CASP15 meeting, Akra Hotel, Antalya, Turkey, December 10-13, 2022

# Prediction of TS and protein assembly by trRosettaX2 and AlpahFold2

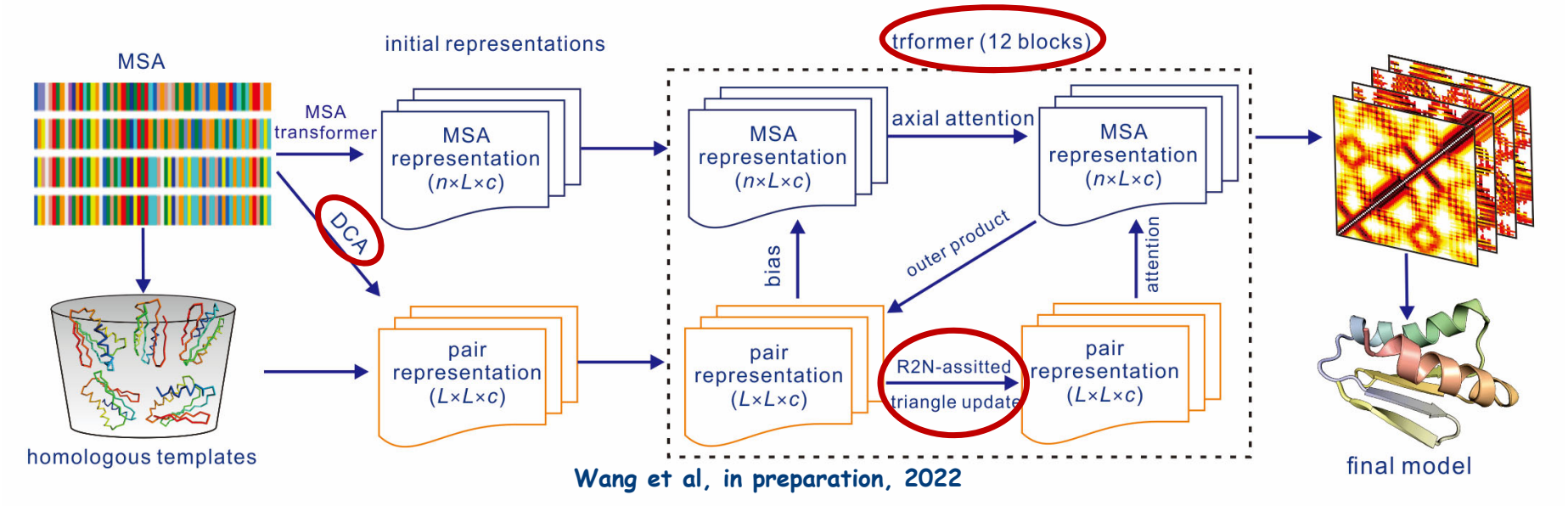
Jianyi Yang

Shandong University

http://yanglab.qd.sdu.edu.cn/

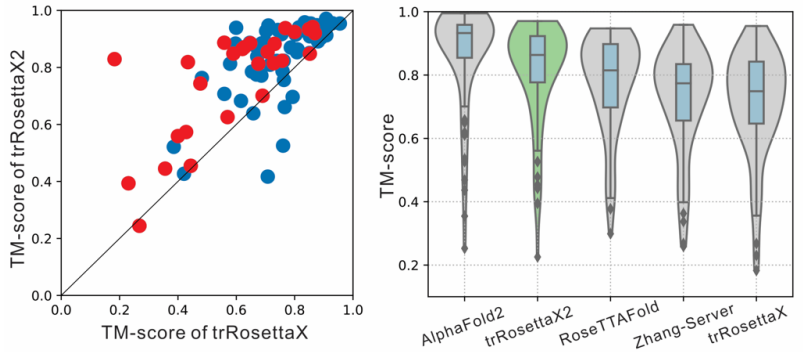
# CONTENTS

# trRosettaX2



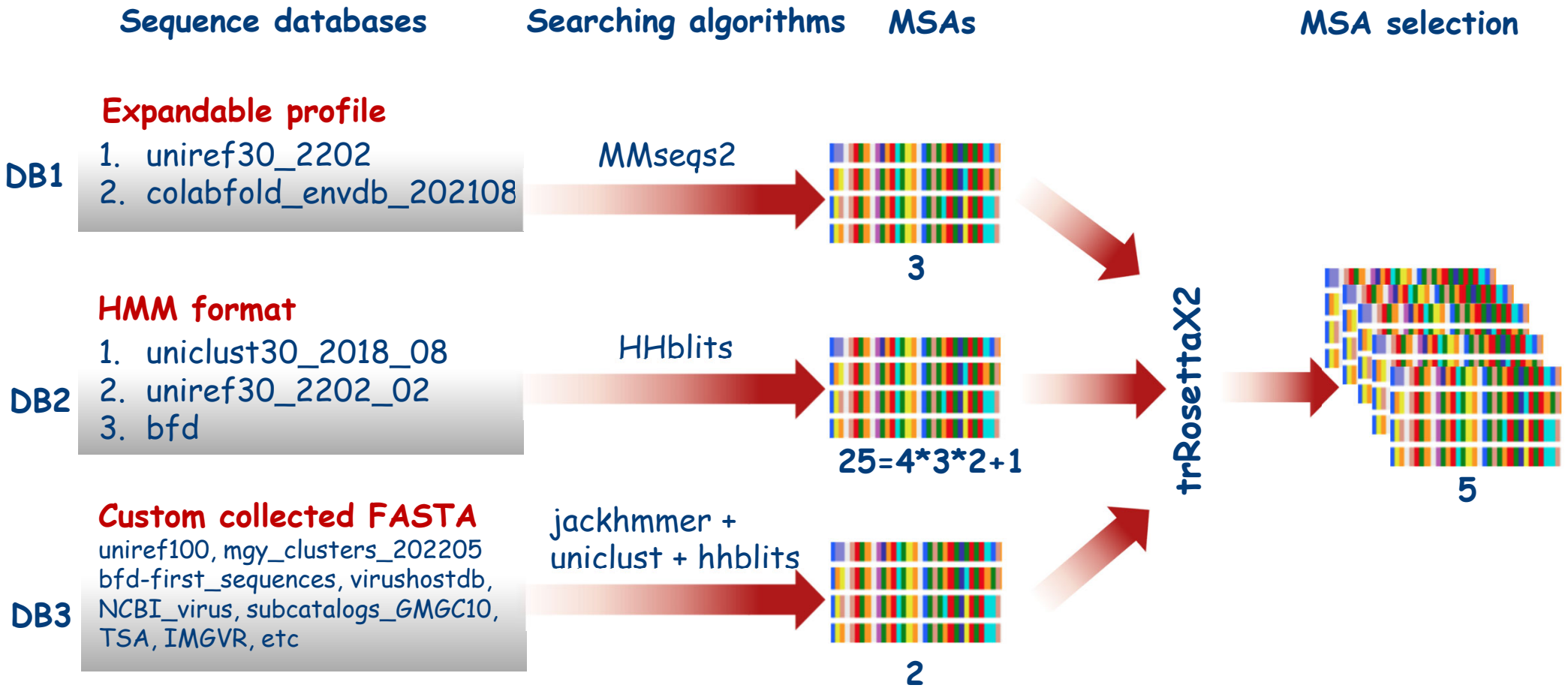

Wang et al, in preparation, 2022

CASP14 targets




Peng et al, Current Opinion in Structural Biology, 2022

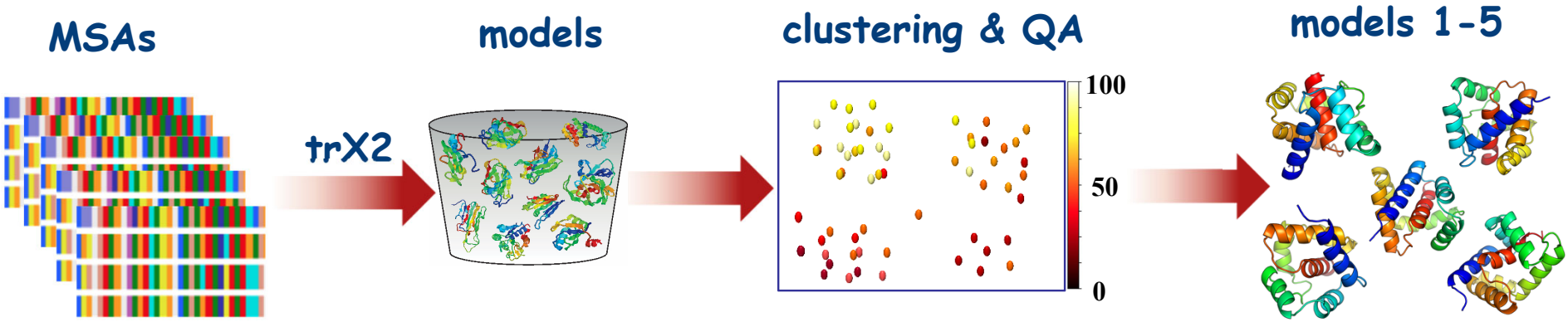
# Method: MSA generation & selection




Note: long disorder regions (by DISOPRED3) are removed before MSA generation and modeling

# Method: TS prediction by trRosettaX2

Single protein



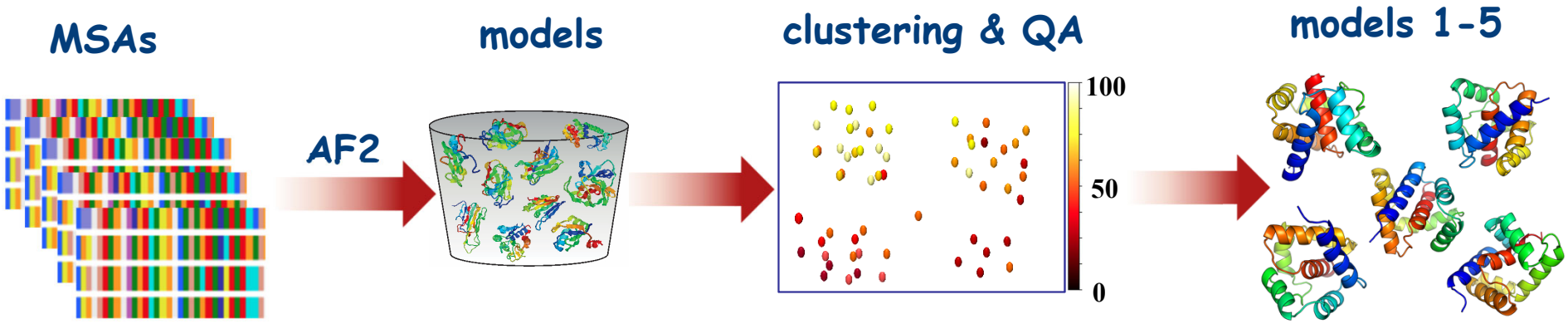

works for most regular targets, otherwise

Note:

- QA was done by a single-model based method DeepUMQA
- We tried trRosettaX-Single, when no MSA is available

# Method: TS prediction by AlphaFold2

Single protein

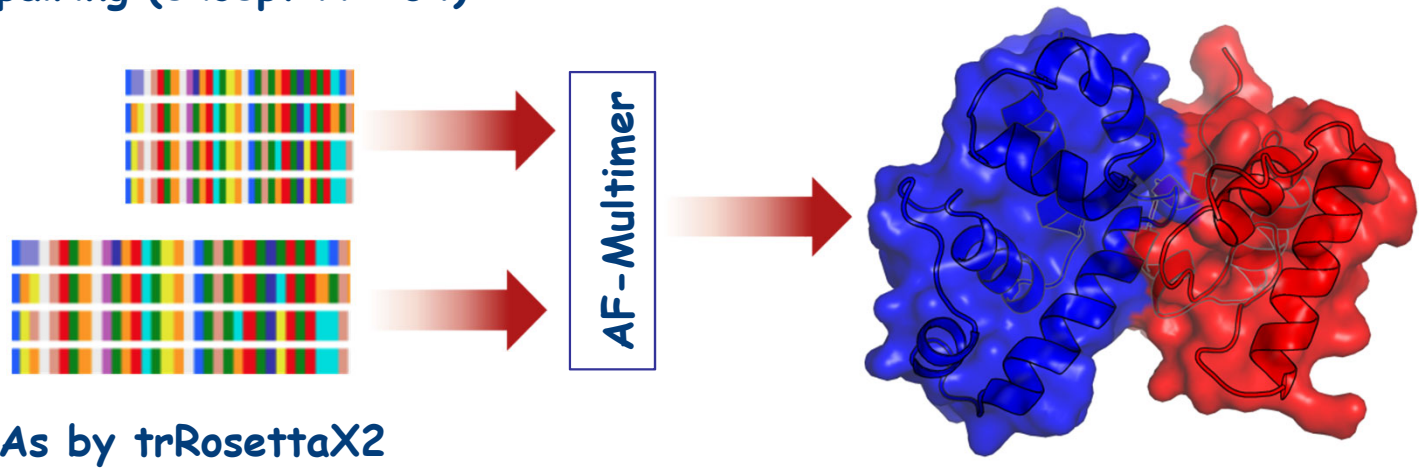



Rank trX2 and AF2 models by QA score

# Protein assembly prediction by AlphaFold-Multimer

Protein assembly

No MSA pairing (except H1134)




MSAs by trRosettaX2

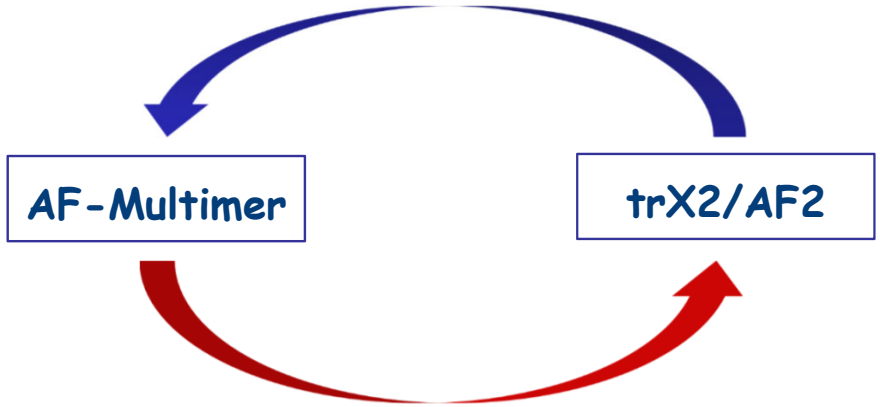
Rank models based on "iptm+ptm" score

Note: for big targets (e.g., H1111), templates are used

# Interplay between AF-Multimer and trX2/AF2

For protein assembly

Provide TS model as template, if complex model is bad (e.g., H1129)



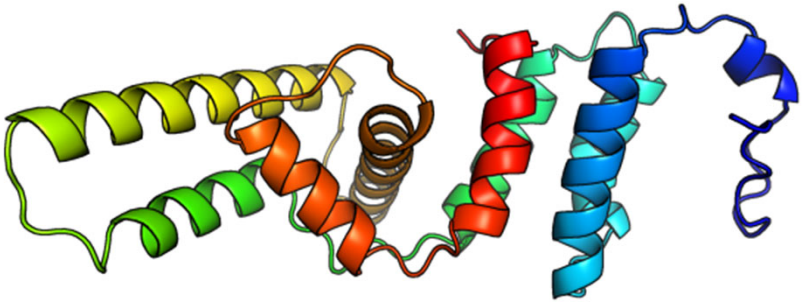

Deduce TS model from complex model, if TS model is bad (e.g., H1137)

# CONTENTS

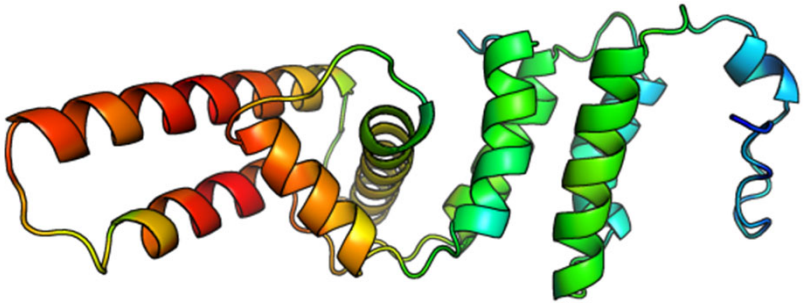
# What went right? T1130-D1

**no** homologous sequences could be detected from DB1 & DB2

Estimated TM-score in trX2: 0.35; pLDDT in AF2: 55



AF2 model color from N to C terminal



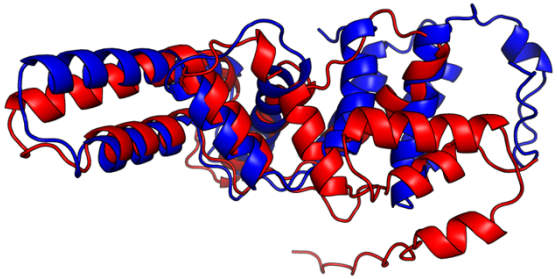
Color by pLDDT (red is high, blue is low)

real TM-score: ~0.5

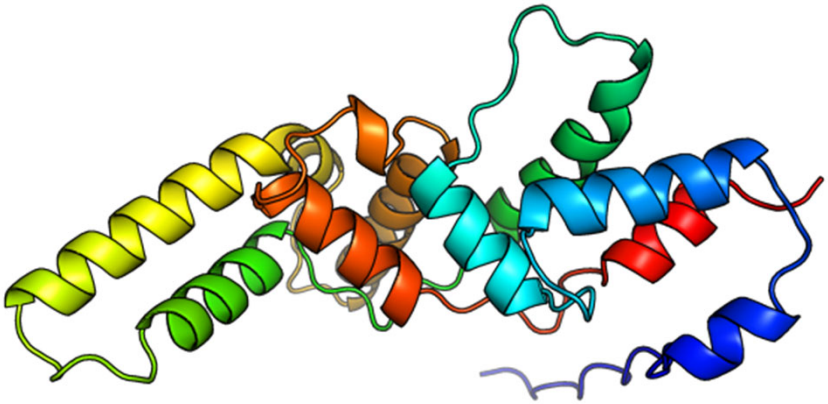
# What went right? T1130-D1

~400 homologous sequences from DB3

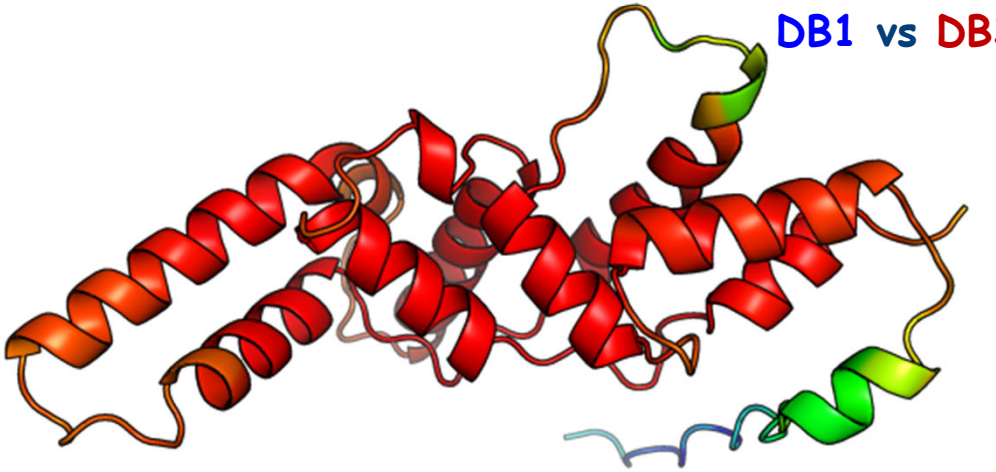
Estimated TM-score in trX2: ~0.9, pLDDT in AF2: ~90



DB1 vs DB3



AF2 model color from N to C terminal



Color by pLDDT (red is high, blue is low)

real TM-score: ~0.97

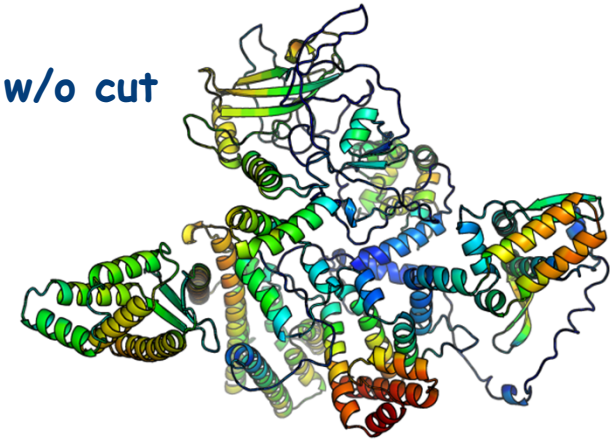
# What went right? T1125(1200 AAs)

Only one full-length sequence hit was found

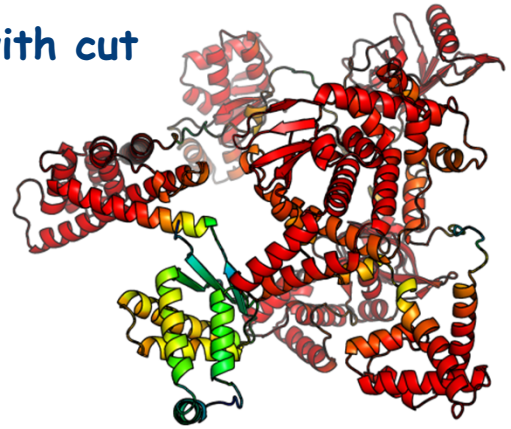
- Cut into 7 domains based on an in-house approach UniDoc (under revision)
- Assembly domain MSAs
- Use domain models as custom templates

| domain | #homo seqs |
|---|---|
| 1-139 | 2 |
| 140-325 | 58 |
| 326-466 | 6 |
| 467-607 | 6 |
| 608-816 | 440 |
| 817-935 | 195 |
| 936-1200 | 9 |
| 1-1200 | 740 |

w/o cut



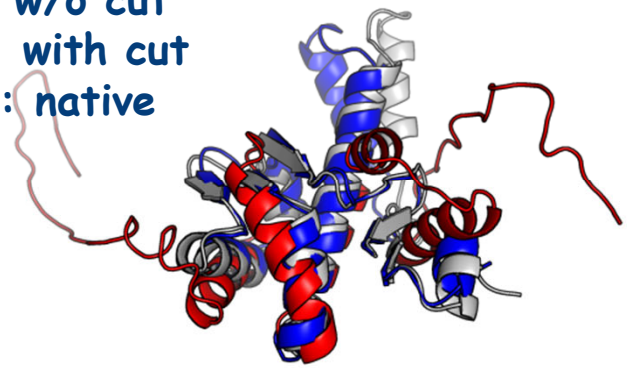
with cut



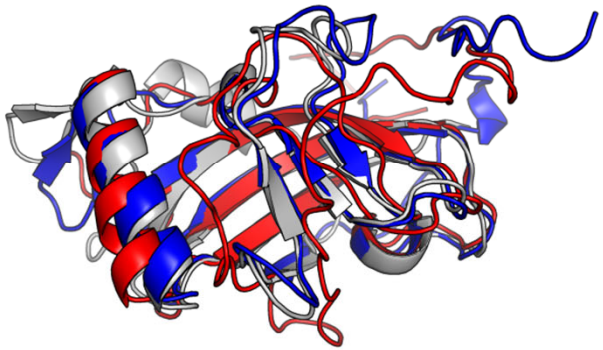
Color by pLDDT (red is high, blue is low)

# What went right? T1125(1200 AAs)

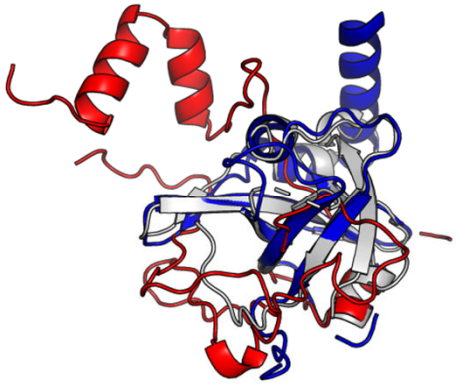
Red: w/o cut
Blue: with cut
Grey: native



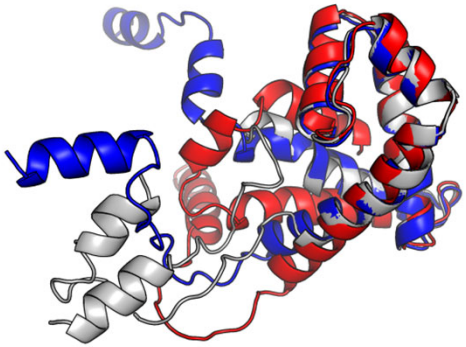
D1: 0.25/0.82



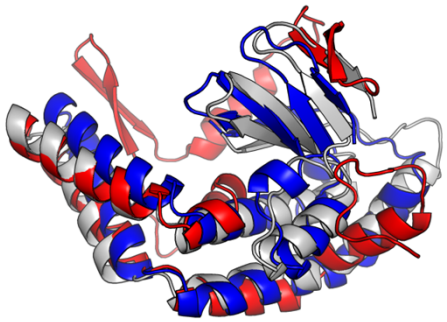
D2: 0.52/0.75



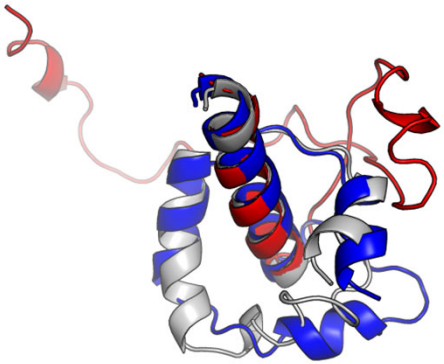
D3: 0.22/0.78



D4: 0.58/0.68



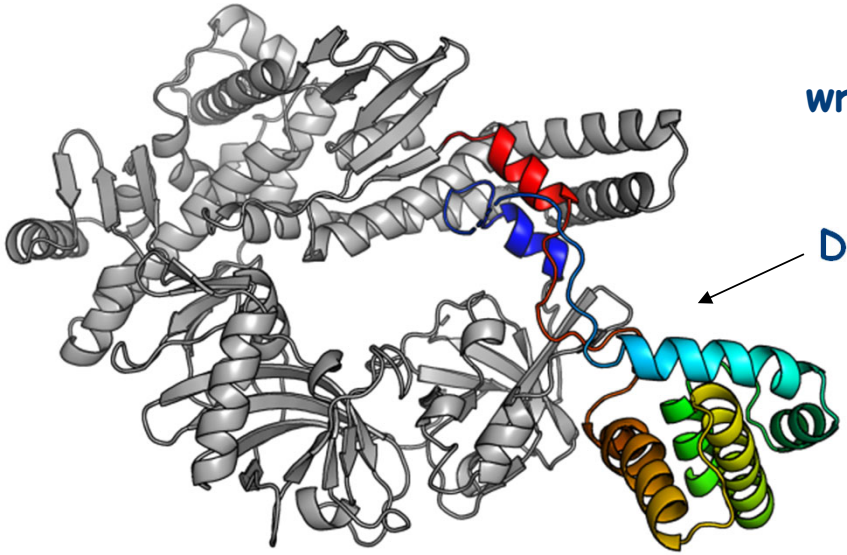
D5: 0.4/0.66



D6: 0.32/0.67

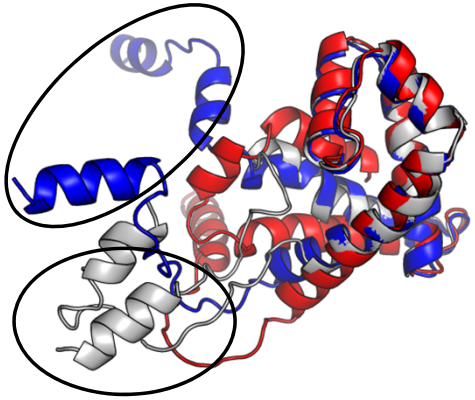
# What went wrong? T1125(1200 AAs)

wrong cut can affect the domain orientations (the whole target TM-score is ~0.3)




T1125-experimental structure

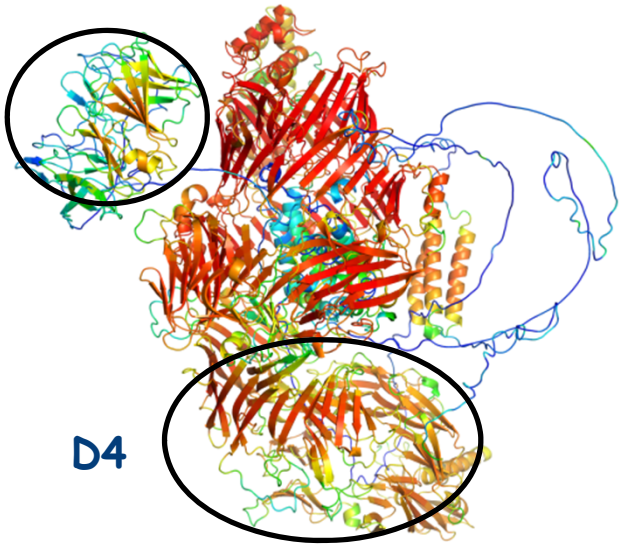
| domain | #homo seqs |
|---|---|
| 1-139 | 2 |
| 140-325 | 58 |
| 326-466 | 6 |
| 467-607 | 6 |
| 608-816 | 440 |
| 817-935 | 195 |
| 936-1200 | 9 |



# What went right? T1169(3364 AAs)

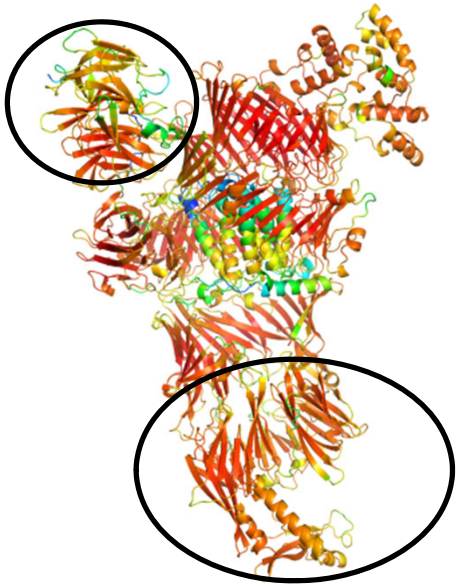
- too big to be modeled with high confidence
- Remove disordered regions: 1-26 (wrong), 2907-3364

w/o remove




Color by pLDDT

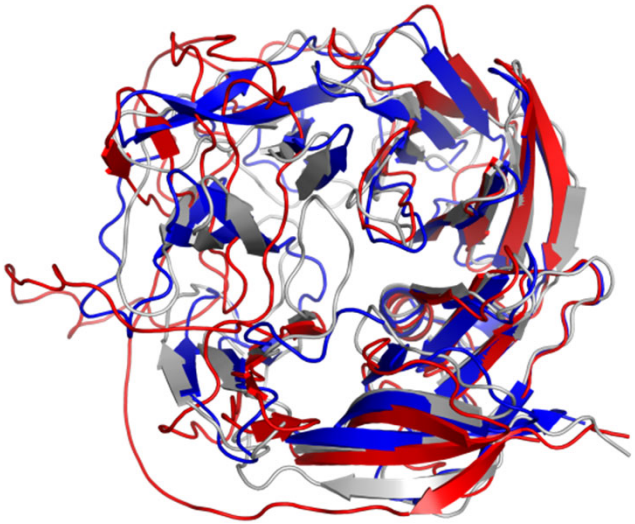
with remove



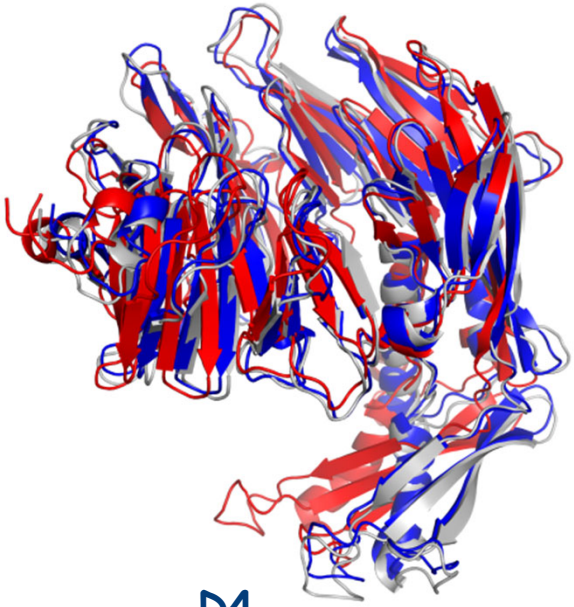
pLDDT increases from ~70 to ~80

# What went right? T1169(3364 AAs)

| TM-score | D1 (1-345) | D2 (1302-2735) | D3 (378-699,1223-1301) | D4 (700-1222) |
|---|---|---|---|---|
| w/o remove | 0.56 | 0.93 | 0.93 | 0.77 |
| with remove | 0.76 | 0.96 | 0.96 | 0.94 |



D1



D4

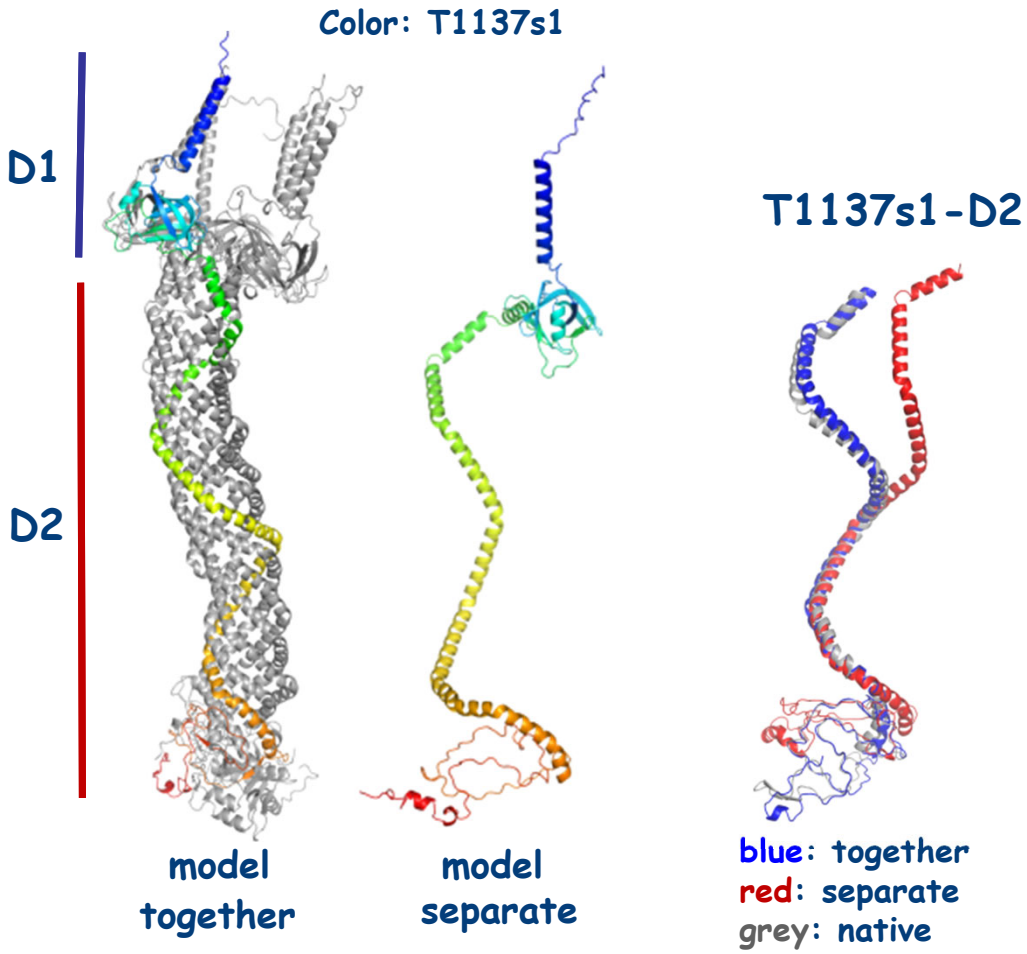
Note: Residues 1-26 were wrongly removed due to wrong disorder prediction

# What went right? T1137s1-s6

## Model S1-S6 together is important

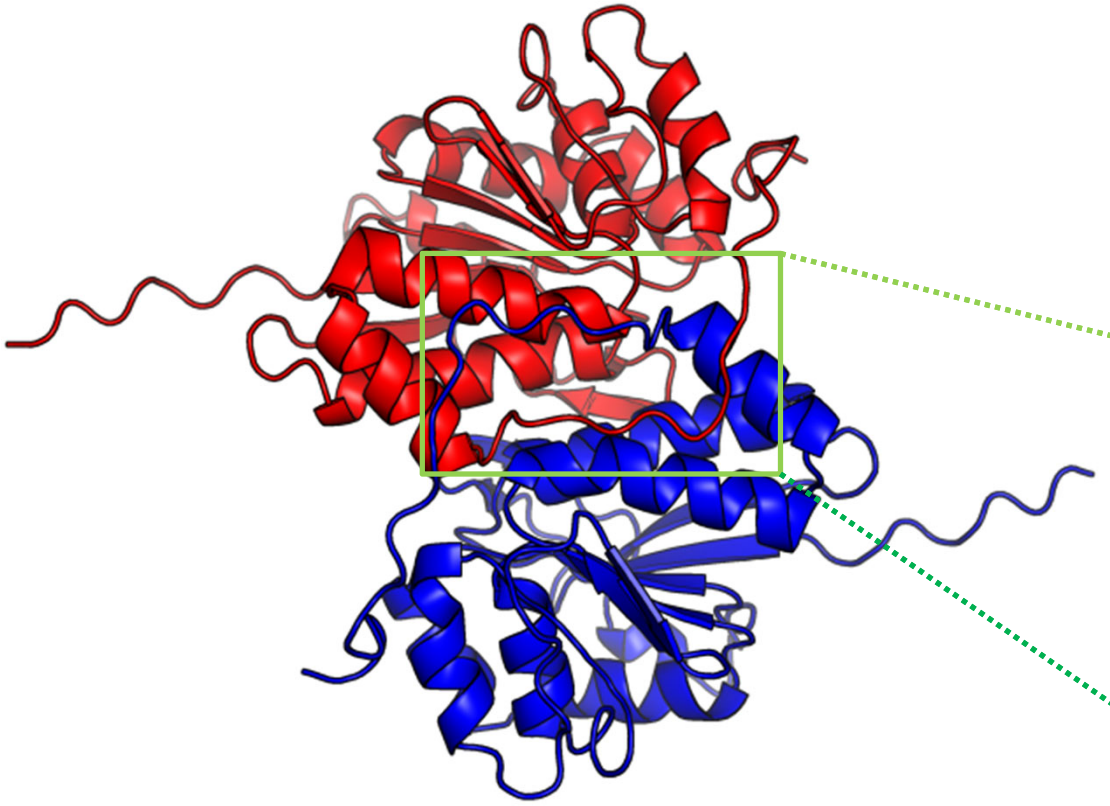
| TM-score | separate | together |
|---|---|---|
| T1137s1-D1 | 0.811 | 0.904 |
| T1137s1-D2 | 0.36 | 0.806 |
| T1137s2-D1 | 0.867 | 0.899 |
| T1137s2-D2 | 0.448 | 0.729 |
| T1137s3-D1 | 0.705 | 0.828 |
| T1137s3-D2 | 0.314 | 0.854 |
| T1137s4-D1 | 0.936 | 0.975 |
| T1137s4-D2 | 0.450 | 0.853 |
| T1137s4-D3 * | 0.874 | 0.170 |
| T1137s5-D1 | 0.936 | 0.954 |
| T1137s5-D2 | 0.408 | 0.862 |
| T1137s6-D1 | 0.860 | 0.881 |
| T1137s6-D2 | 0.503 | 0.903 |

* wrongly removed residues due to wrong disorder prediction



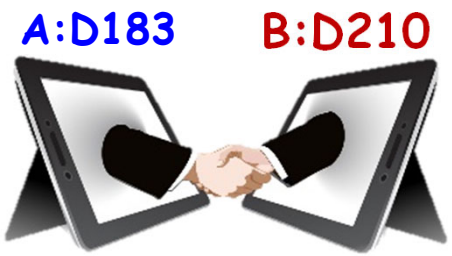
# T1110o vs T1109o

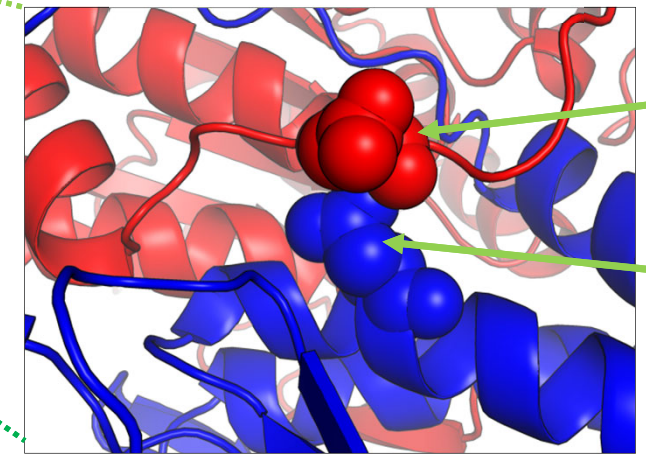
## T1110o is the wild-type, easy to predict



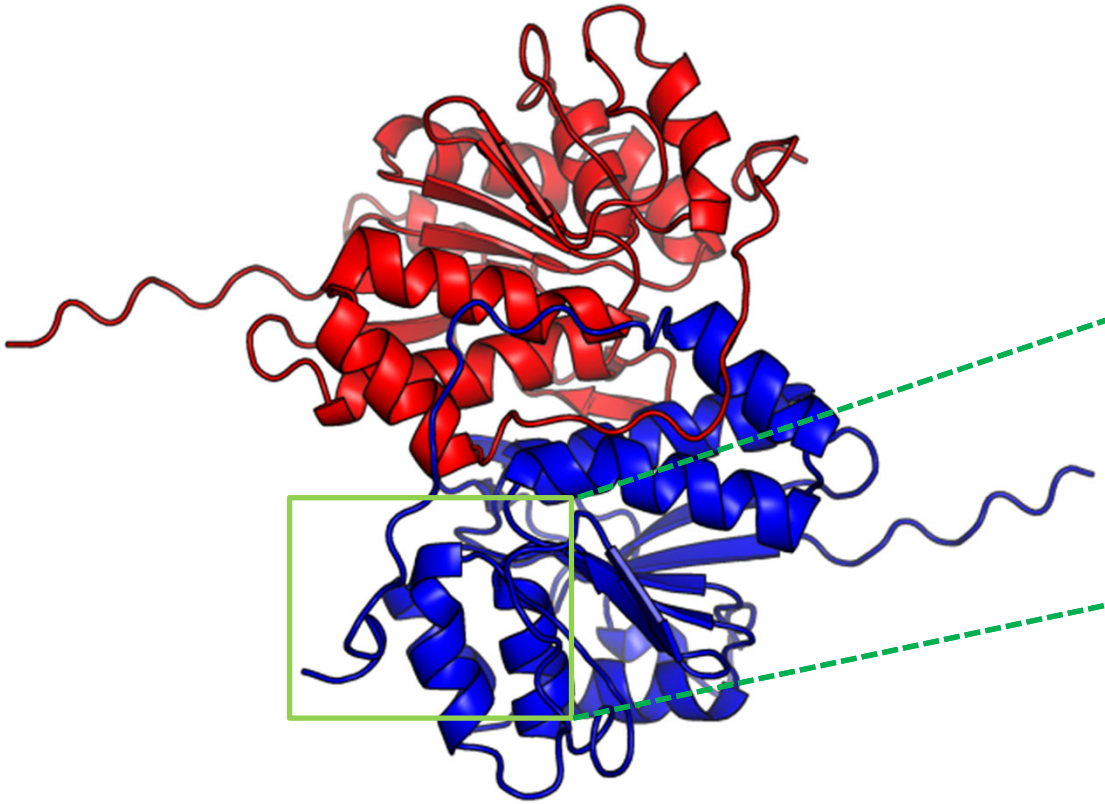
DockQ score > 0.9

**Observation 1: inter-chain interaction**



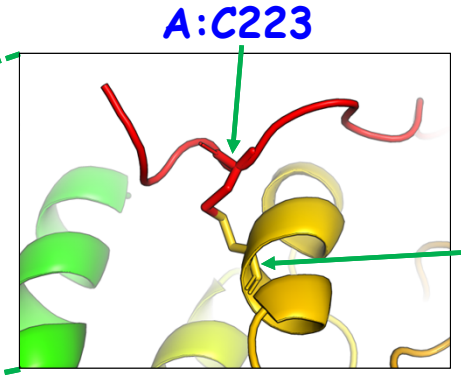



# T1110o vs T1109o



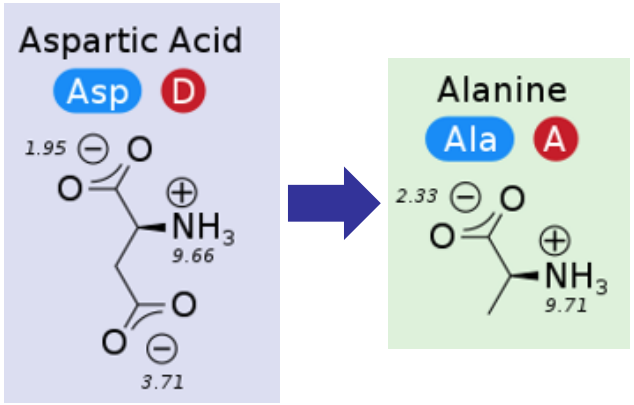
DockQ score > 0.9

Observation 2: intra-chain disulfide bond



# T1110o vs T1109o

T1109o is a "disruptive mutation D183A" of T1110o

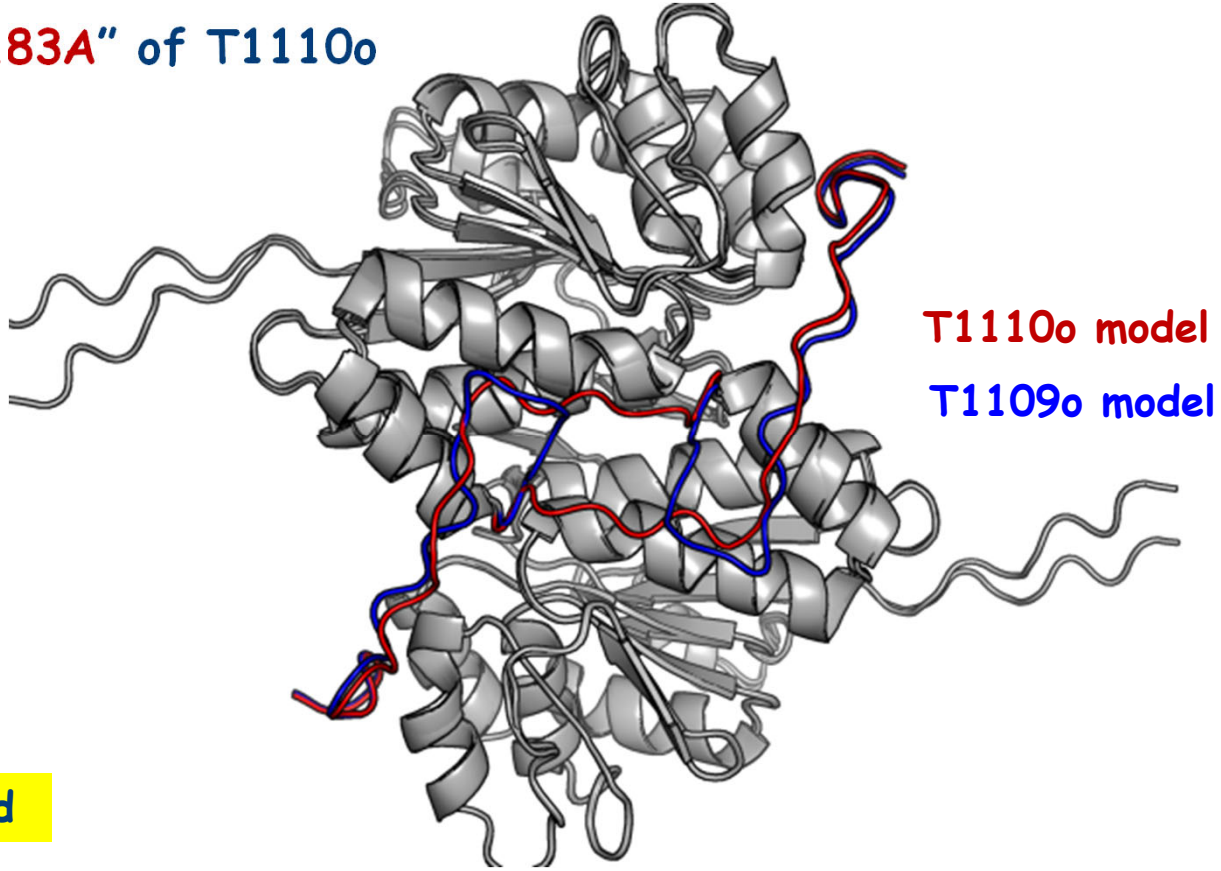



Observation 1: inter-chain interaction

A:A183 ..... B:D210

Observation 2: intra-chain disulfide bond

inter-chain disulfide bond

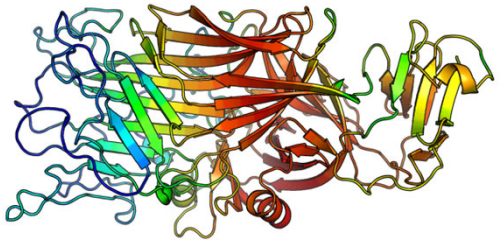



Model built with MSA from DB3
DockQ score: ~0.7

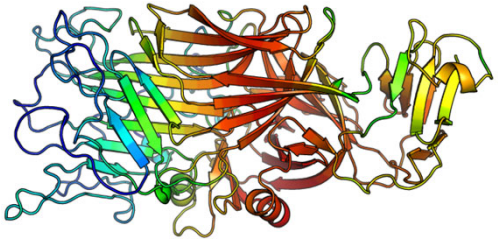
# H1129

T1129s2

DB1 MSA: 26 seqs



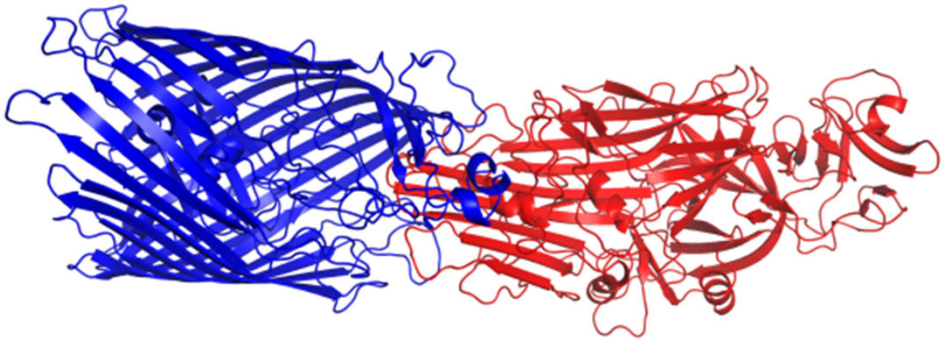
DB2 MSA: 22 seqs



DB3 MSA: 441 seqs

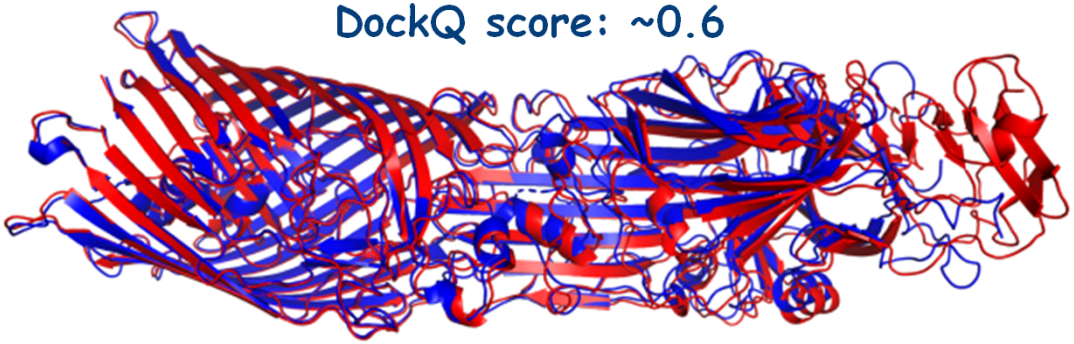
Yang-Server model

Use DB3 MSA and Yang-Server model as a template for AF-Multimer



iptm+ptm: 80 vs ~30 without template

DockQ score: ~0.6



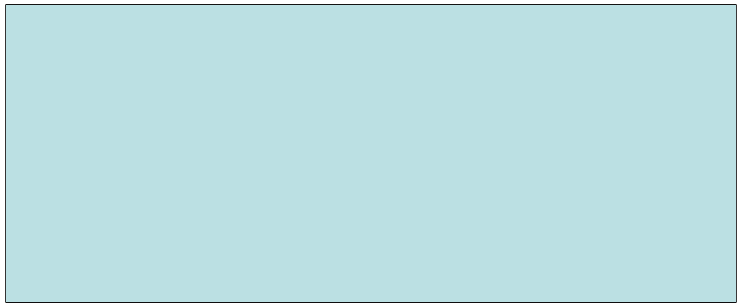
Blue: native
Red: model

# What went wrong?

T1131(173 AAs)

T1122 (241 AAs)

—no homologous sequences could be detected
—hard to fold with both trX2 and AF2

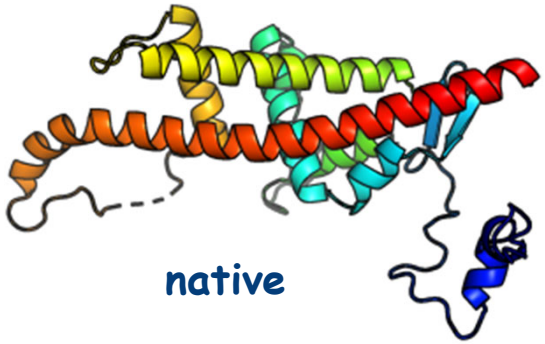


native

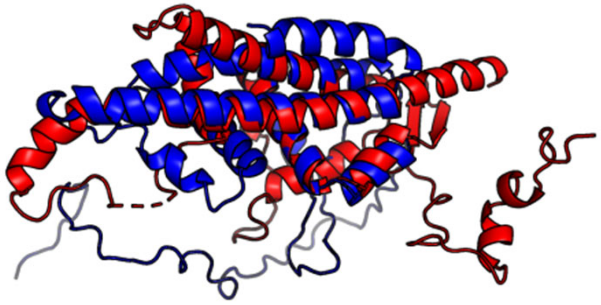


best TM-score is ~0.3

red: native
blue: model



native



best TM-score is ~0.5

# CONTENTS

# Conclusion

- MSA curation is helpful for hard targets
- PDB templates are not necessary for TS prediction
- MSA pairing is not necessary for protein assembly
- Homologous templates are important for big protein assembly (H1111)

- Single-sequence folding is still challenging (T1122, T1131)
- Protein assembly is still challenging (e.g., H1142)
- Dynamic structure is challenging

# Acknowledgments

## CASP organizers

Wenkai Wang

Hong Wei

Yang Zhang

David Baker

## Yang lab members

National Nature Science Foundation of China

# Thank you!

# Questions?


山东大学