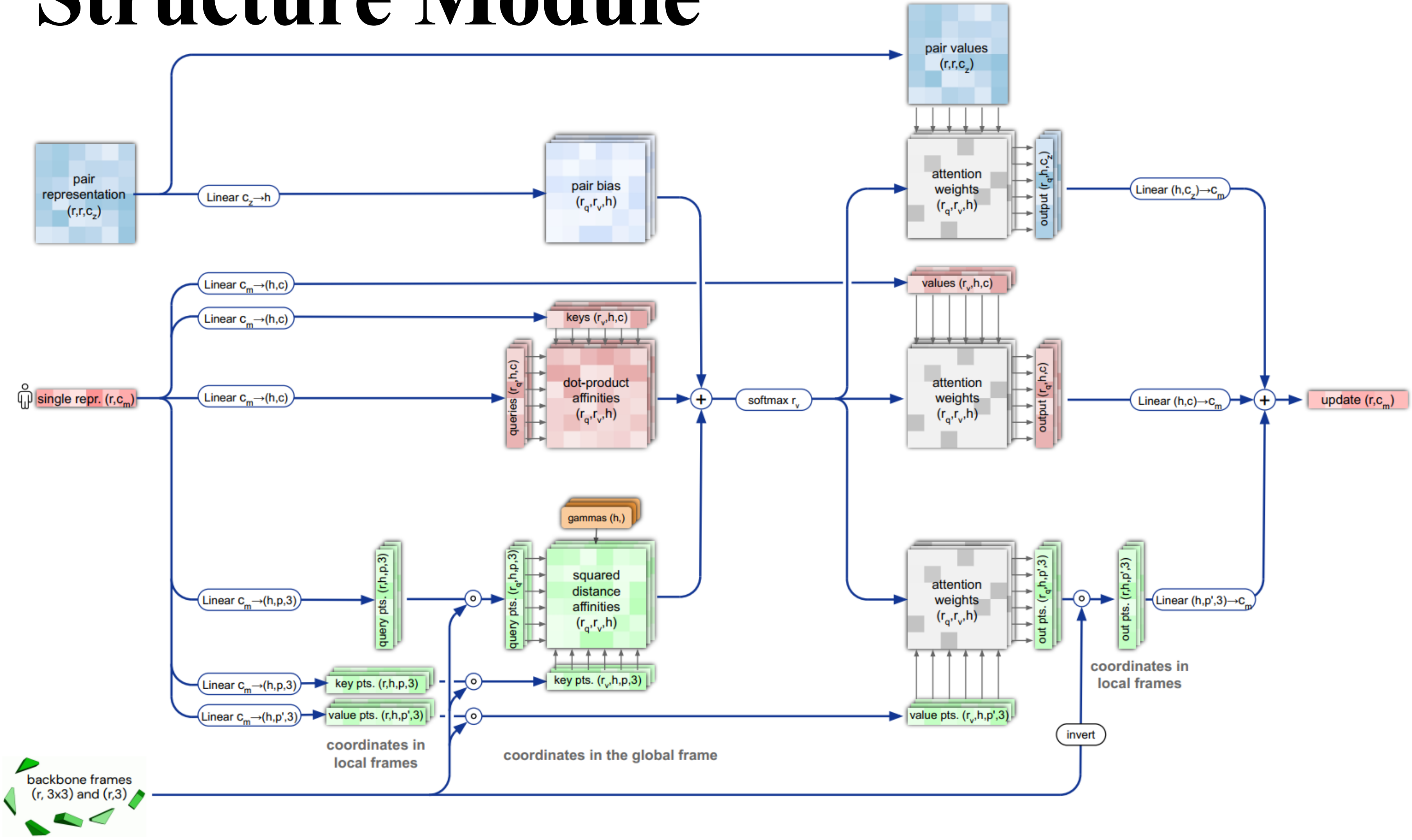
Edges update



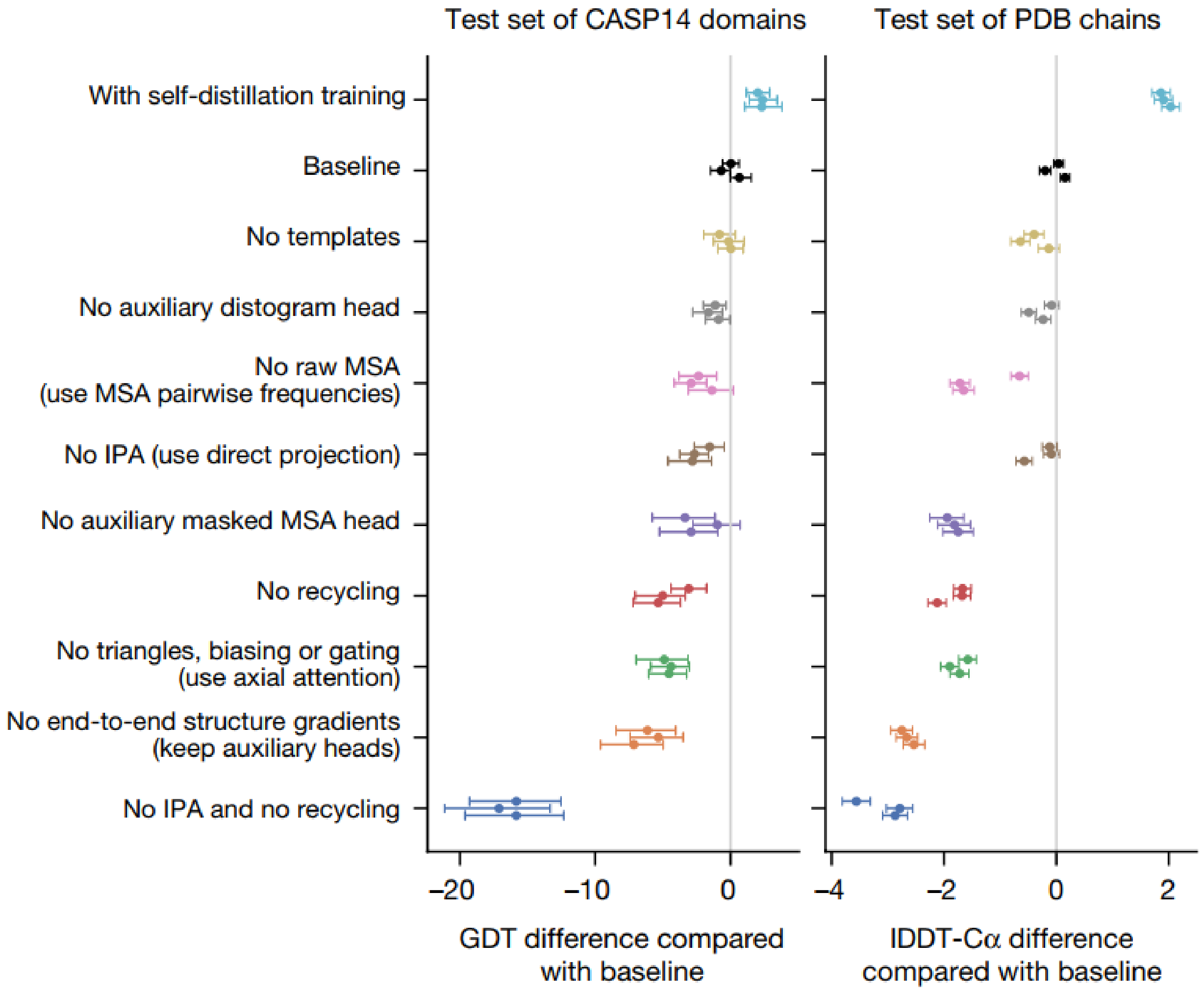
Structure Module



Structure Module

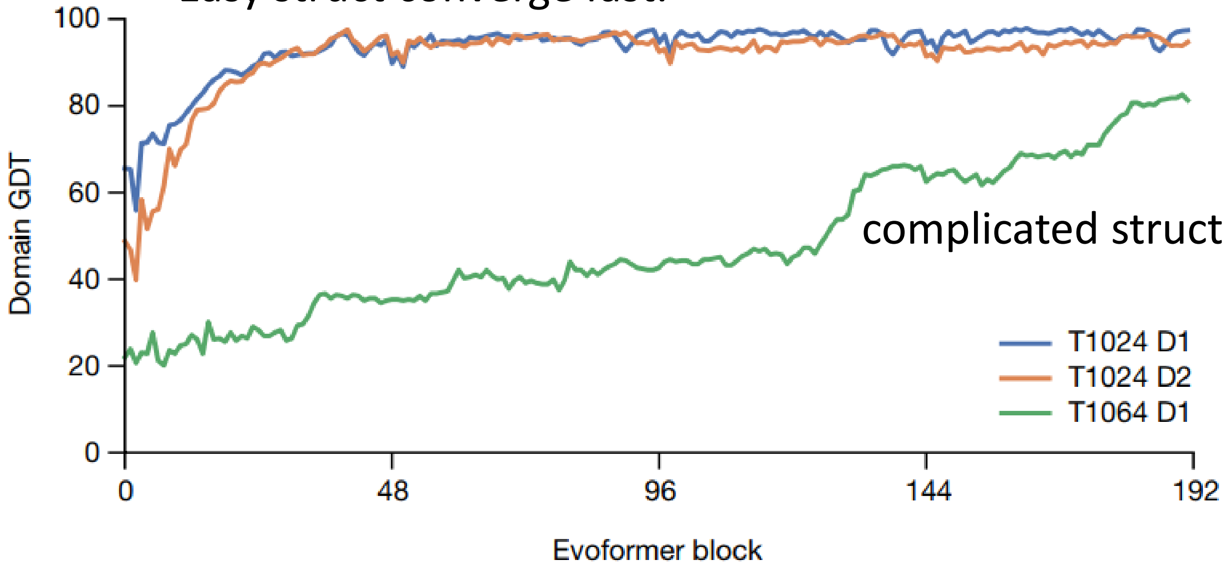


Interpreting the neural network



The contribution to accuracy of components of AlphaFold

Interpreting the neural network

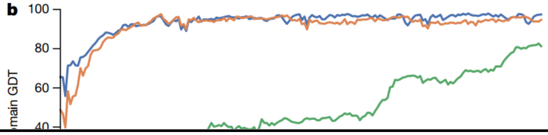


Easy struct converge fast.

complicated struct converge slow.

GDT score of intermediates alone with network depth.

T1024



T1064