D-I-TASSER: Integrating Deep Learning with multi-MSAs and threading alignments for protein structure prediction

(Groups “UM-TBM” and “Zheng”)

Wei Zheng[1,2], Qiqige Wuyun[3], and Peter L. Freddolino[1,2]

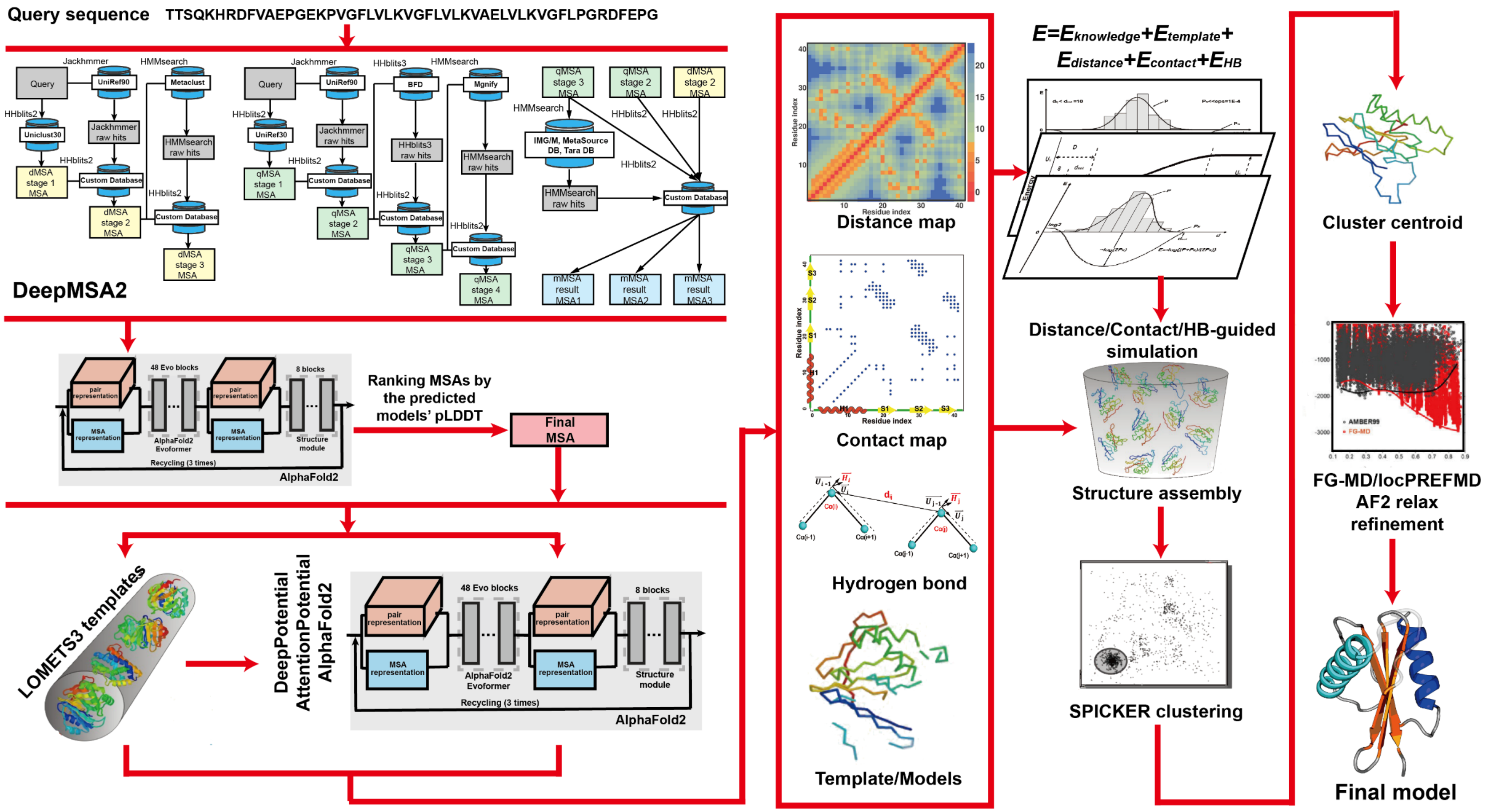
[1]Department of Computational Medicine and Bioinformatics , University of Michigan

[2]Department of Biological Chemistry, University of Michigan

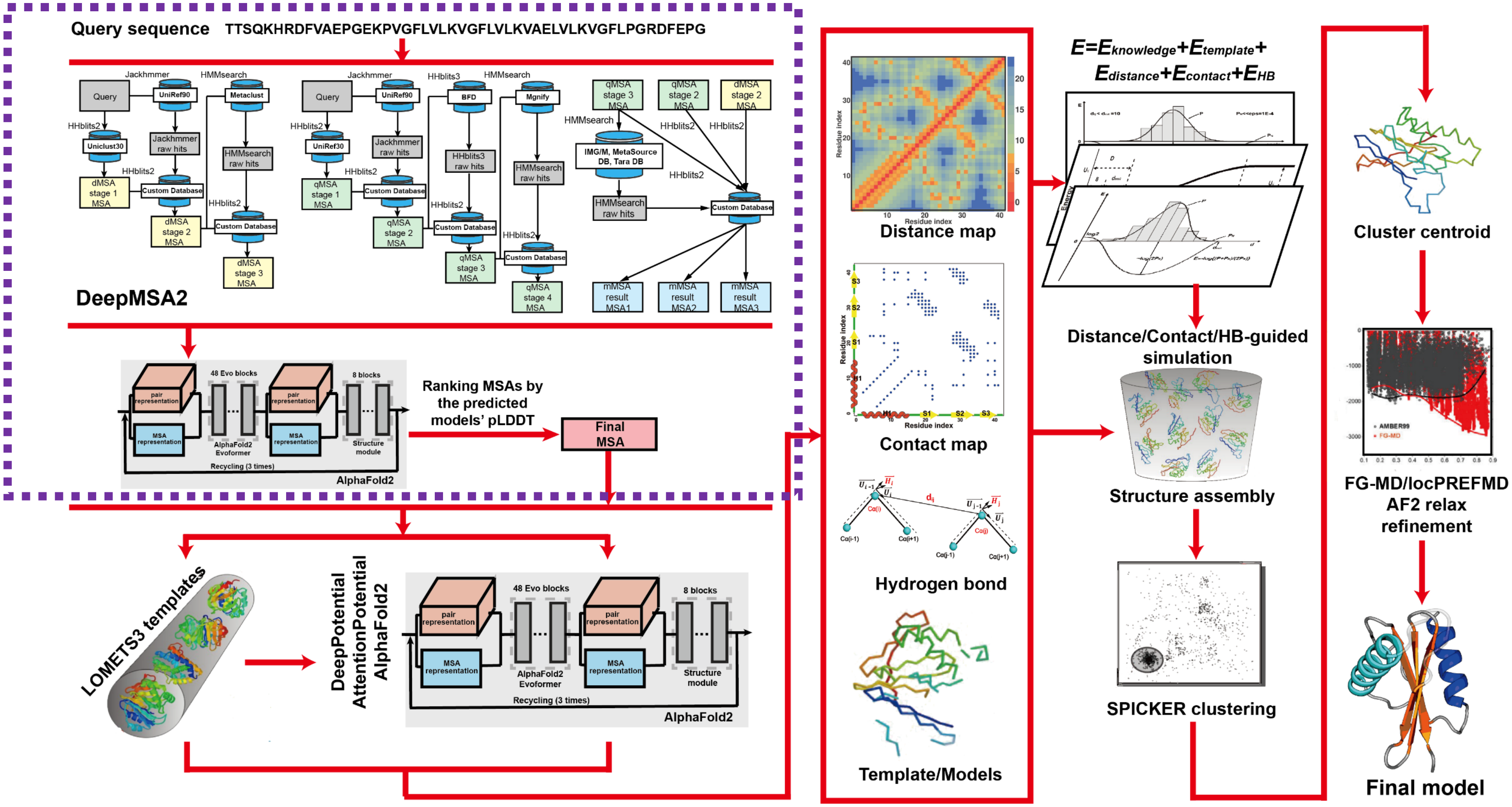
[3]Department of Computer Science and Engineering, Michigan State University

Methods

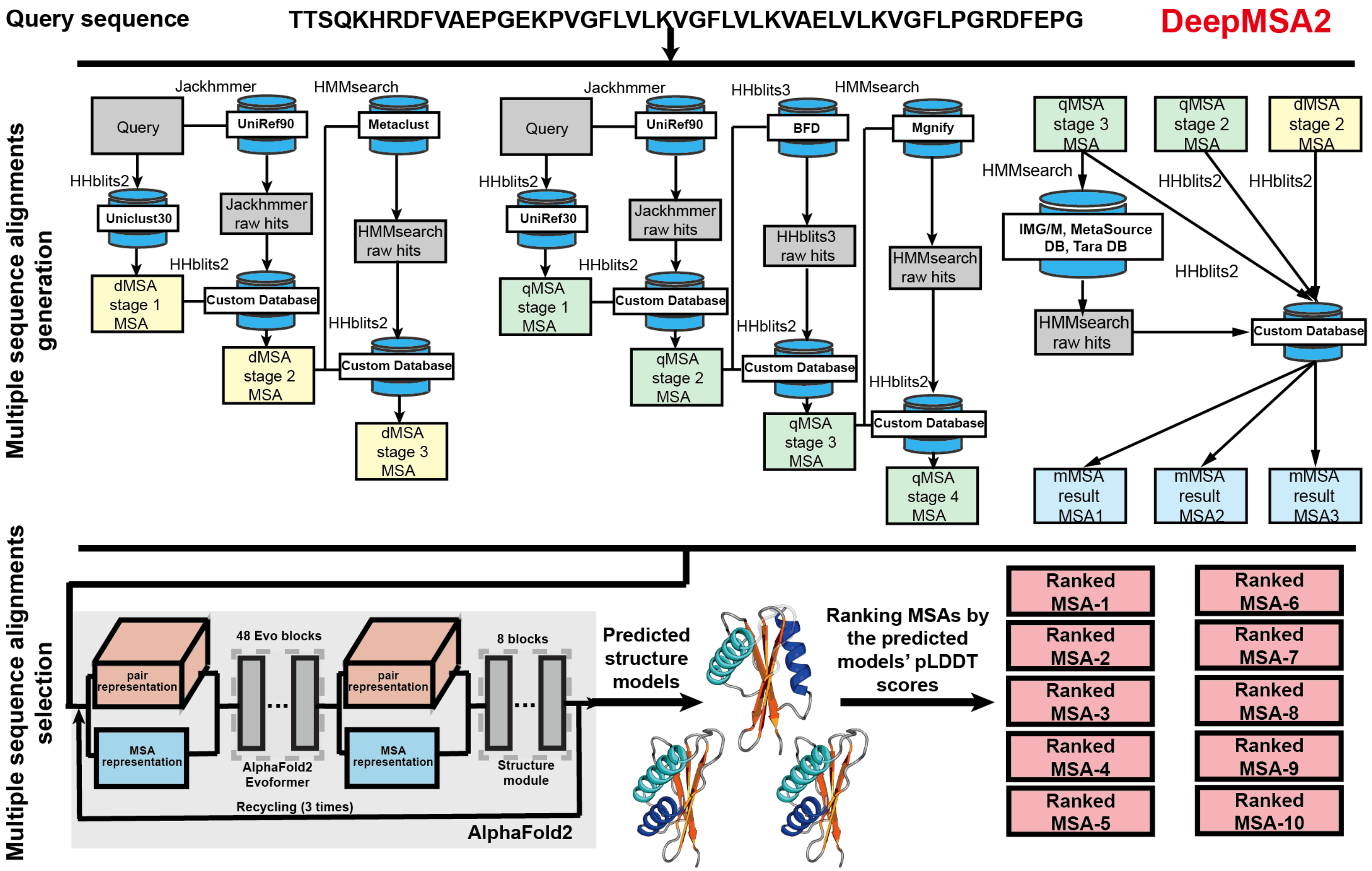
UM-TBM server built from D-I-TASSER for single-chain protein modeling



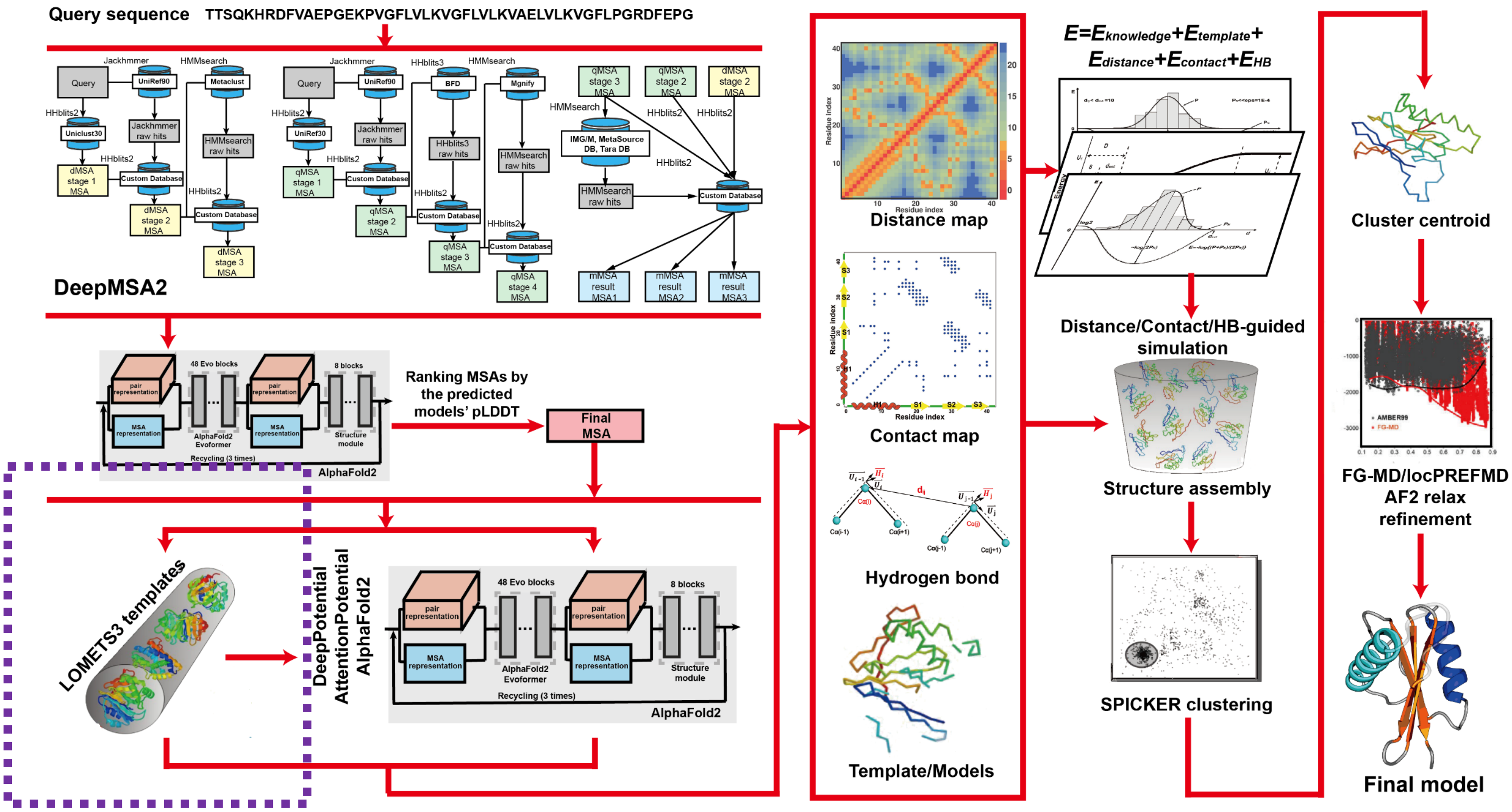
UM-TBM server built from D-I-TASSER for single-chain protein modeling



DeepMSA2 for monomeric MSA construction



UM-TBM server built from D-I-TASSER for single-chain protein modeling



Query sequence TTSQKHRDFVAEPGEKPVGFLVLKVGFLVLKVAELVLKVGFLPGRDFEPG

DeepMSA2

Ranking MSAs by the predicted models' pLDDT

Final MSA

LOMETS3 templates

DeepPotential

AttentionPotential

AlphaFold2

$E = E_{knowledge} + E_{template} + E_{distance} + E_{contact} + E_{HB}$

Distance map

Contact map

Hydrogen bond

Template/Models

Distance/Contact/HB-guided simulation

Structure assembly

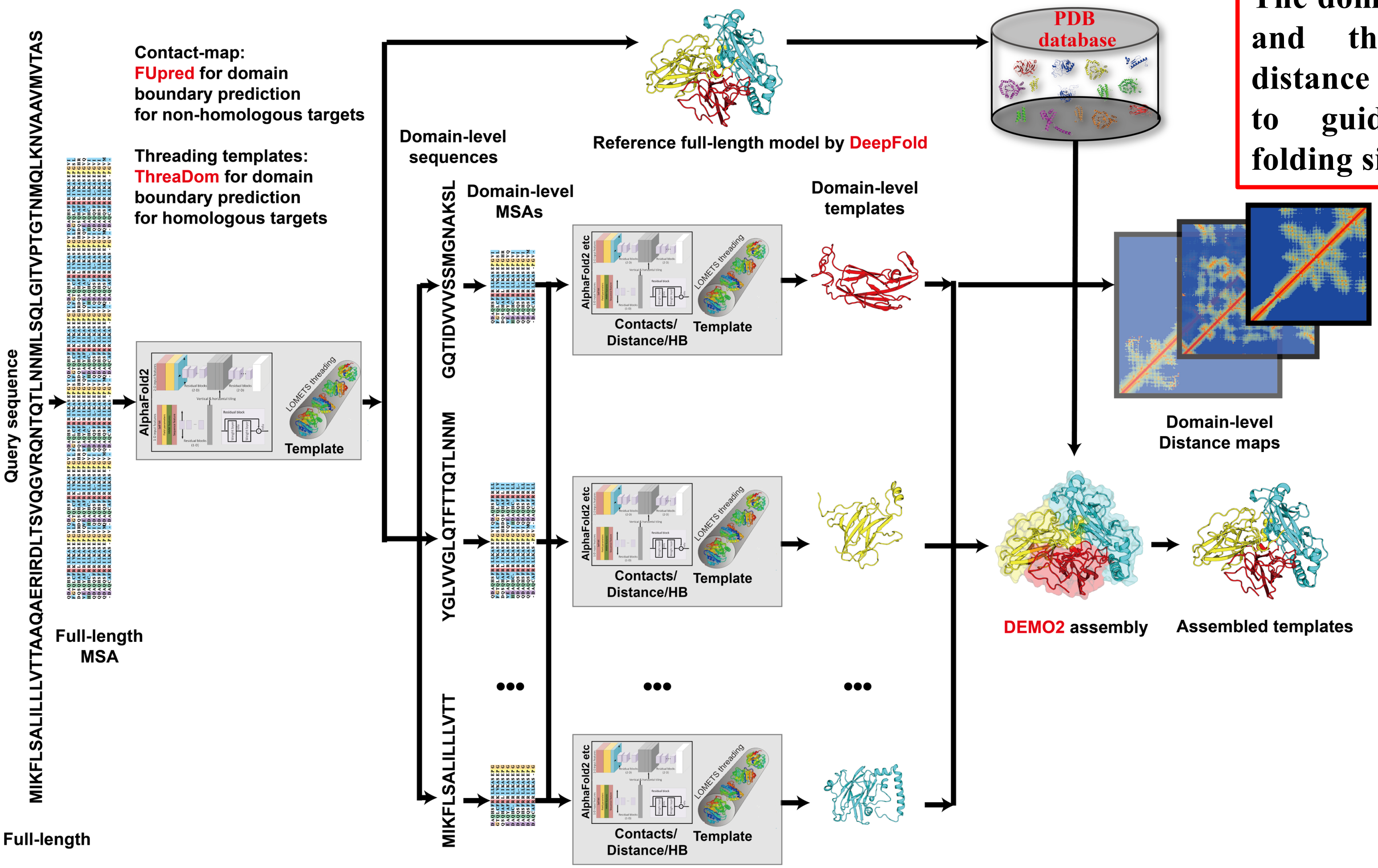
SPICKER clustering

Cluster centroid

FG-MD/locPREFMD AF2 relax refinement

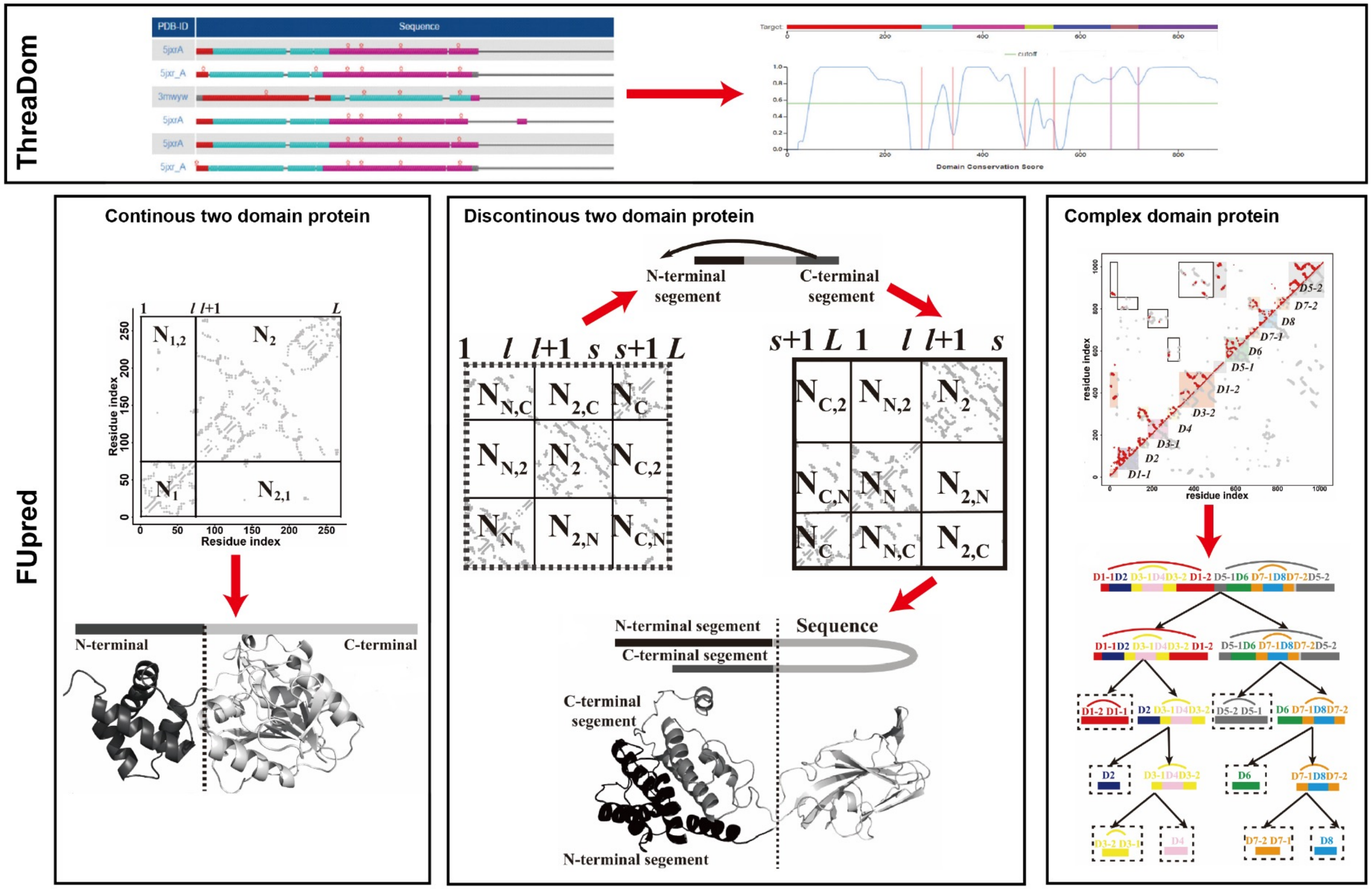
Final model

LOMETS3 for threading template detection

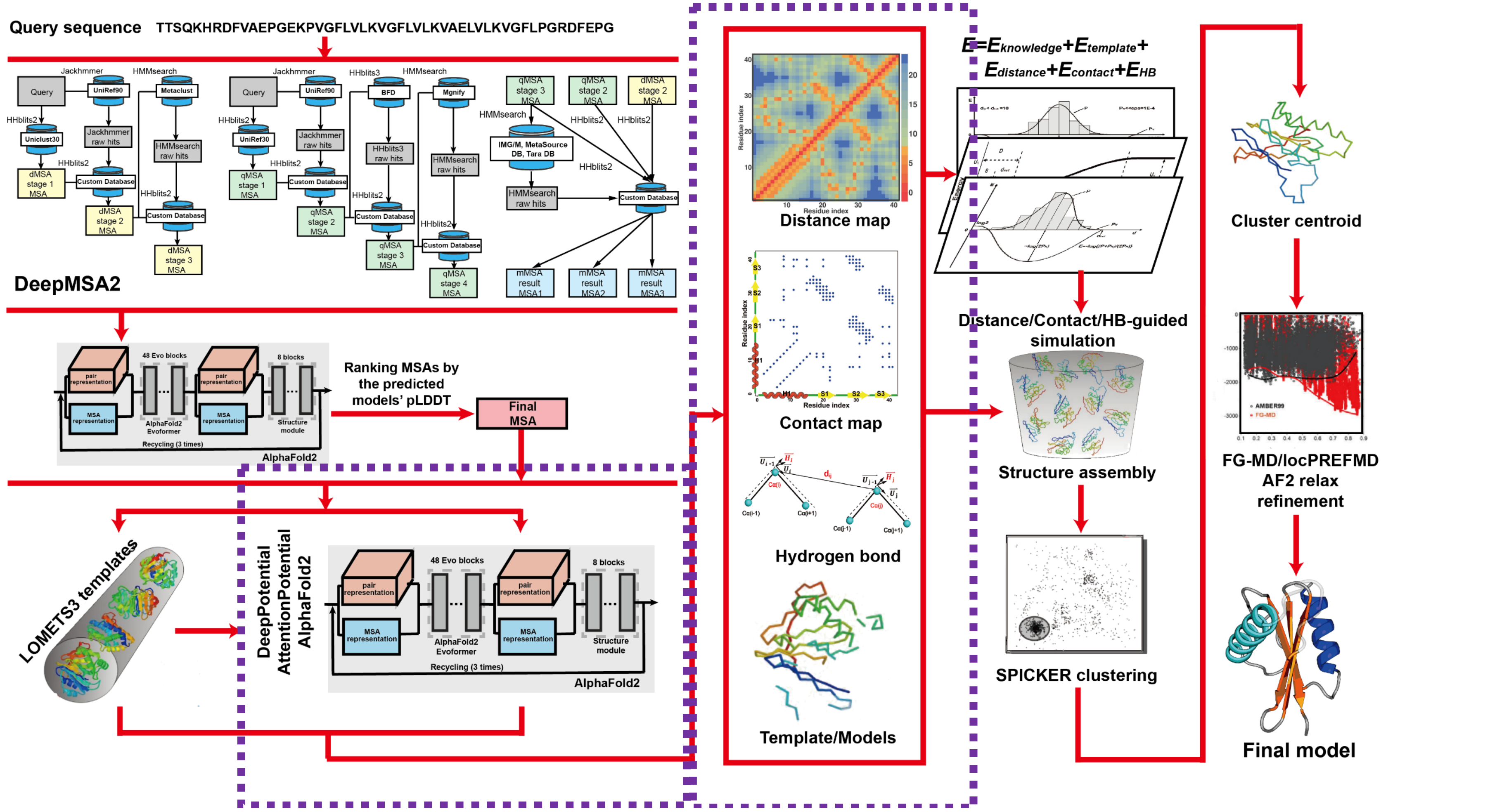


The domain-level distance maps and the whole chain-level distance map will be combined to guide the D-I-TASSER folding simulation

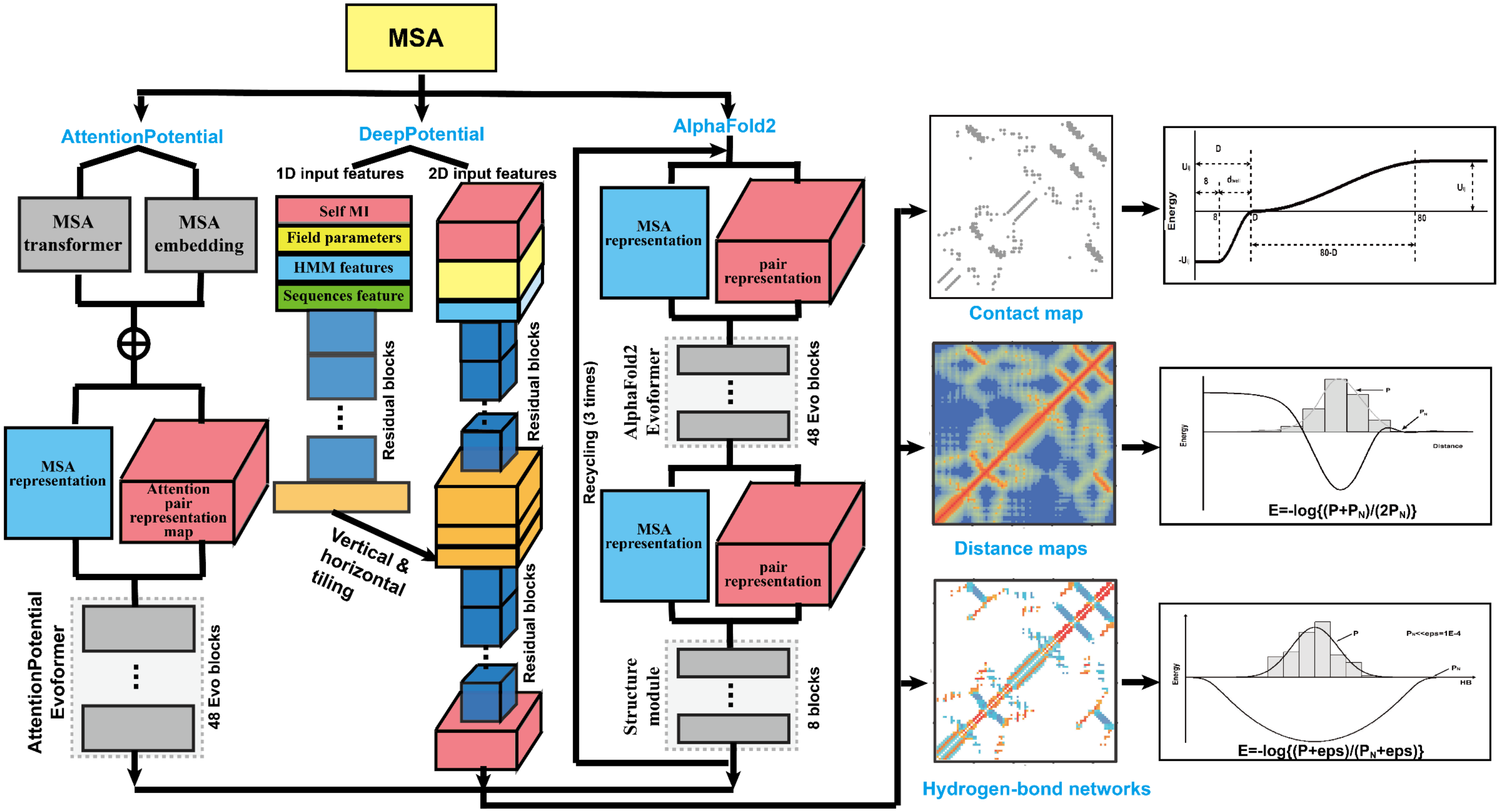
Domain boundary prediction in UM-TBM server



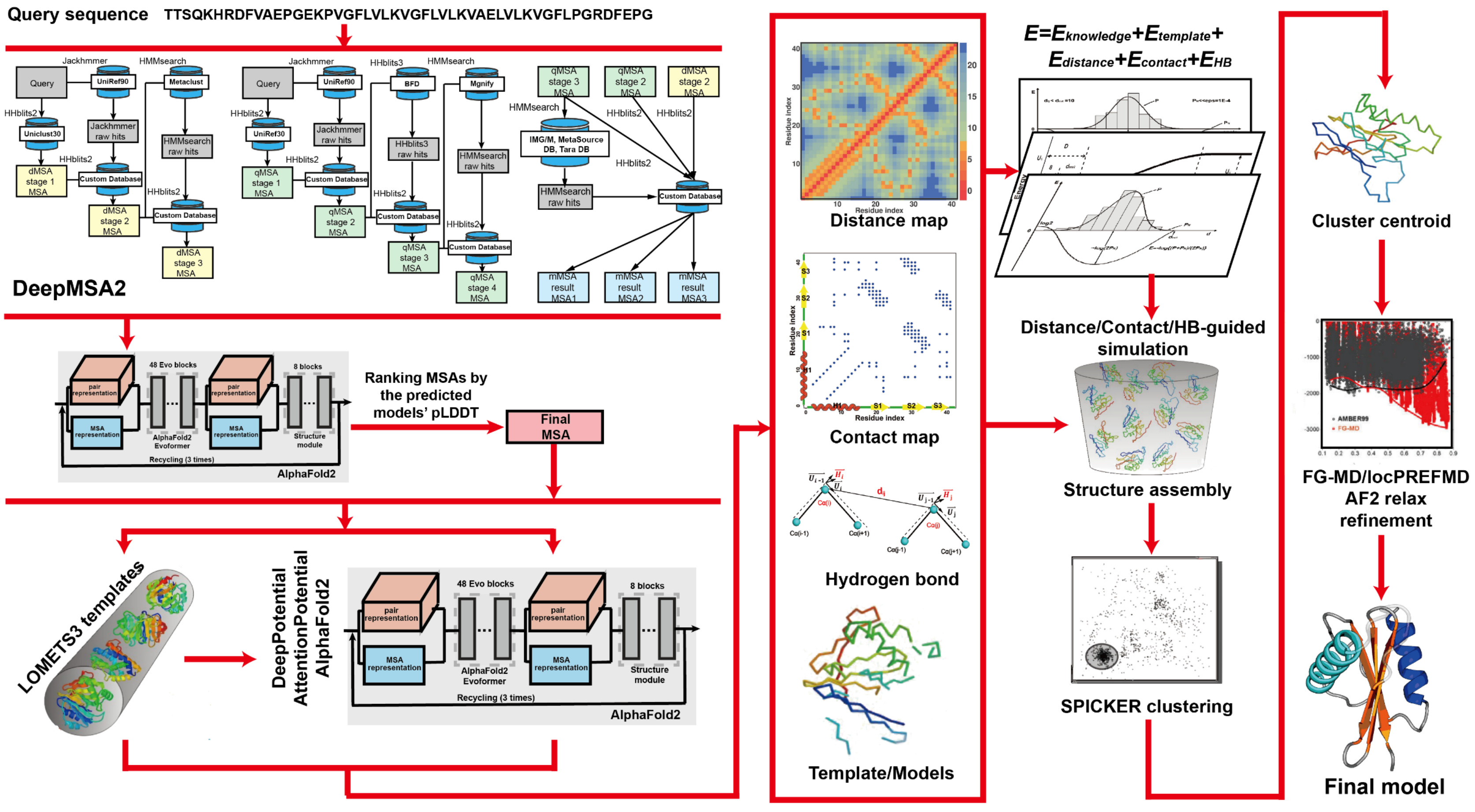
UM-TBM server built from D-I-TASSER for single-chain protein modeling



Deep learning-based spatial restraints prediction



UM-TBM server built from D-I-TASSER for single-chain protein modeling



Query sequence TTSQKHRDFVAEPGEKPVGFLVLKVGFLVLKVAELVLKVGFLPGRDFEPG

DeepMSA2

Ranking MSAs by the predicted models' pLDDT

Final MSA

LOMETS3 templates

DeepPotential AttentionPotential AlphaFold2

$E=E_{knowledge}+E_{template}+E_{distance}+E_{contact}+E_{HB}$

Distance map

Contact map

Hydrogen bond

Template/Models

Distance/Contact/HB-guided simulation

Structure assembly

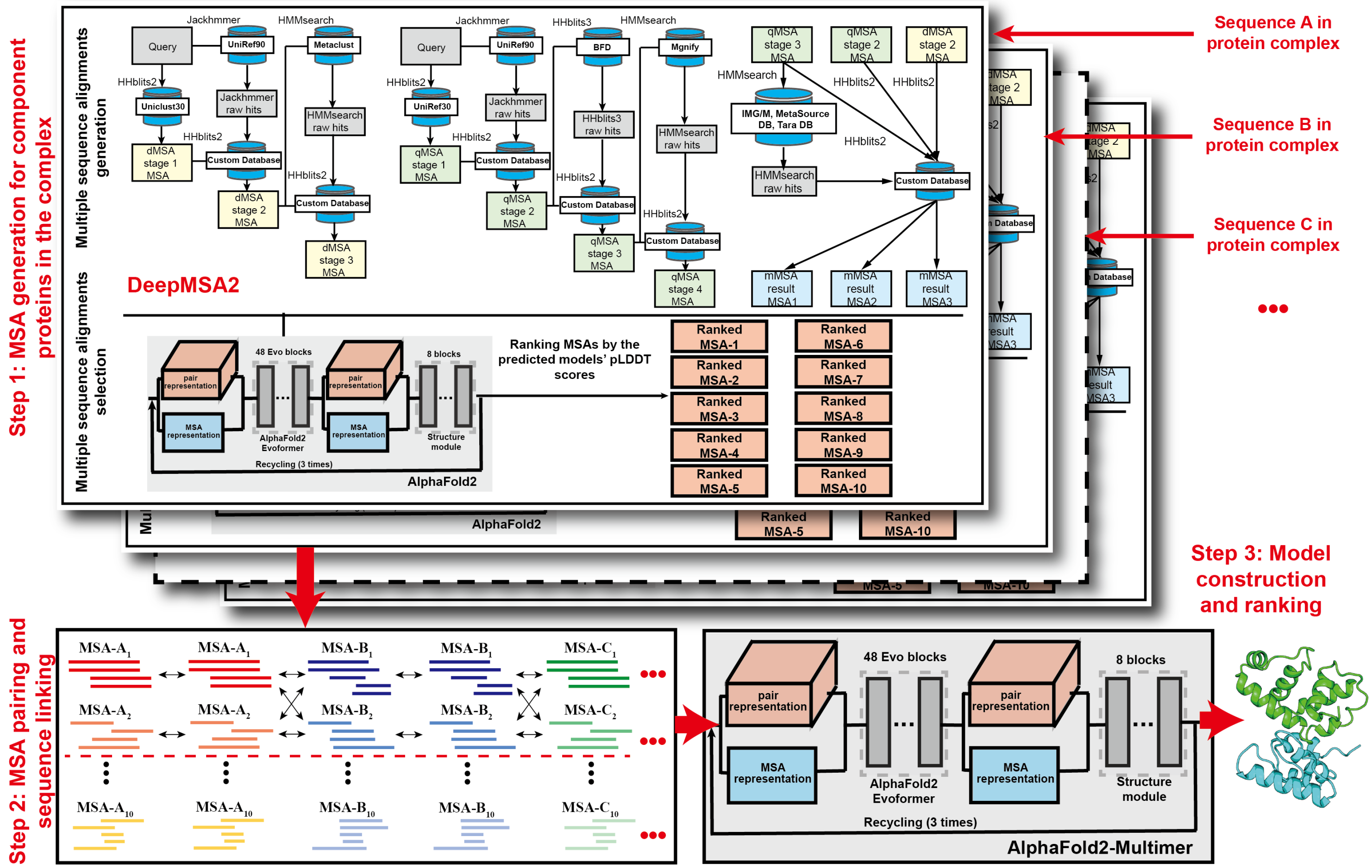
SPICKER clustering

Cluster centroid

FG-MD/locPREFMD AF2 relax refinement

Final model

‘Zheng’ built on DeepMSAFold-Multimer for protein complex modeling



Step 1: MSA generation for component proteins in the complex

Sequence A in protein complex

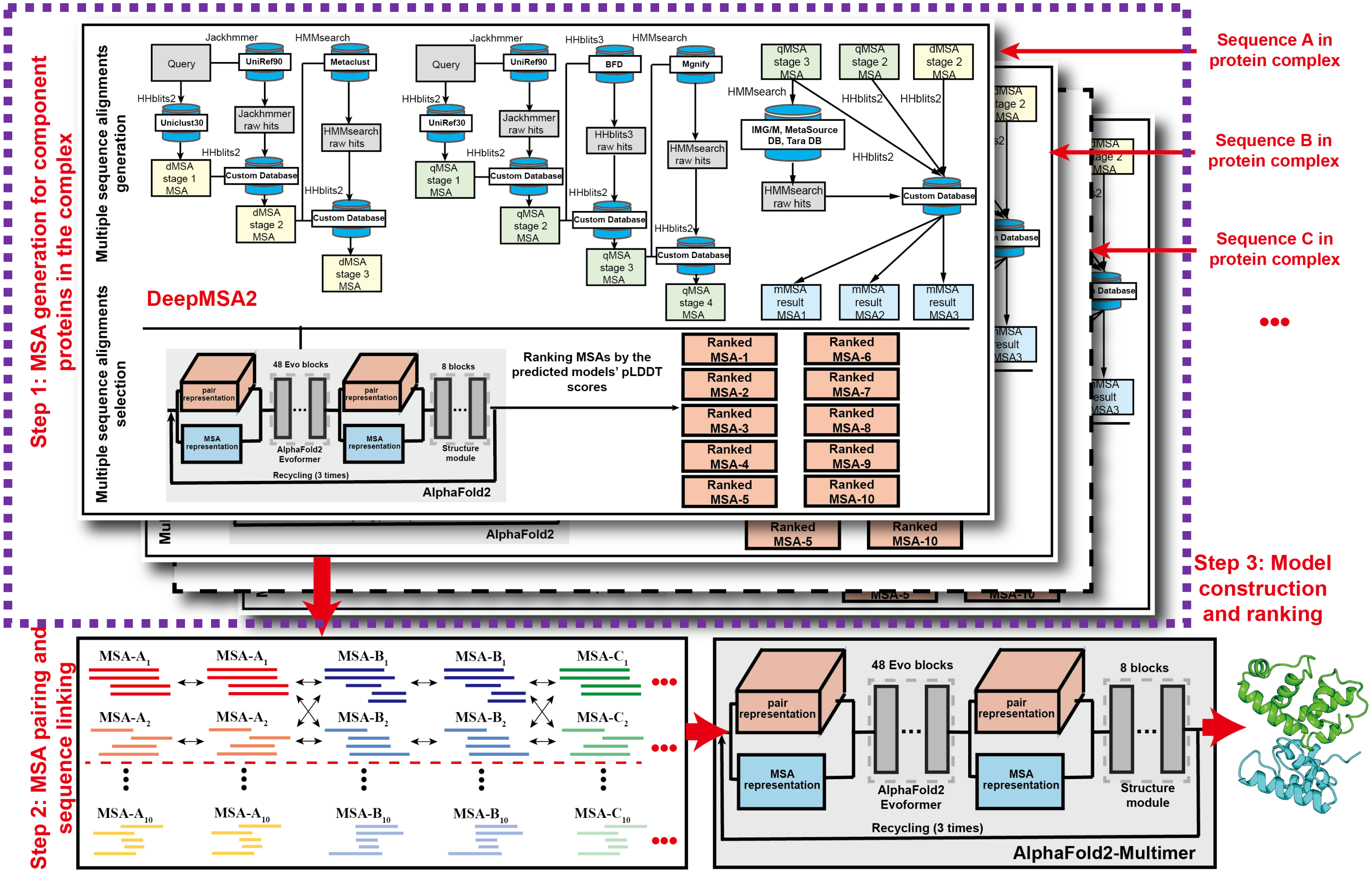
Sequence B in protein complex

Sequence C in protein complex

Step 2: MSA pairing and sequence linking

Step 3: Model construction and ranking

Zheng built on DeepMSAFold-Multimer for protein complex modeling



Step 1: MSA generation for component proteins in the complex

Step 2: MSA pairing and sequence linking

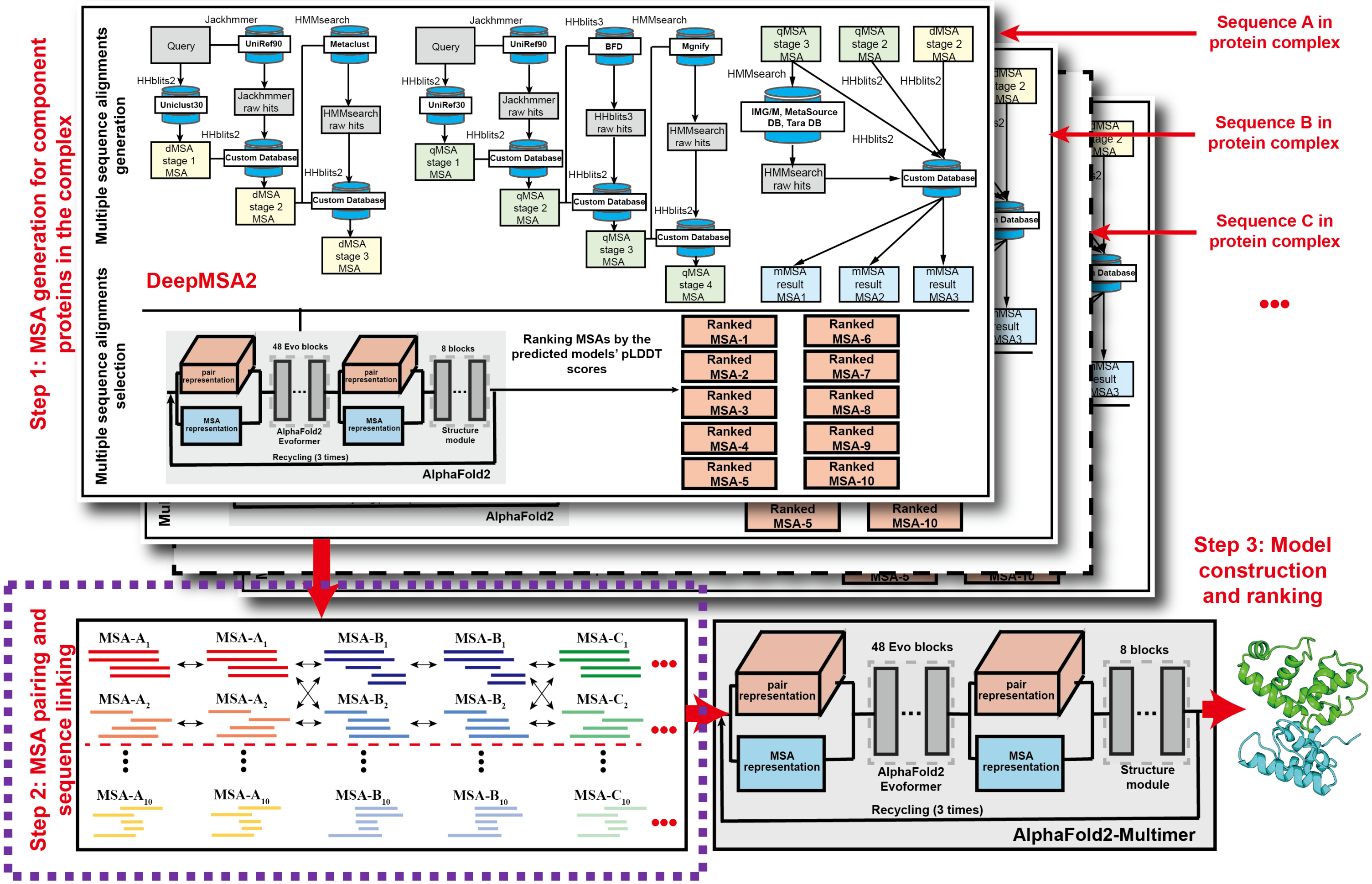
Step 3: Model construction and ranking

Sequence A in protein complex

Sequence B in protein complex

Sequence C in protein complex

Zheng built on DeepMSAFold-Multimer for protein complex modeling



Step 1: MSA generation for component proteins in the complex

Sequence A in protein complex

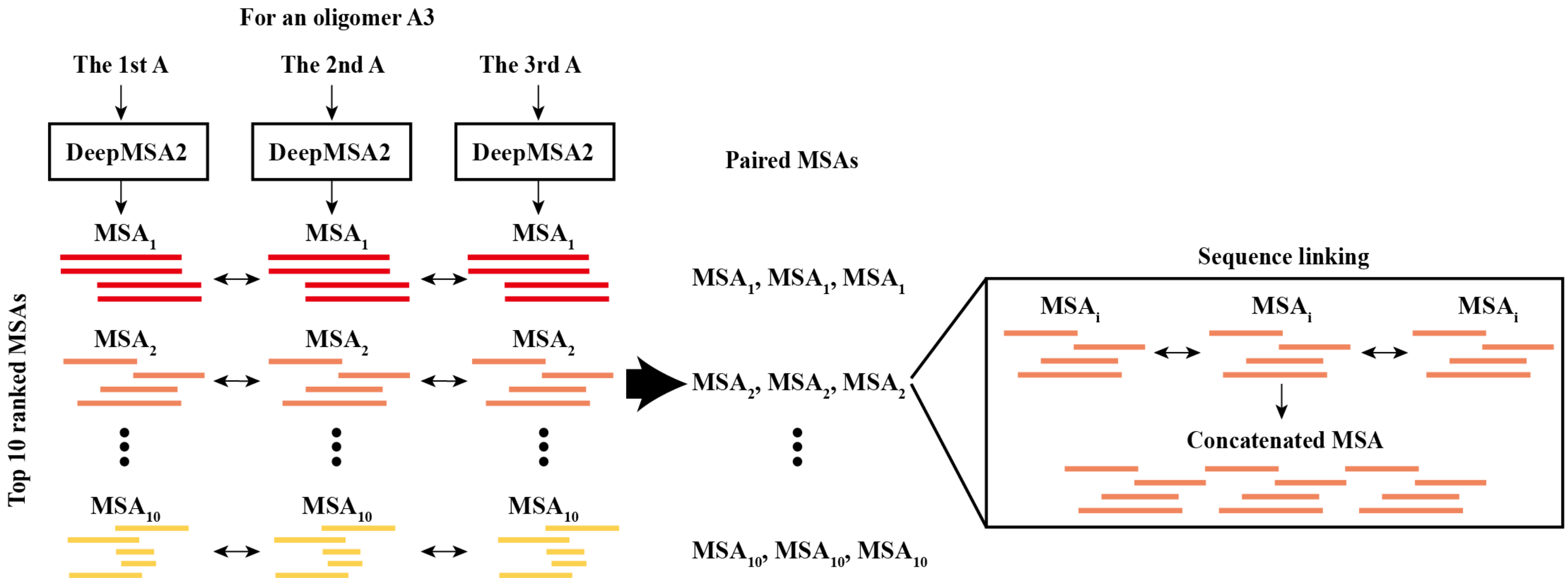
Sequence B in protein complex

Sequence C in protein complex

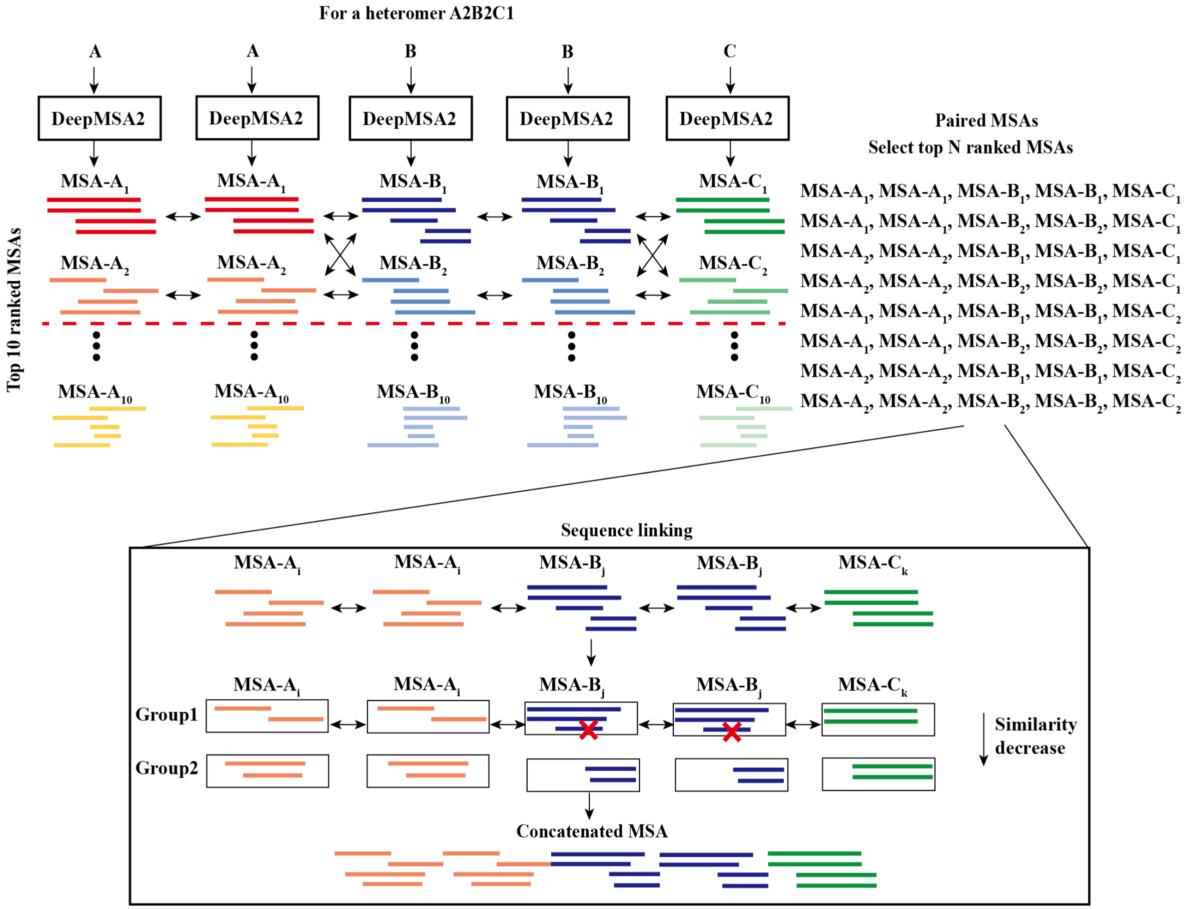
Step 2: MSA pairing and sequence linking

Step 3: Model construction and ranking

Homo-oligomer MSA paring and sequence linking



Heteromer MSA pairing and sequence linking

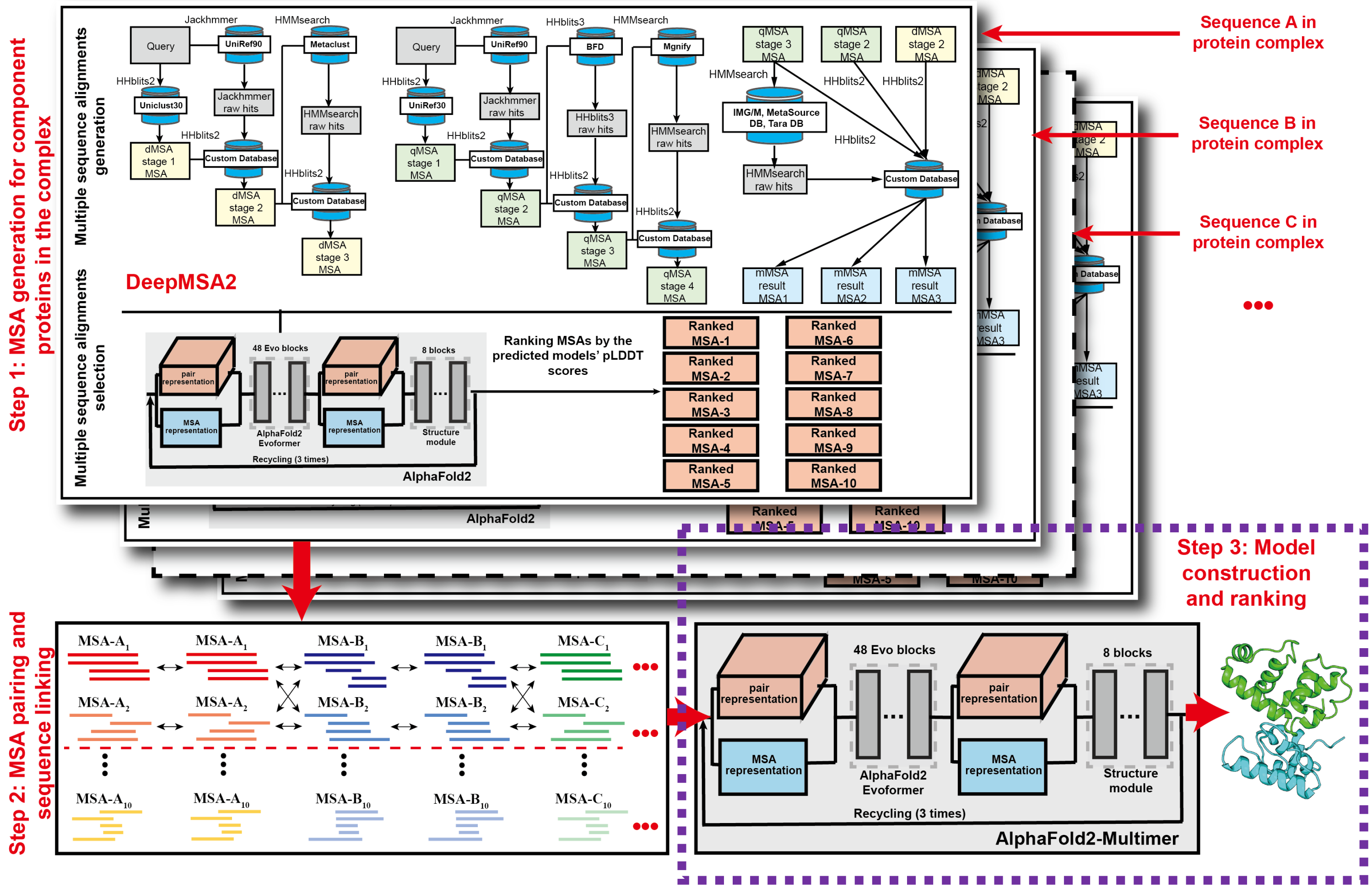


Paired MSAs
Select top N ranked MSAs

MSA-A_1, MSA-A_1, MSA-B_1, MSA-B_1, MSA-C_1
MSA-A_1, MSA-A_1, MSA-B_2, MSA-B_2, MSA-C_1
MSA-A_2, MSA-A_2, MSA-B_1, MSA-B_1, MSA-C_1
MSA-A_2, MSA-A_2, MSA-B_2, MSA-B_2, MSA-C_1
MSA-A_1, MSA-A_1, MSA-B_1, MSA-B_1, MSA-C_2
MSA-A_1, MSA-A_1, MSA-B_2, MSA-B_2, MSA-C_2
MSA-A_2, MSA-A_2, MSA-B_1, MSA-B_1, MSA-C_2
MSA-A_2, MSA-A_2, MSA-B_2, MSA-B_2, MSA-C_2

$$P = N^M \leq 100$$

Zheng is based on DeepMSAFold-Multimer for protein complex modeling



Step 1: MSA generation for component proteins in the complex

Sequence A in protein complex

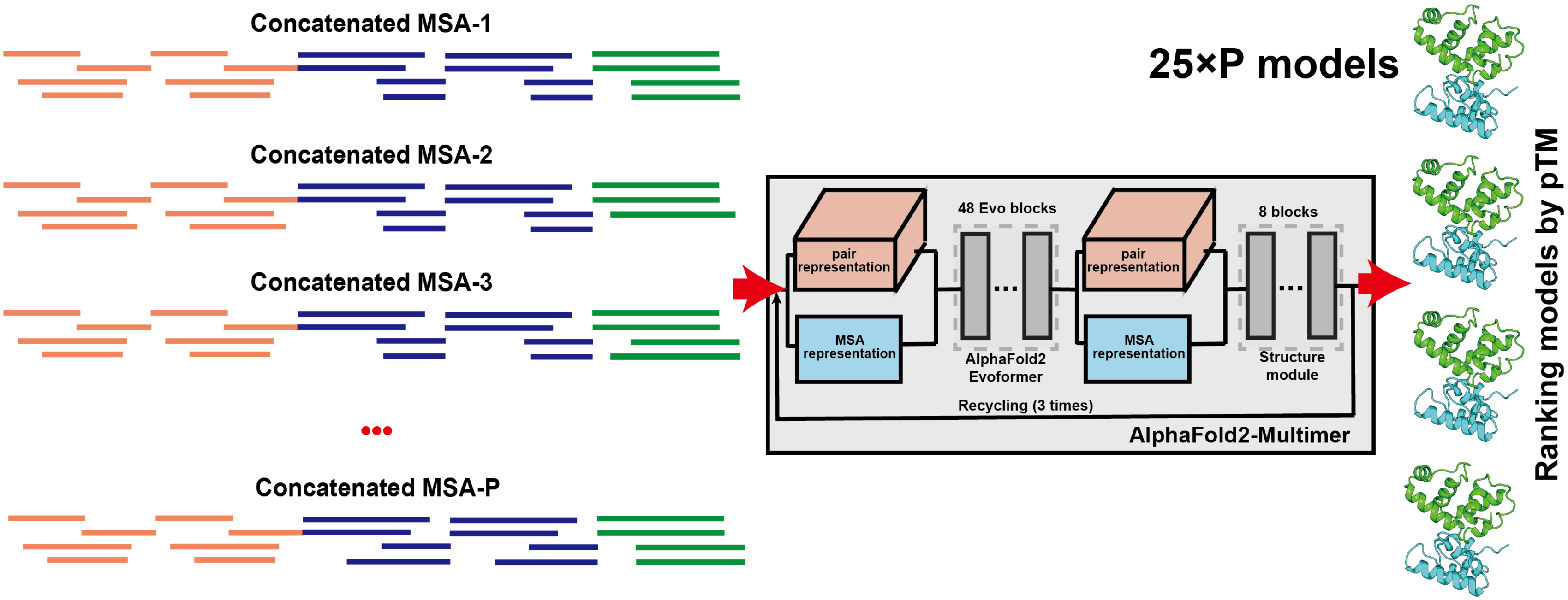
Sequence B in protein complex

Sequence C in protein complex

Step 2: MSA pairing and sequence linking

Step 3: Model construction and ranking

Complex model generation based on Multi-MSAs



Concatenated MSA-1

Concatenated MSA-2

Concatenated MSA-3

•••

Concatenated MSA-P

48 Evo blocks

pair representation

MSA representation

AlphaFold2 Evoformer

8 blocks

Structure module

Recycling (3 times)

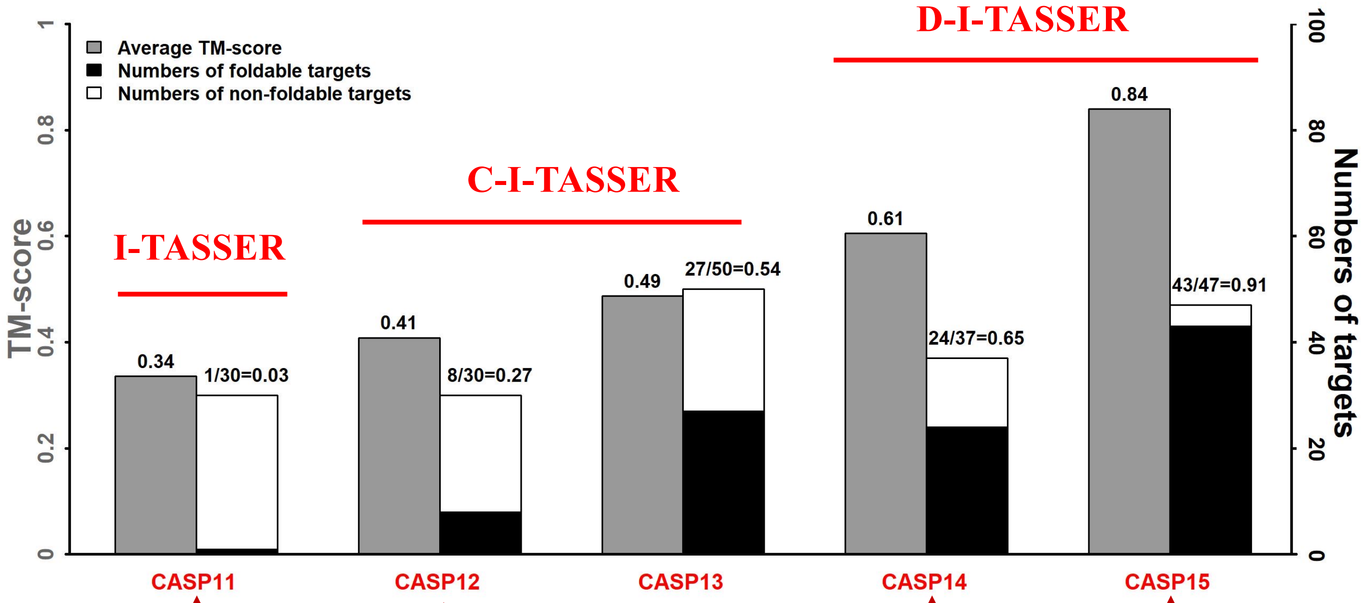
AlphaFold2-Multimer

25×P models

Ranking models by pTM

Results for UM-TBM (D-I-TASSER) server

Summary of FM targets folded by I-TASSER series algorithm



Fragment assembly-based

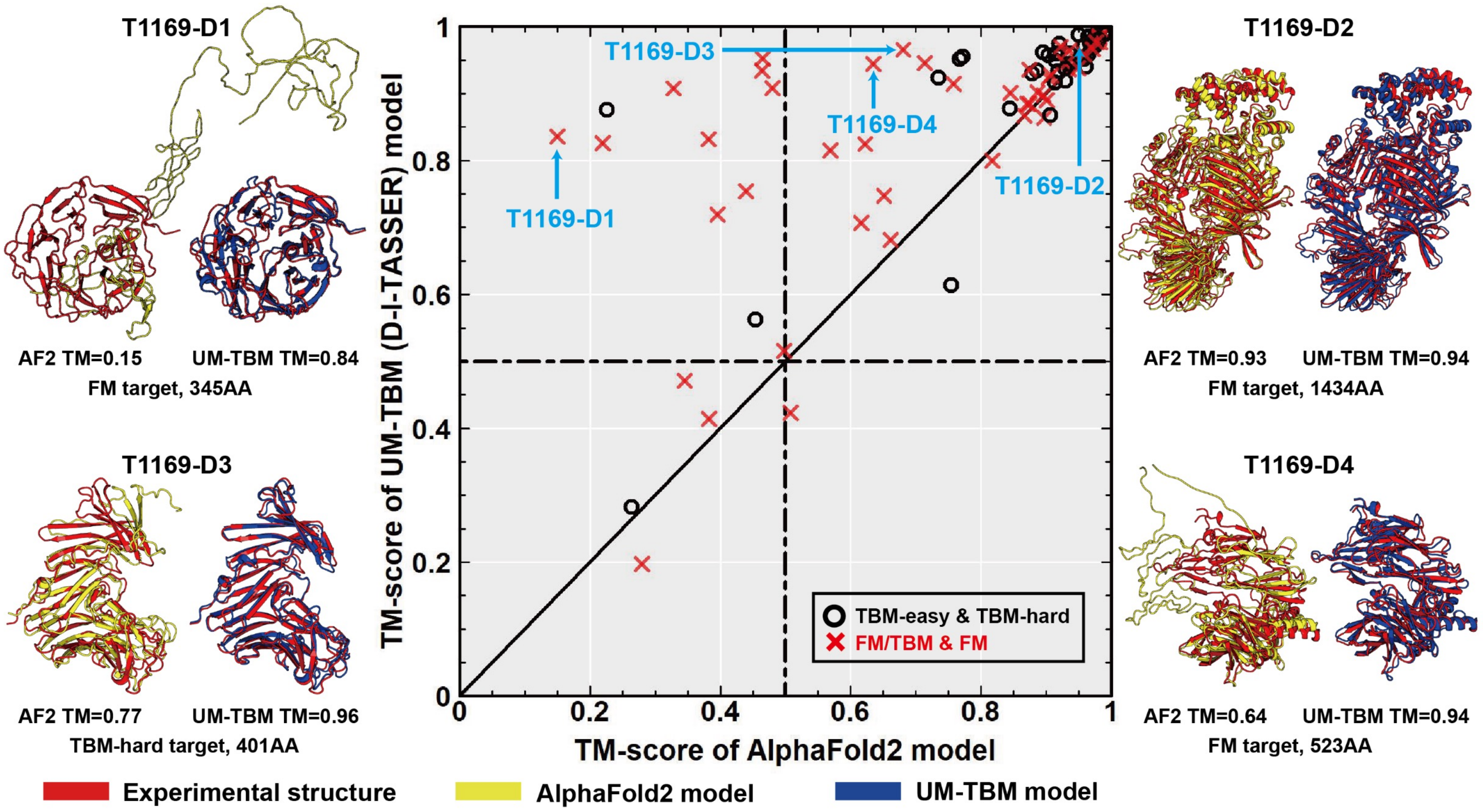
Direct coupling analysis-based contacts added

Deep-learning-based contacts added

Deep-learning-based distances added (DeepPotential)

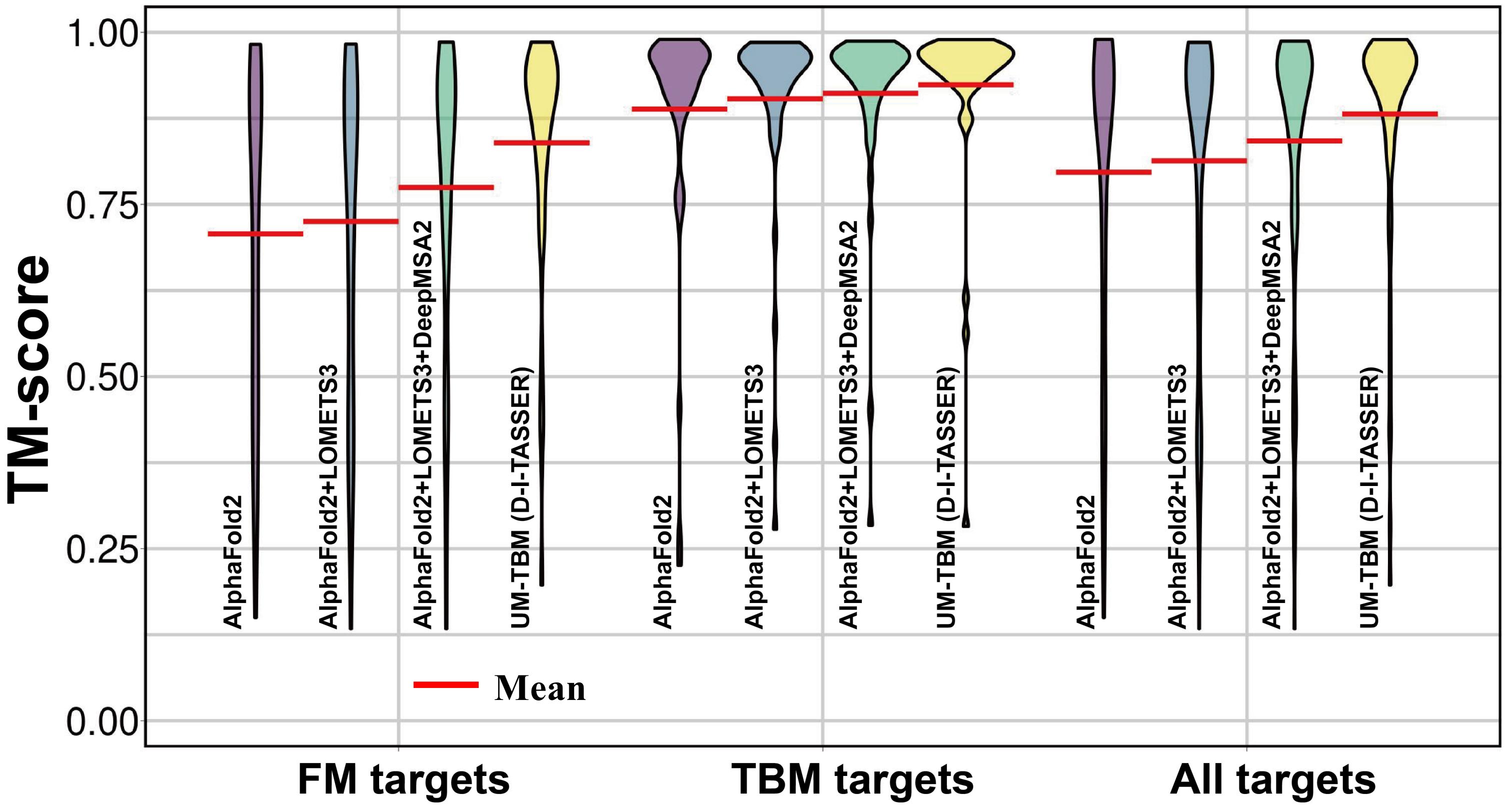
Deep-learning-based distances updated (DeepPotential, AttentionPotential, AlphaFold2)

D-I-TASSER vs standard AlphaFold2 on CASP15 domains



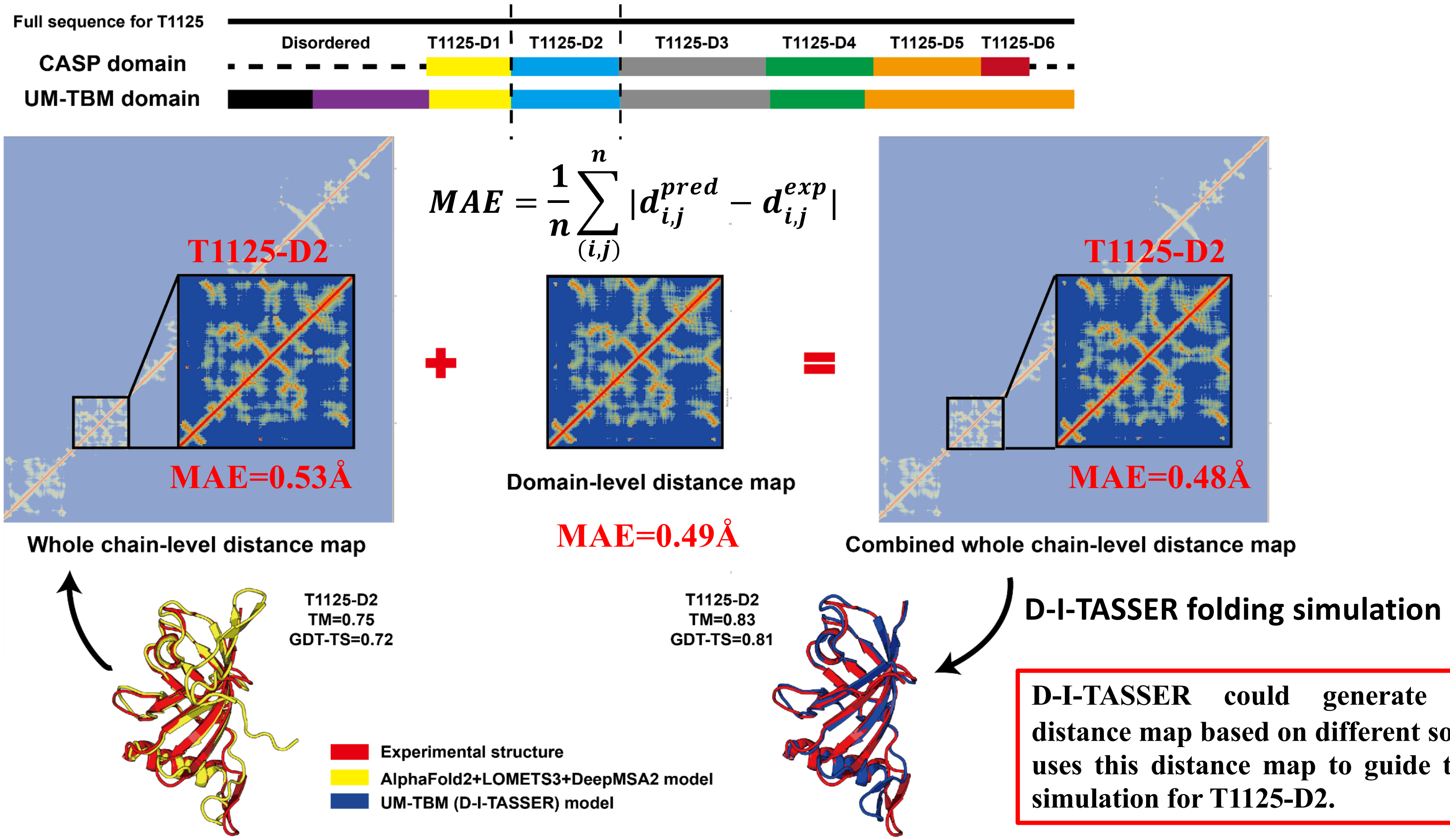
Standard AlphaFold2 data are taken from CASP15 NBIS-AF2-standard server.

Impact of different components in D-I-TASSER



Alphafold2 (with LOMETS3 and DeepMSA2) models are used as the initial conformation of D-I-TASSER simulations, thus the further improvement represents the contribution of D-I-TASSER folding simulation and other modules used in the UM-TBM pipeline.

A case study of T1125-D2 to highlight the advance of D-I-TASSER folding

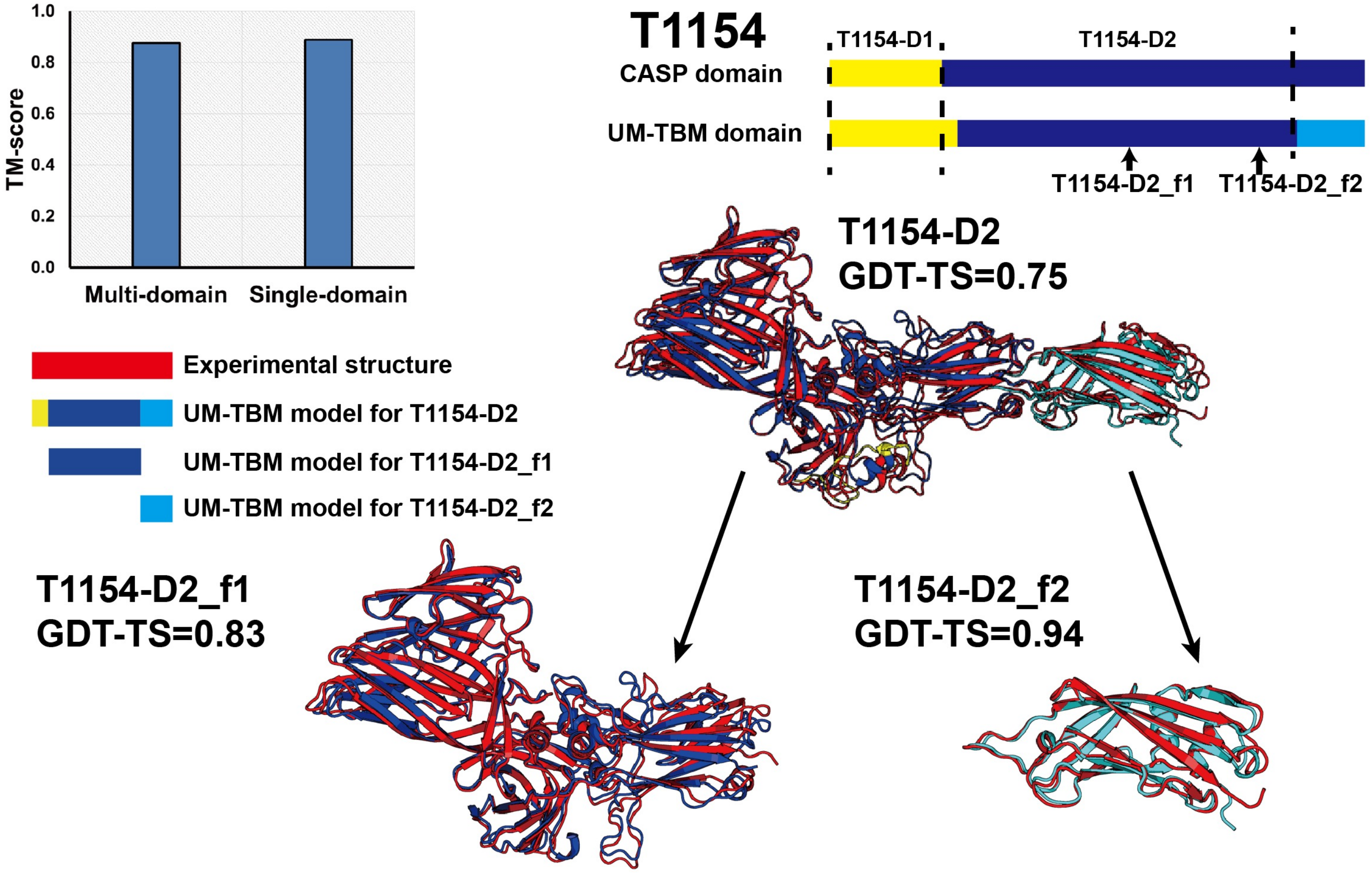


$$MAE = \frac{1}{n}\sum_{(i,j)}^{n} |d_{i,j}^{pred} - d_{i,j}^{exp}|$$

🟥 Experimental structure
🟨 AlphaFold2+LOMETS3+DeepMSA2 model
🟦 UM-TBM (D-I-TASSER) model

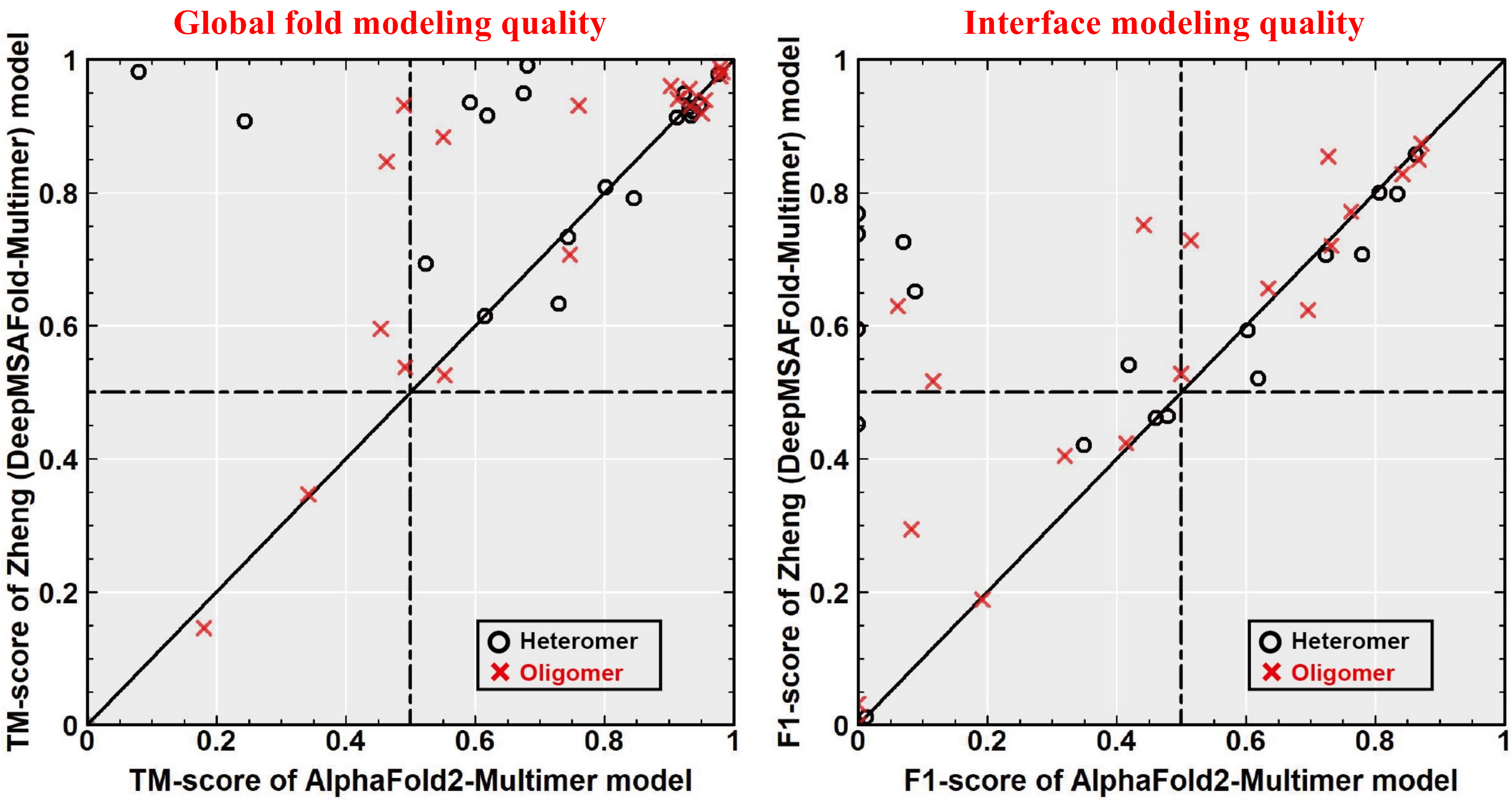
D-I-TASSER could generate a better distance map based on different sources, and uses this distance map to guide the folding simulation for T1125-D2.

Domain partition problem in T1154-D2



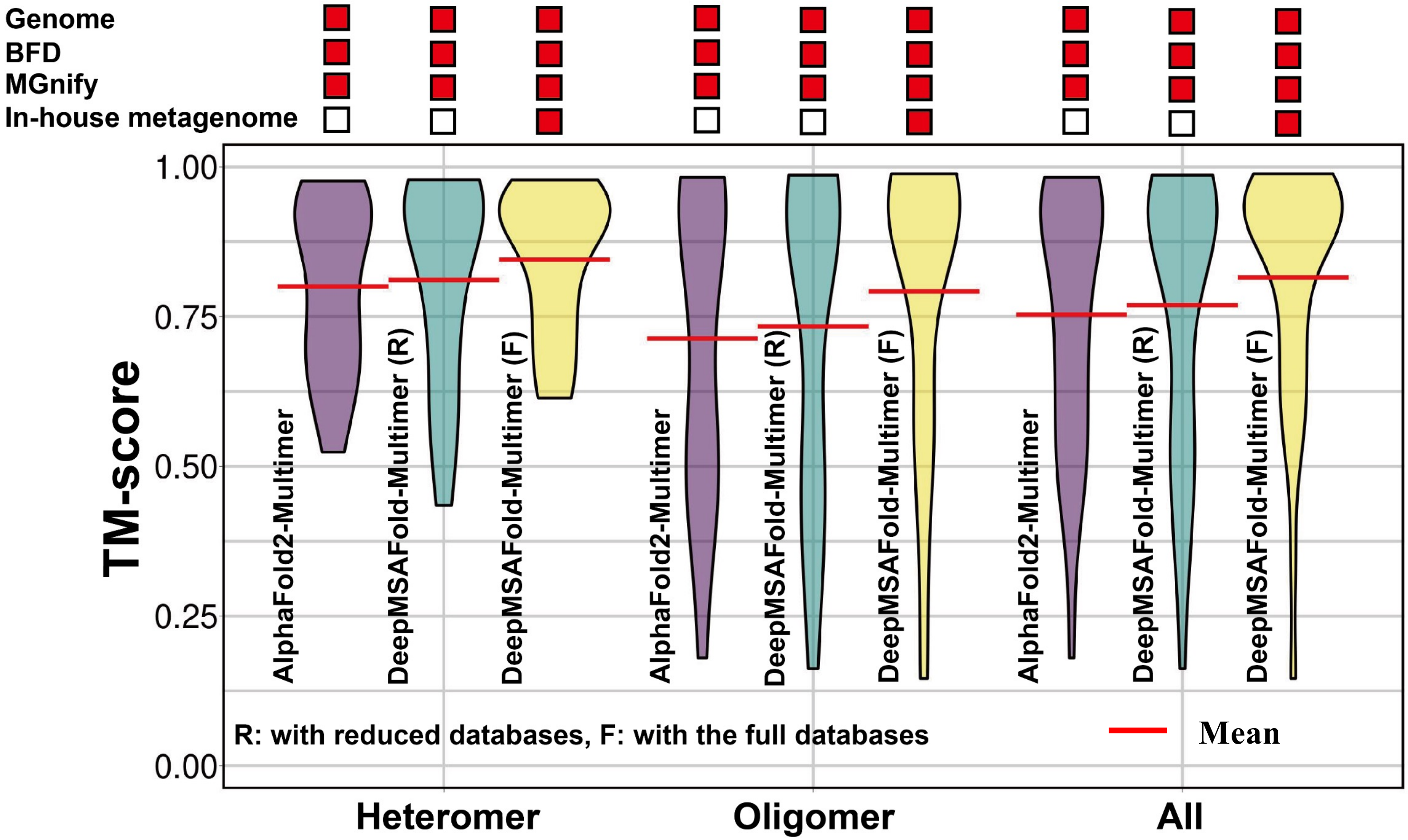
Results for ‘Zheng’ (DeepMSAFold-Multimer) group

DeepMSAFold-Multimer vs standard AlphaFold2-Multimer

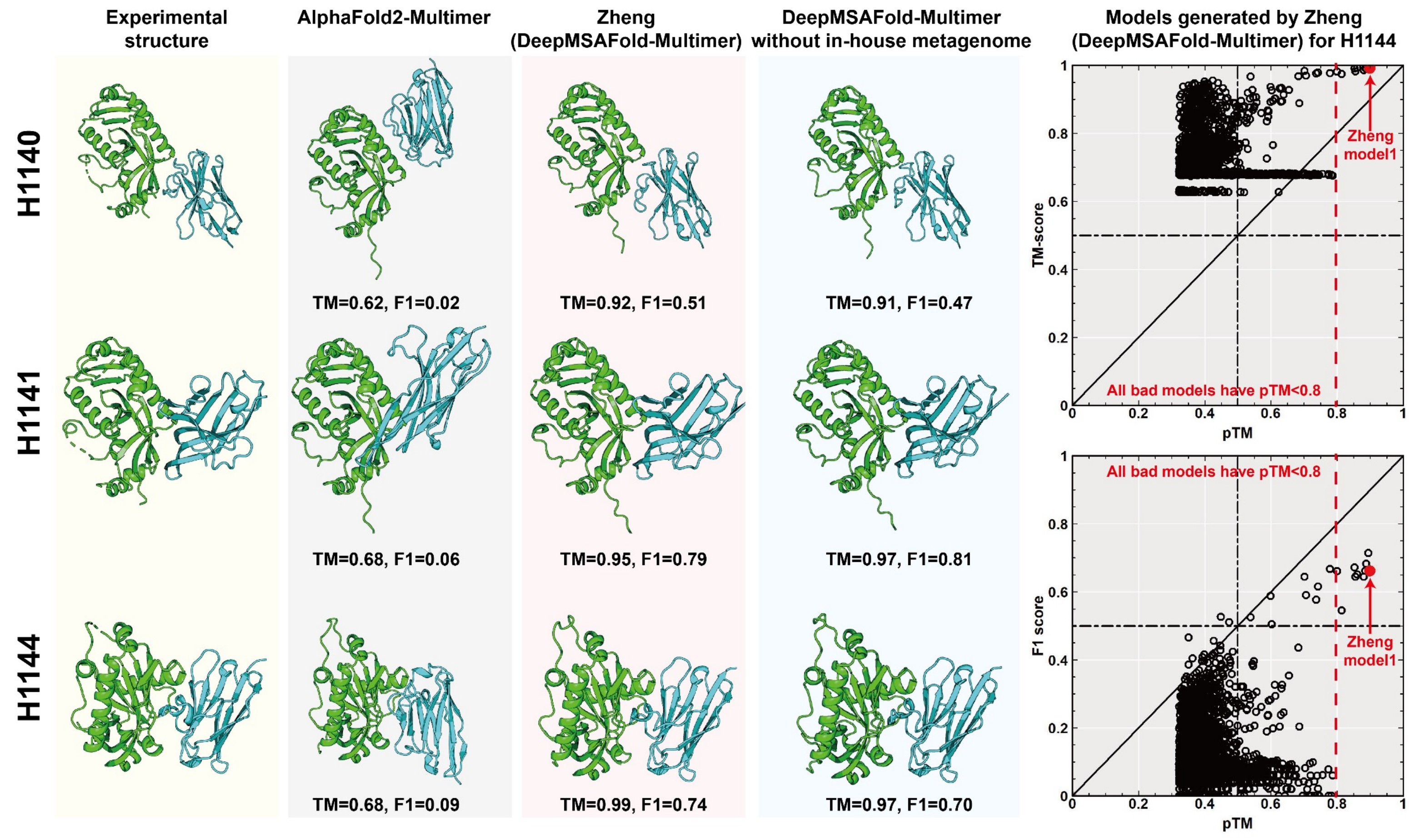


Standard AlphaFold2 data are taken from CASP15 NBIS-AF2-multimer server.

Impact of the MSA generation and database for complex modeling



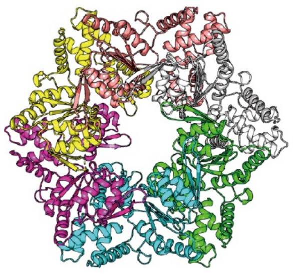
High quality model generation for Nanobody-antigen by pairing MSAs



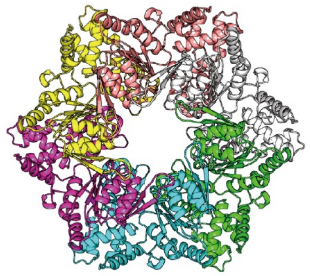
DeepMSAFold-Multimer works well for producing high quality models by paring the different MSAs and ranking the model with pTM.

Model ranking problem with pTM score in H1172

T1170o, A6, 1908AA

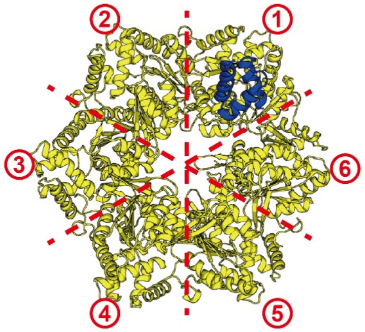


Experimental structure

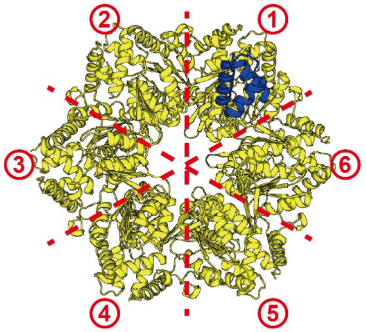


Zheng model1
pTM=0.793
TM=0.93, F1=0.58

H1171, A6B1, 1956AA

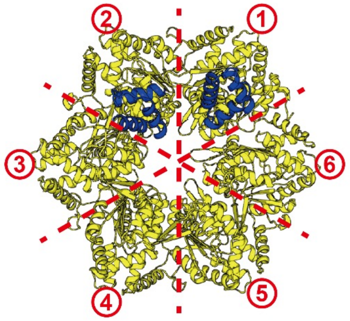


Experimental structure

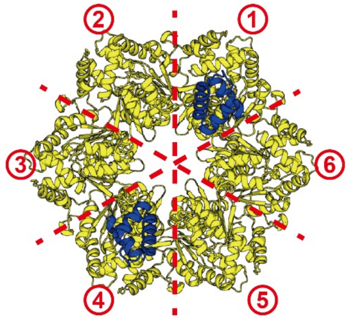


Zheng model1
pTM=0.764
TM=0.93, F1=0.51

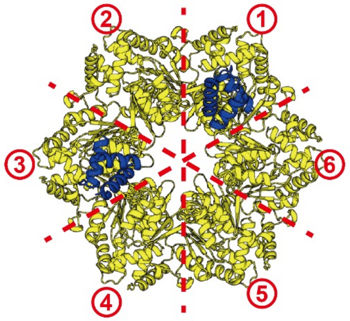
H1172, A6B2, 2004AA



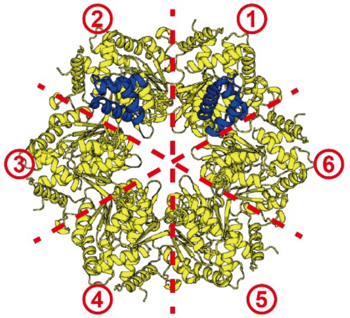
Experimental structure



Zheng model1
pTM=0.746
TM=0.91, F1=0.54



Zheng model2
pTM=0.738
TM=0.91, F1=0.50



Zheng model3
pTM=0.733
TM=0.93, F1=0.54

The pTM scores are not sufficient for identifying the best model among models that have similar pTM scores.

Summary

- What went right
  - D-I-TASSER algorithm that integrates threading templates, structure model-ranked MSAs, deep learning-based spatial restraints, and an optimized folding system with comprehensive force field works well for protein monomer structure prediction.
  - DeepMSA2 with high-quality MSA generating, ranking, and paring system helps improve the model quality for both protein monomer and protein complex.
- What went wrong
  - AlphaFold2's QA score (pLDDT and pTM) has the ability to distinguish models that are significantly better, but it is not sensitive enough to rank high-quality models.
  - Domain partition is still meaningful and important for large multi-domain protein modeling, however the accuracy of domain boundary prediction is still not yet satisfactory.

Acknowledgements

Yang Zhang

Yang Li

Robin Pearce

Jonathan Poisson

NSF
National Science Foundation

NIH
National Institutes of Health

M | ARC
Advanced Research Computing
(ARC) at the
University of Michigan

POWERED BY
XSEDE
Extreme Science and Engineering
Discovery Environment

C
A
S
P

Thank you
Q&A