CASP16 Conference

Team FZZH

EnsembleFold: Alternative conformation prediction using multi-MSA strategy and structural clustering

12/04/2024

Speaker: Wei Zheng

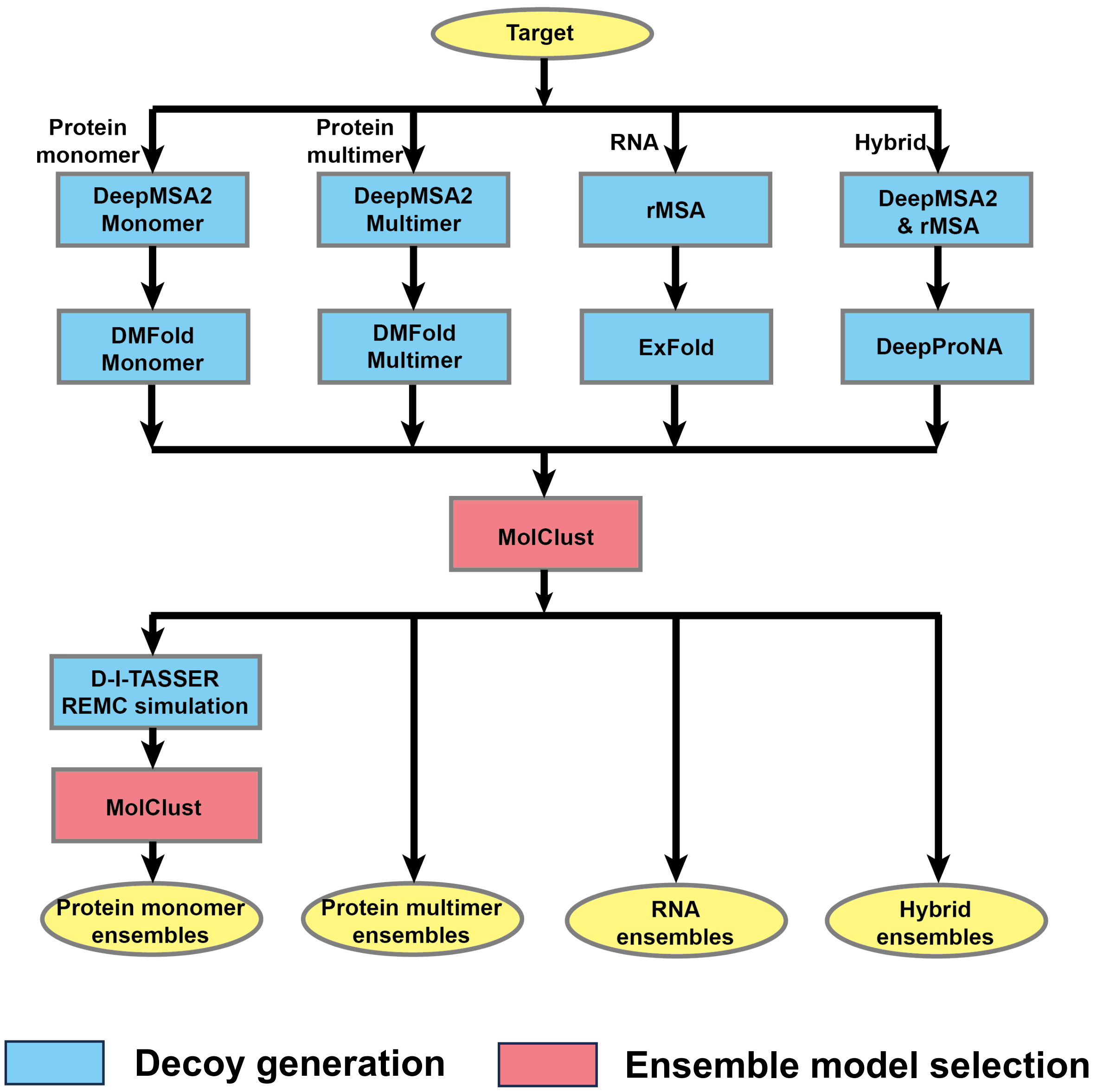
Wei Zheng, Qiqige Wuyun, Quancheng Liu, Chunxiang Peng,
Wentao Ni, Gang Hu, Ziying Zhang, Xiaogen Zhou, Lydia Freddolino

Outline

1 **Methods**

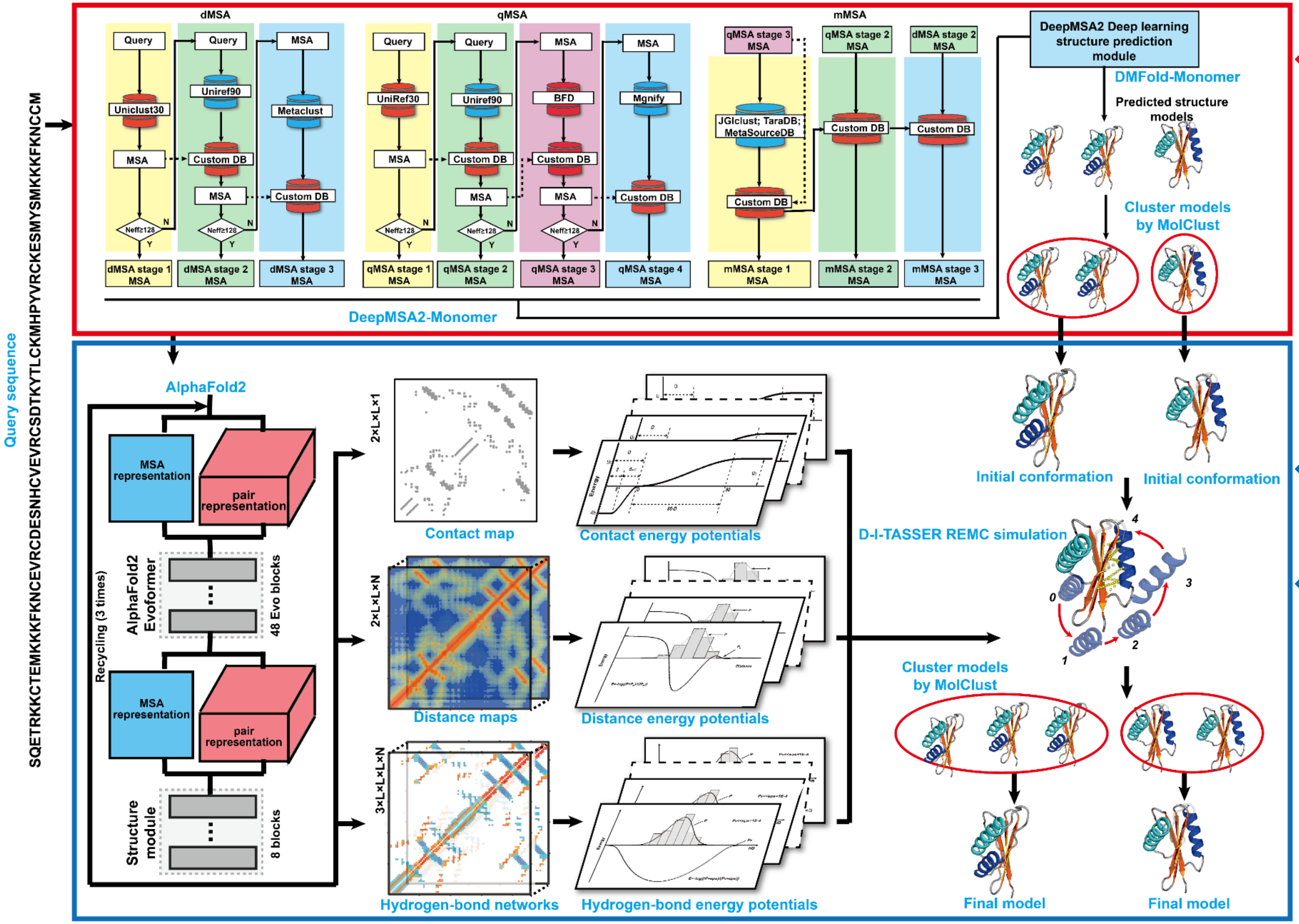
2 Results

Methods: Overall pipeline of EnsembleFold for ensemble targets



- Zheng-Server, Zheng-Multimer, Zheng, MIEnsembles-Server, and NKRNAs participated in CASP16
- MIEnsembles-Server (server group) and Zheng (human group) focus on ensemble targets
- Same pipeline, Zheng has longer running time and more combinations of MSAs
- Four different pipelines for handling protein monomer, protein complex, RNA, and hybrid targets

Methods: Protein monomer ensemble prediction by EnsembleFold



DeepMSA2-Monomer & DMFold-Monomer

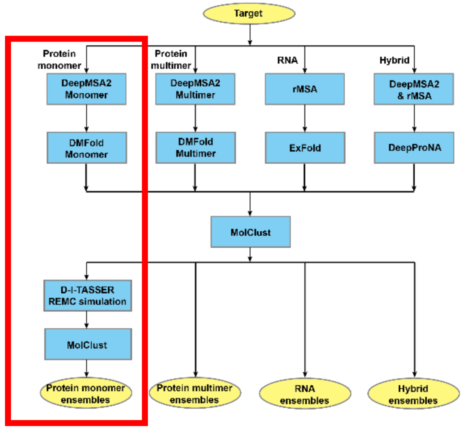
Key points:

1. Using all MSAs from DeepMSA2
2. Using DMFold models and spatial restraints from all representative models in replica exchange Monte Carlo (REMC) simulation
3. Clustering decoys from REMC simulation

D-I-TASSER REMC simulation

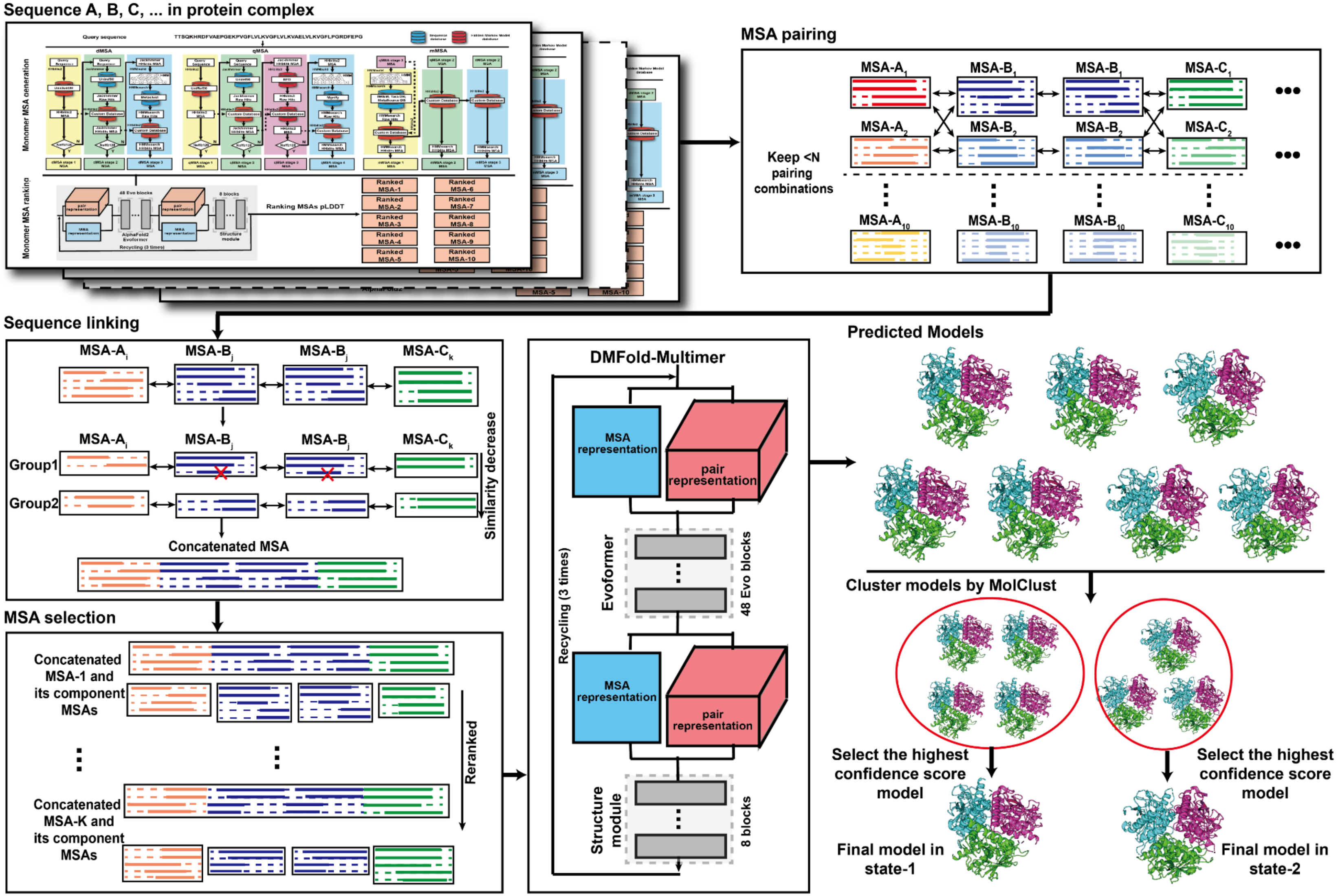
Structural clustering by MolClust

1. SPICKER
2. US-align



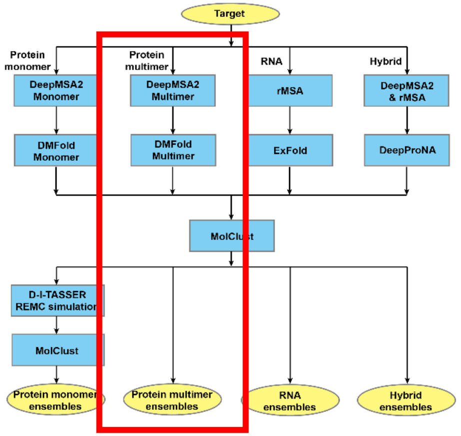
For targets: T1214, T1200 and T1300.

Methods: Protein multimer ensemble prediction by EnsembleFold



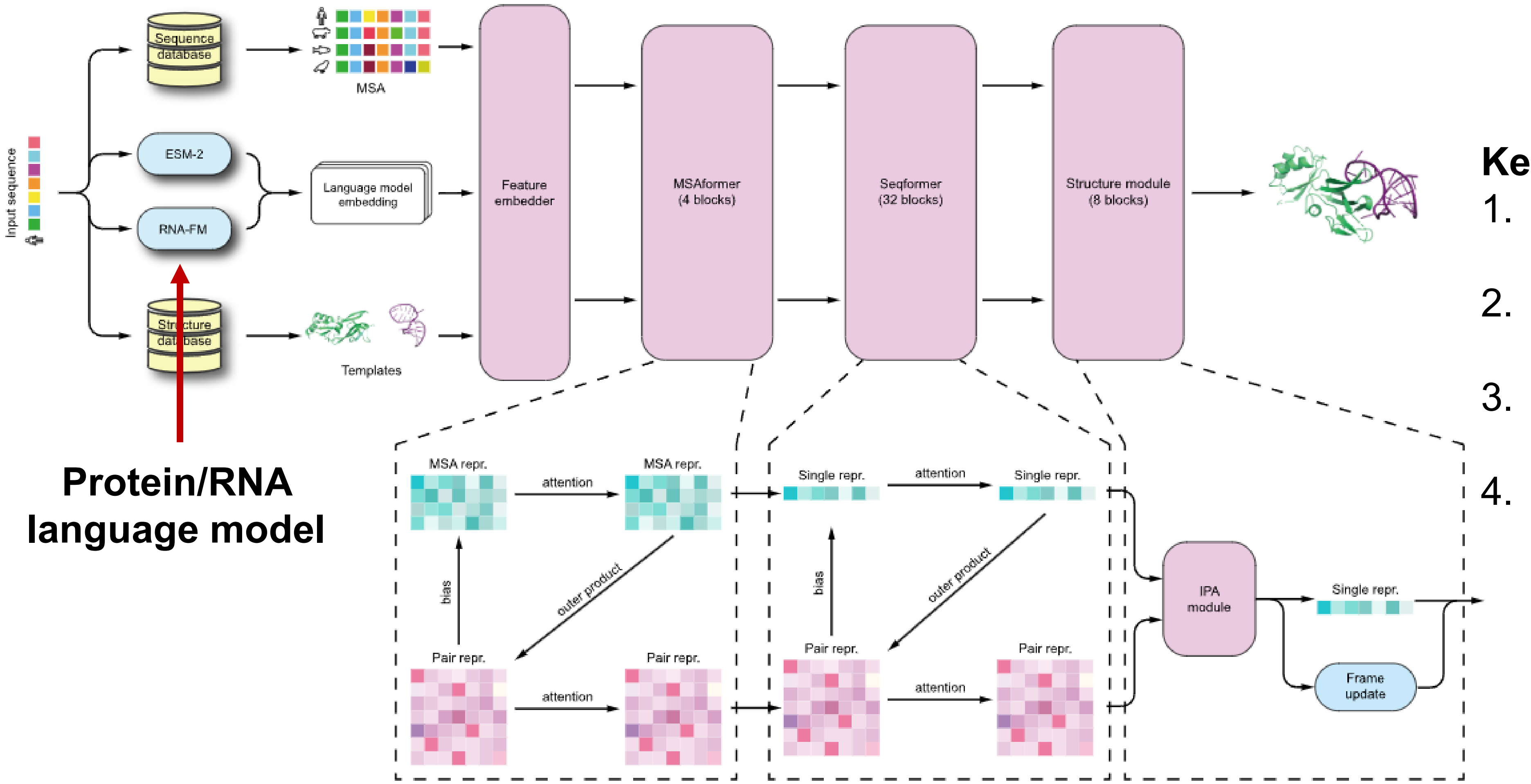
Key points:

1. Larger metagenomes than CASP15 version
2. More combinations of MSA pairing
3. Sampling strategy in modeling stage: using the template or not, opening the drop up rate or not, and using different alphafold2 pre-trained parameters (v1 v2 v3).
4. Clustering models by structural similarity, rank by highest confidence score of the members



For targets: T1249v1/v2 and T1294v1/v2.

Methods: Hybrid ensemble prediction by EnsembleFold

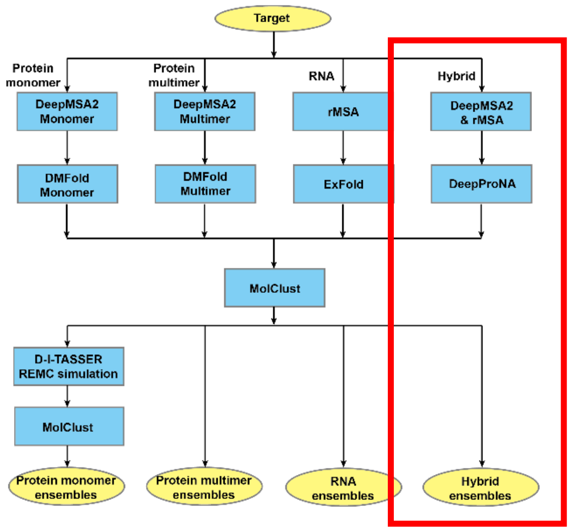


Protein/RNA language model

Wentao Ni

Key points:

1. Modified from AlphaFold2 pipeline
2. Using Protein/RNA language model
3. Using multiple sets of MSAs as input
4. Clustering the models



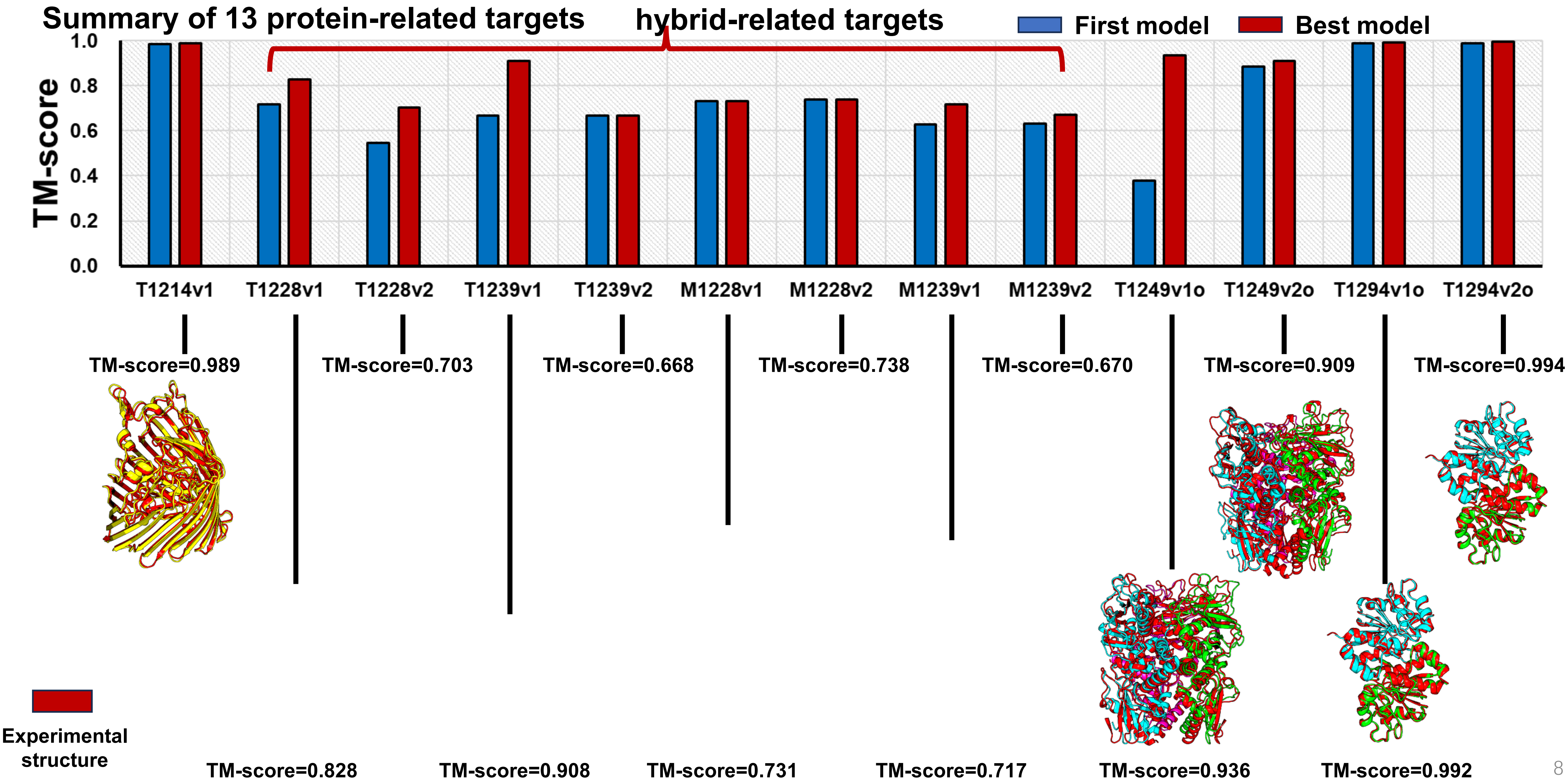
For targets: M1228v1/v2, T1228v1/v2, M1239v1/v2, and T1239v1/v2.

Outline

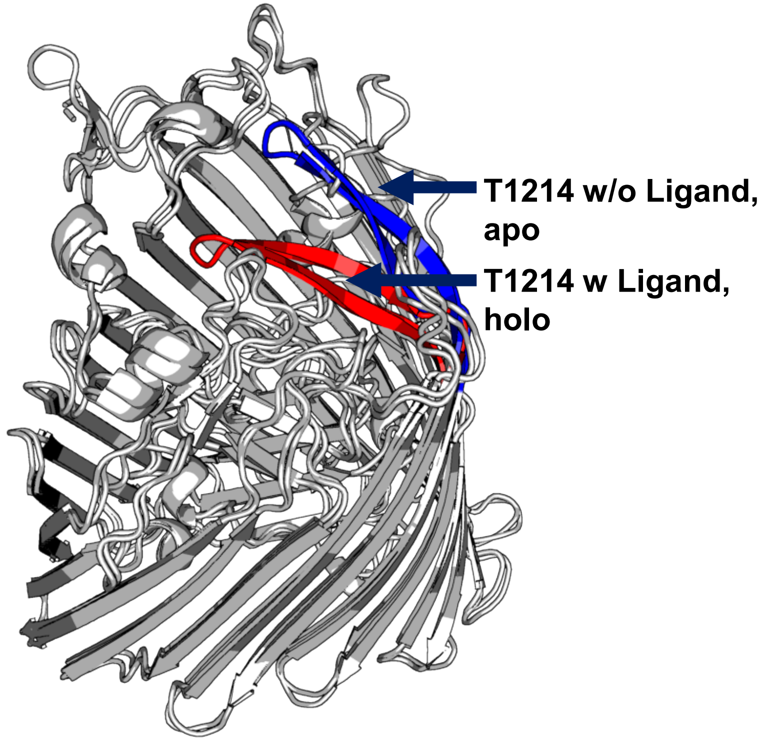
1 Methods

2 **Results**

Results: Overall results of protein-related ensemble targets



Results: T1214, what went right?

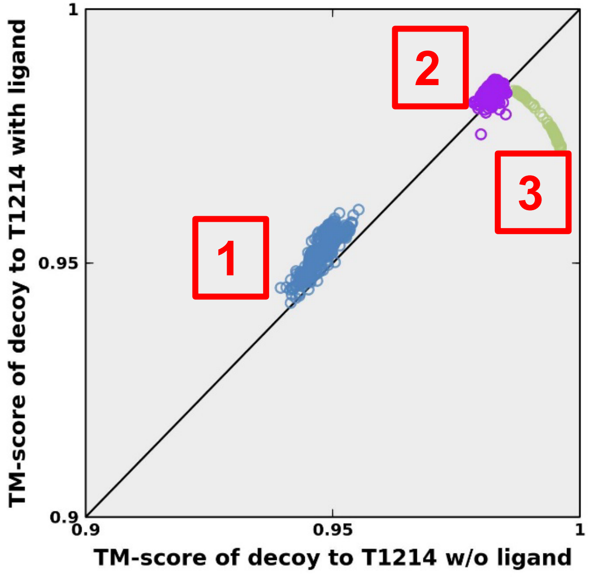


Experimental structures

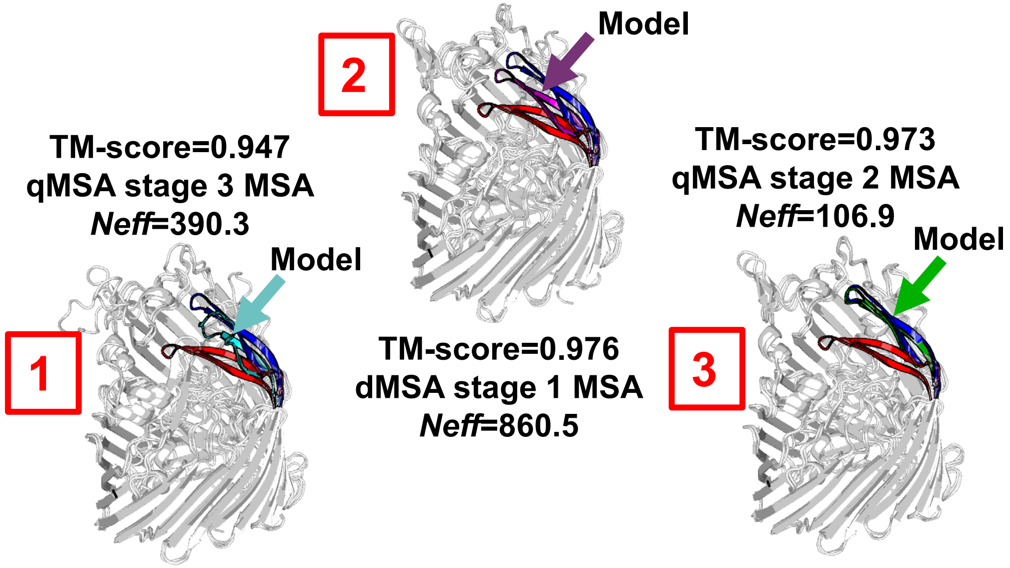
Model this target without ligand

- The model from each MSA corresponds to one ‘state’
- REMC simulation helps create a diverse set of models

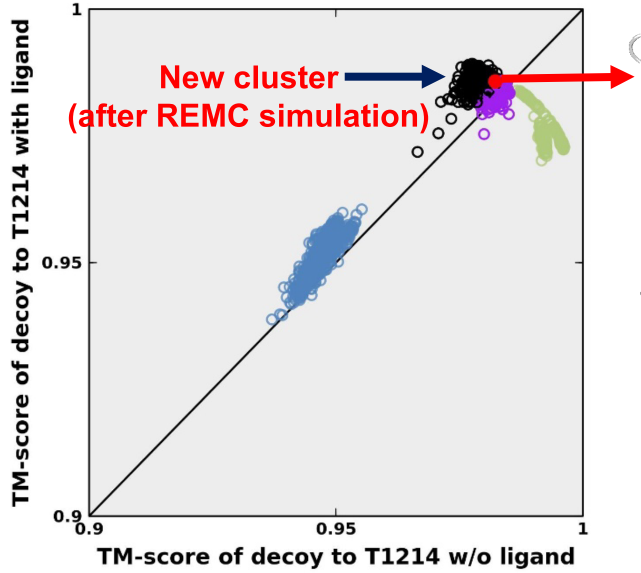
MolClust result for DMFold-Monomer decoys



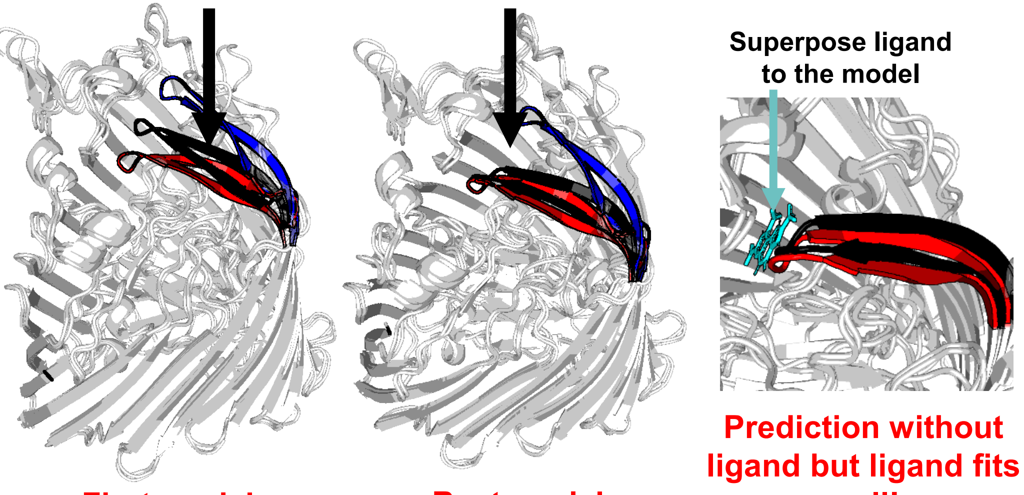
DMFold-Monomer models



MolClust result for EnsembleFold decoys



EnsembleFold models



First model TM-score=0.986

Best model TM-score=0.989

Prediction without ligand but ligand fits well!

Results: M1228/T1228 what went right?

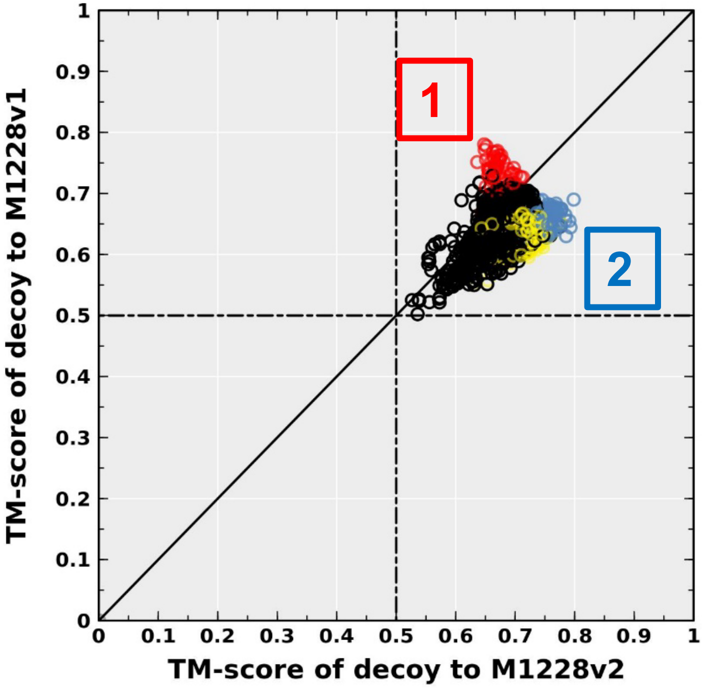
Experimental structures

M1228v1

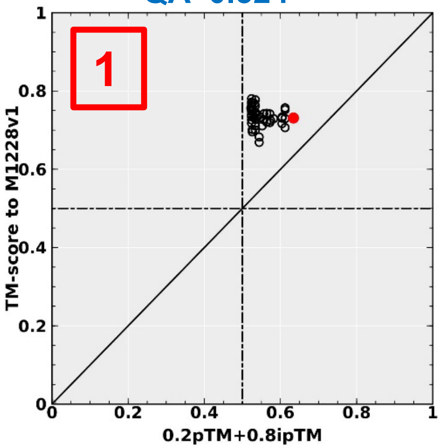
M1228v2

protein with both states shown

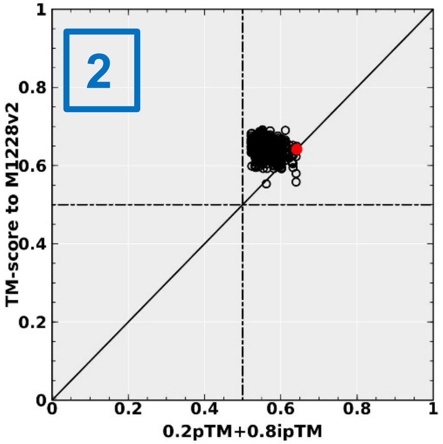
MolClust result for EnsembleFold decoys



Best decoy for M1228v1
TM-score=0.780
QA=0.524



Best decoy for M1228v2
TM-score=0.798
QA=0.612



Model1 for T1228v1
TM-score=0.718

Model1 for T1228v2
TM-score=0.545

Structural clustering works well for picking the correct model of each state in this case

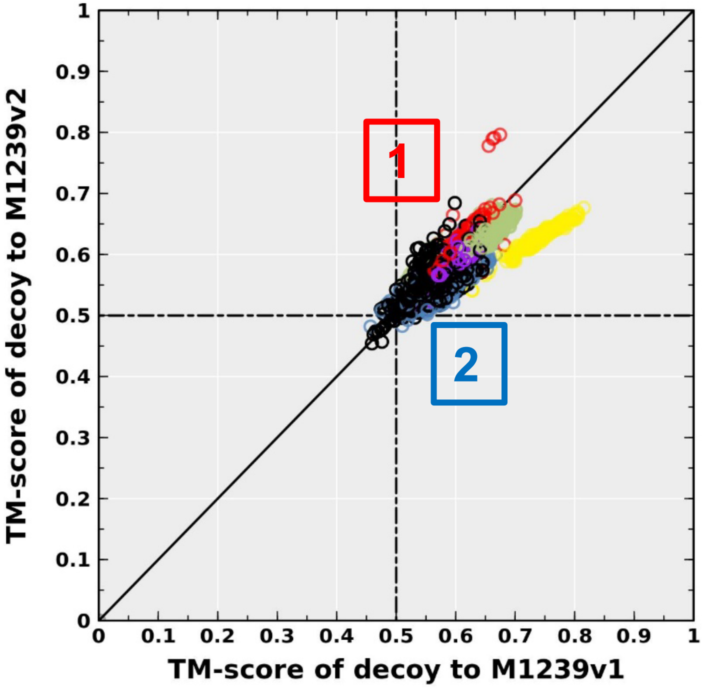
Results: M1239/T1239, what went right?

Experimental structures

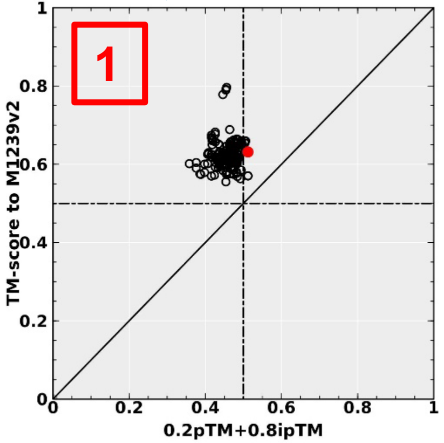
M1239v2

M1239v1

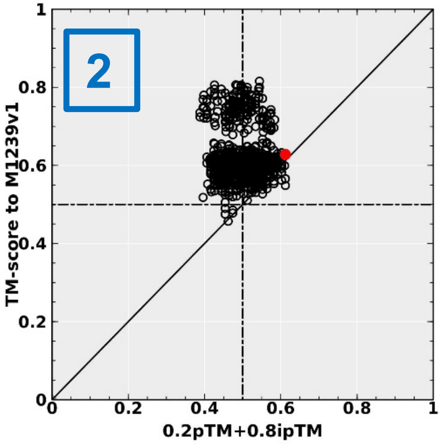
MolClust result for EnsembleFold decoys



Best decoy for M1239v2
TM-score=0.796
QA=0.456



Best decoy for M1239v1
TM-score=0.816
QA=0.544



Model1 for T1239v2
TM-score=0.668

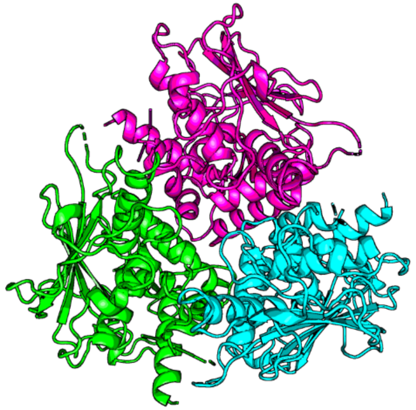
Model1 for T1239v1
TM-score=0.668

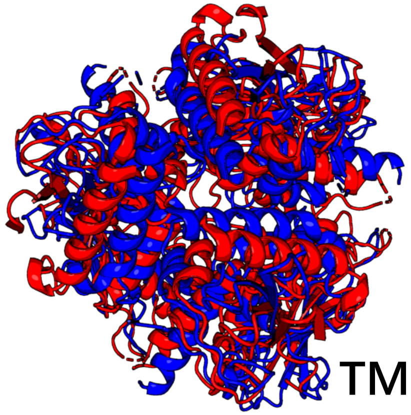
Structural clustering works well for picking the correct model of each state in this case

protein with both states shown

Results: T1249o, what went wrong?

Experimental structure
Trimer



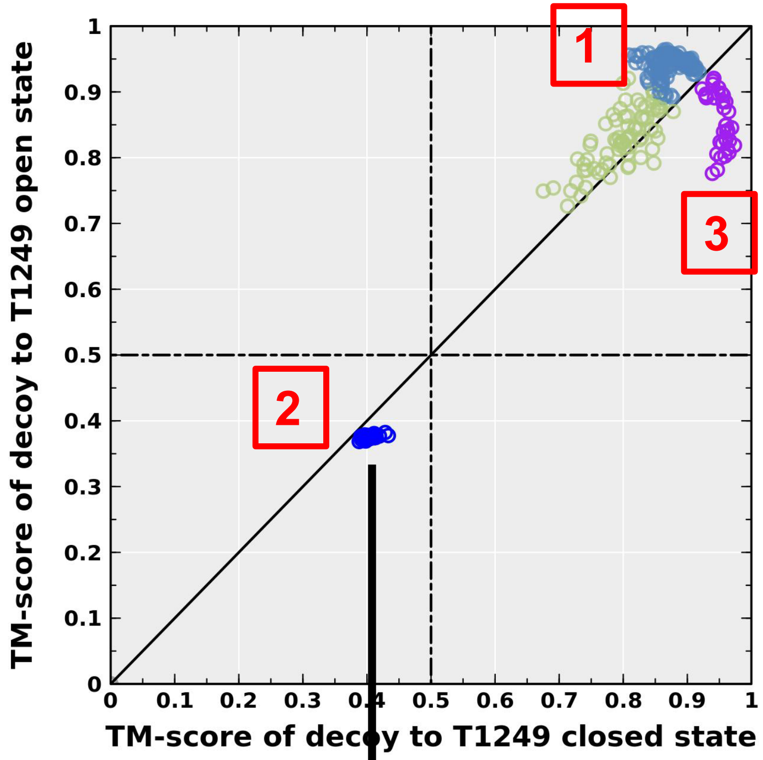


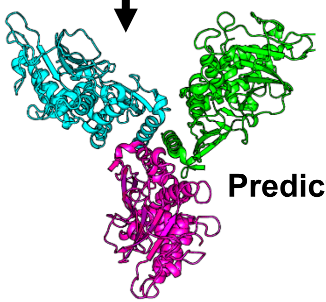
TM-score=0.83

Experimental structures

🟥 Open state

🟦 Closed state

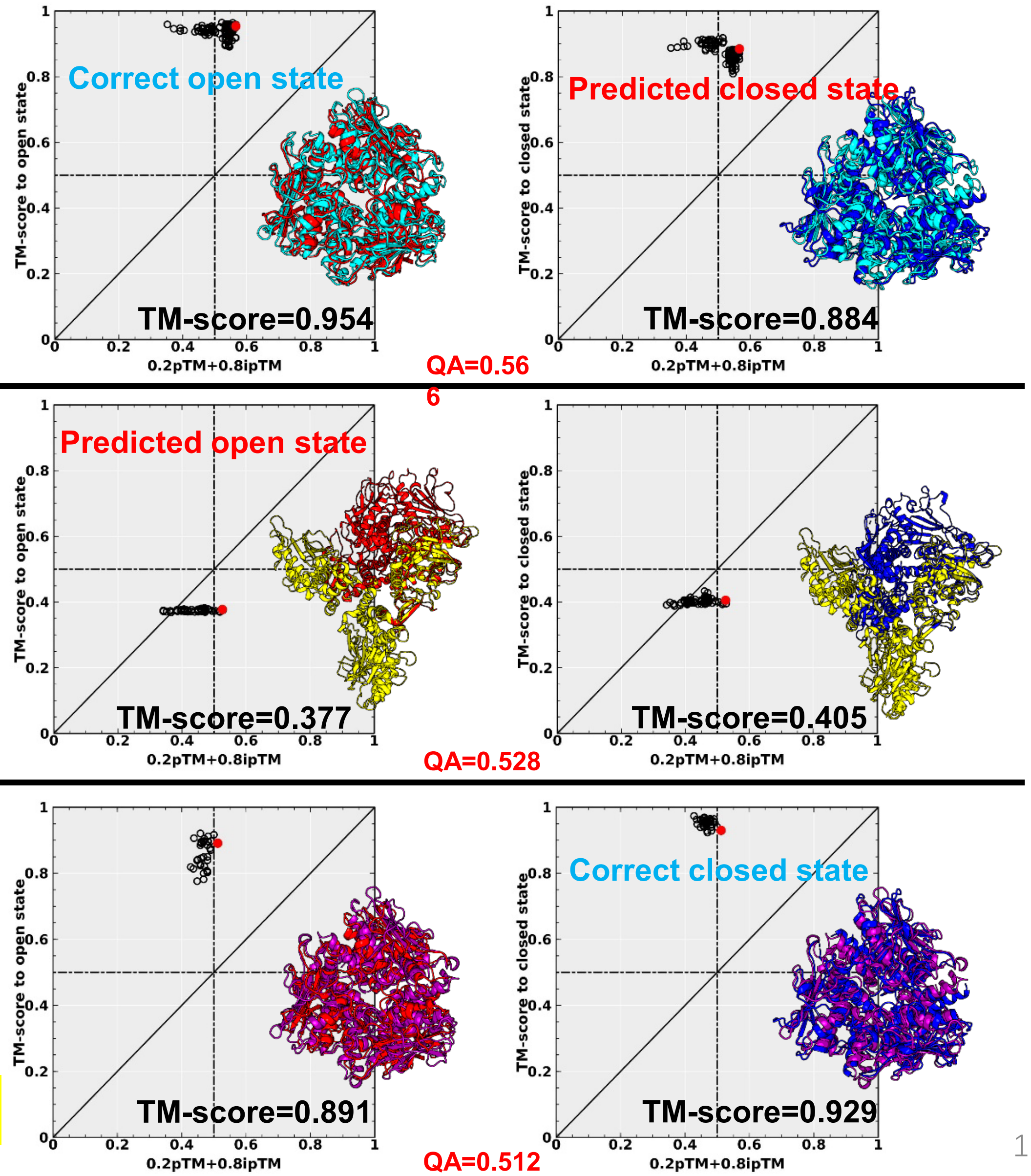




Predicted open state

Open state

Closed state



Cluster1

Correct open state

TM-score=0.954

Predicted closed state

TM-score=0.884

QA=0.56
6

Cluster2

Predicted open state

TM-score=0.377

TM-score=0.405

QA=0.528

Cluster3

TM-score=0.891

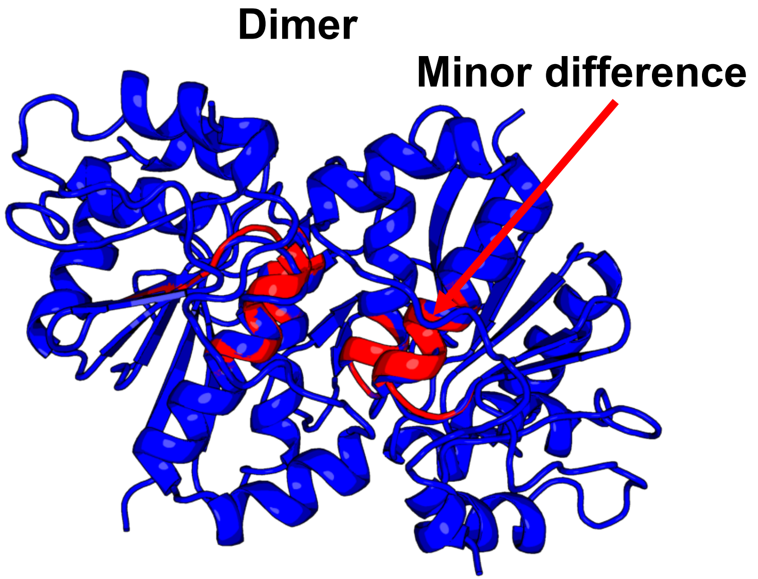
Correct closed state

TM-score=0.929

QA=0.512

Incorrect QA score ranking prevents us from selecting correct models

Results: T1294o, what went wrong?

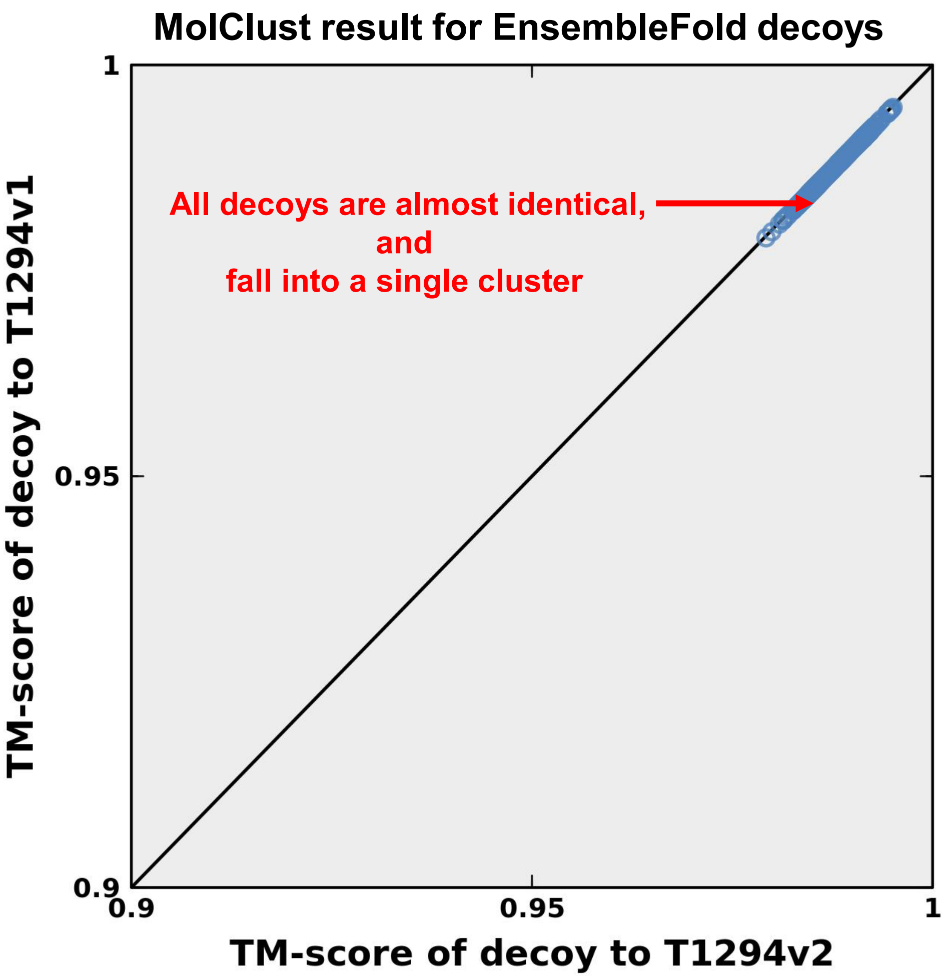


Experimental structures

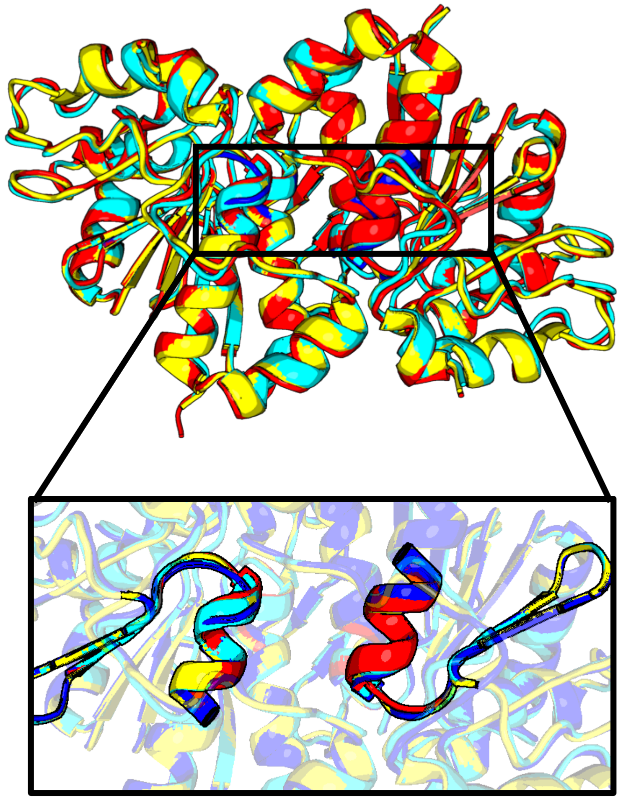
🟥 T1294v1

🟦 T1294v2

TM-score=0.999 between two states



All decoys are almost identical, and fall into a single cluster



Almost identical in the selected region

TM-score=0.988 to T1294v1
TM-score=0.989 to T1294v2

Predicting ensemble structures with minor variations remains highly challenging

Future direction: dynamic selection of clustering regions and thresholds may be necessary to emphasize sampling of important regions in candidate clusters

Summary

What went right by EnsembleFold?

- Diverse sets of MSAs help create models with multiple states for ensemble targets
- Knowledge-based REMC simulation helps create diverse set of models
- Structural clustering works well for picking the correct model of each state in most cases

What went wrong by EnsembleFold?

- Current confidence scores are not sensitive enough for selecting correct state model
- Predicting ensemble structures with minor variations remains highly challenging

Acknowledgements

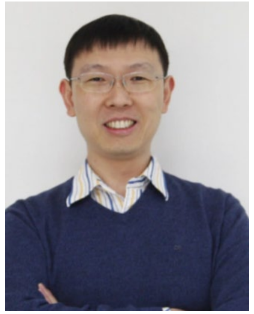
Team FZZH

Freddolino & Zheng & Zhou & Hu Team

Dr. Lydia Freddolino
Umich

Dr. Wei Zheng
Umich -> NK

Dr. Xiaogen Zhou
ZJUT

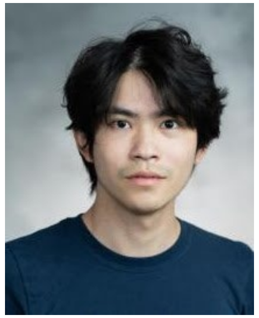


Dr. Gang Hu
NK

Dr. Pengshuo Yang
SDFNU

Dr. Qiqige Wuyun
MSU

Dr. Chunxiang Peng
Umich



Quancheng Liu
Umich

Wentao Ni
NK

Ziying Zhang
ZJUT

CASP 16

NIGMS
ACCESS

NIAID

NSF

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

Umich: University of Michigan
NK: Nankai University
ZJUT: Zhejiang University of Technology
MSU: Michigan State University
SDFNU: Shandong First Medical University

Thank you!

Q&A